# Análisis de la Legibilidad y Comprensibilidad en Documentos Académicos

por

**José Medardo Tapia Téllez**

Tesis sometida como requerimiento parcial

para obtener el grado de

**MAESTRÍA EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES**

por el

Instituto Nacional de Astrofísica, Óptica y Electrónica

febrero, 2022

Tonantzintla, Puebla, México

Asesores:

**Dr. Aurelio López López**

INAOE

Coordinación de Ciencias Computacionales

INAOE

©INAOE 2022

Todos los derechos reservados

El autor otorga al INAOE el permiso para

reproducir y distribuir copias de este trabajo

en su totalidad o en partes

# Análisis de la Legibilidad y Comprensibilidad en Documentos Académicos

## José Medardo Tapia Téllez

## Resumen

Esta investigación es un análisis lingüístico-computacional de textos académicos a través de su Comprensibilidad y Legibilidad. En nuestro trabajo nos referimos a Comprensibilidad como la facilidad o dificultad para entender un texto, y la Legibilidad la entendemos como la facilidad de lectura de un texto para un lector con un determinado nivel educativo. Utilizamos tres medidas de Comprensibilidad: Carga de Comprensibilidad (que refiere a la correcta secuenciación entre los conceptos), Conectividad (que se enfoca en la conexión entre conceptos) y Dispersión (que describe la densidad de información para el lector por párrafo). Metodológicamente, para Comprensibilidad, de una colección de documentos académicos seleccionados, se obtienen estas tres medidas y con ellas se generan parámetros de valoración, así como un proceso de evaluación que consiste en retroalimentación visual para el reporte parcial de las métricas en un documento nuevo (borrador). Para Legibilidad, se construyen clasificadores a través de aprendizaje de máquina los cuales son capaces de distinguir entres dos textos, cuál de ellos es más o menos legible y es con estos clasificadores se crean evaluadores que son capaces de acotar superior e inferiormente el nivel académico de un nuevo documento (borrador). Nuestras aportaciones son entonces: un procedimiento para evaluar la Comprensibilidad basado en tres medidas validadas a través de una colección de textos, un proceso de evaluación preliminar de Comprensibilidad de borradores de tesis a través de retroalimentación visual, clasificadores/comparadores basados en documentos académicos de diferentes niveles y evaluadores capaces de acotar el nivel académico de un texto. Finalmente y como trabajo a futuro, buscamos la evaluación del proceso de retroalimentación visual, la construcción de clasificadores basados en una LSTMR y la generación de una API capaz de evaluar un documento de tesis en proceso a través de nuestras aportaciones.

# Legibility and Comprehensibility Analysis on Academic Texts

**José Medardo Tapia Téllez**

## Abstract

This research is a computational linguistic analysis of academic texts through its Comprehensibility and Legibility. In our work we refer to Comprehensibility as the ease or difficulty to understand a text, and we understand Legibility as the ease of reading of a text for a reader with a specific academic level. We utilize three Comprehensibility measures: Comprehensibility Burden (which refers to the correct sequentiality between concepts), Connectivity (that focuses in the connection between concepts), and Dispersion (which describes the density of information for the reader). Methologically, for Comprehensibility, from a collection of selected academic documents, we obtain these three measures and, with them, we generate evaluation parameters as well as a evaluation process which consists in visual feedback for the partial report of the metrics in a thesis draft. For Legibility, we build classifiers through the use of machine learning which are capable to distinguish between a couple of texts, which one of them is more or less legible than the other. Based on these classifiers, evaluators are built, which are instances capable of creating top and bottom limits on the academic level of a working thesis document. Our contributions then are: a procedure to evaluate the Comprehensibility based on three measures validated through a collection of texts, a process of preliminary evaluation of thesis drafts based on visual feedback, classifiers based on academic documents of different leves, and evaluators that limit the top and bottom level of an academic thesis draft. Finally, and as future work, we want to evaluate our visual feedback process, modify the construction of classifiers and now base it on a LSTMR, and the generation of an API capable of evaluating a thesis draft through our contributions.

# Dedicatoria

El siguiente trabajo está dedicado a Pina y Medardo, gracias a que ustedes nunca dejaron de creer en mí, yo volví a hacerlo.

# Agradecimientos

Una tesis no se podría escribir sin la ayuda de un mentor, así que en primer lugar me gustaría expresar mi más profundo agradecimiento al Dr. Aurelio López, que a través de su invaluable tiempo, paciencia y conocimiento, logró guiarme a completar este proyecto. Acompañándolo se encontraron tanto el Dr. Montes, como el Dr. Martínez y el Dr. Noureddine, que por medio de sus sugerencias, crítica constructiva y observaciones perspicaces, hicieron que este proyecto de investigación se convirtiera en una realidad; les extiendo un fuerte agradecimiento. Así mismo, estoy en deuda con todo el profesorado del INAOE, entré con conocimientos superficiales de computación y a través de su cátedra, tarea y proyectos, ahora me dedico de lleno al área; esto no lo hubiese logrado sin sus enseñanzas. Aprendí tanto de los profesores del INAOE como de mis compañeros de maestría, les agradezco todo este conocimiento compartido, su amistad y todos los buenos momentos que pasamos juntos en el instituto. También agradecer al INAOE, a que funciona como una máquina bien aceitada debido a todas las personas que están detrás de él, es importante reconocer el esfuerzo de todos y cada uno de estos individuos en su labor diario.

Quiero continuar expresando mis agradecimientos mirando ahora hacia mi familia, este paso en mi carrera no hubiese sido posible sin el incansable apoyo de mi Padre, el cariño de Pina, la instrumental guía de Lutz y la fe ciega de José Emilio y Víctor Hugo; ustedes han sido claves en mi desarrollo humano y profesional. Así mismo, sin el apoyo y soporte que me brindó TaniBú al comienzo y final de este proyecto, no hubiese sido posible terminarlo, el cariño y amor nunca deben ser subestimados en este tipo de tareas. Por último, también tengo que agradecer a mis amigos, es tan importante trabajar duro, como descansar, vaya que ustedes, a través de risas, aligeran el desgastante proceso que es escribir una tesis de maestría.

# Índice general


**1. Introducción** **1**
- 1.1. Temática . . . 1
  - 1.1.1. Legibilidad . . . 2
  - 1.1.2. Comprensibilidad . . . 4
- 1.2. Motivación . . . 4
- 1.3. A quién está dirigido . . . 6
- 1.4. Problemática y Justificación . . . 6
  - 1.4.1. Problemática . . . 6
  - 1.4.2. Justificación . . . 7
- 1.5. Objetivos . . . 7
  - 1.5.1. Objetivo General . . . 7
  - 1.5.2. Objetivos Específicos . . . 7
- 1.6. Contribuciones . . . 8
  - 1.6.1. Comprensibilidad . . . 8
  - 1.6.2. Legibilidad . . . 9
- 1.7. Organización del Documento . . . 9

**2. Marco Teórico** **11**
- 2.1. Aspectos del Lenguaje . . . 11
  - 2.1.1. Lenguaje . . . 11
  - 2.1.2. Palabras . . . 13
  - 2.1.3. Oración . . . 13
  - 2.1.4. Párrafo . . . 13
- 2.2. Lingüística . . . 14
  - 2.2.1. Lingüística Computacional . . . 14
  - 2.2.2. Procesamiento de Lenguaje Natural (PLN) . . . 14
  - 2.2.3. Etiquetado de Partes del Discurso . . . 15
- 2.3. Documentos Académicos . . . 15
  - 2.3.1. Secciones Seleccionadas de la Tesis . . . 16
- 2.4. Entropía de Shannon . . . 17
- 2.5. Retroalimentación Visual . . . 17
  - 2.5.1. Nube de Términos . . . 17
  - 2.5.2. Teoría de Grafos . . . 18
  - 2.5.3. Gráfica de Barras . . . 19
- 2.6. Aprendizaje de Máquina . . . 19
  - 2.6.1. Máquina de Soporte Vectorial Lineal . . . 20
  - 2.6.2. $K$-vecinos más cercanos (KNN) . . . 21
  - 2.6.3. Perceptrón . . . 21

2.6.4. Exactitud . . . 22
2.6.5. Evaluación Automática de Texto . . . 23

**3. Trabajo Relacionado 25**
3.1. Comprensibilidad . . . 25
3.2. Legibilidad . . . 29

**4. Metodología en Comprensibilidad 36**
4.1. Representación . . . 36
4.1.1. Secuencias de Oraciones y Conceptos . . . 36
4.1.2. Grafo de Conceptos . . . 37
4.2. Medidas . . . 37
4.2.1. Carga de Comprensibilidad (Secuencialidad) . . . 37
4.2.2. Conectividad . . . 39
4.2.3. Dispersión . . . 40
4.2.4. Ejemplo de Obtención de Representación . . . 41
4.3. Retroalimentación Visual . . . 42
4.3.1. Nubes de Términos . . . 43
4.3.2. Grafos . . . 43
4.3.3. Gráfica de Barras . . . 44

**5. Colección de Datos 46**
5.1. Número de Documentos . . . 46
5.2. Documentos Depurados . . . 47
5.3. Secciones . . . 47
5.4. Estadística para Palabras . . . 48

**6. Evaluación y Retroalimentación Visual en Comprensibilidad 49**
6.1. Evaluación . . . 49
6.1.1. Experimentos . . . 49
6.1.2. Evaluación de la sección “Planteamiento del Problema” . . . 50
6.1.3. ón de la sección “Justificación” . . . 51
6.1.4. Evaluación de la sección “Resultados” . . . 52
6.1.5. Análisis de las Medidas en la Colección de Textos Académicos 53
6.2. Retroalimentación Visual . . . 54
6.2.1. Ejemplo de Retroalimentación Visual en un Documento de TSU 54
6.2.2. Ejemplo de Retroalimentación Visual en un Documento de Doctorado . . . 58
6.2.3. Análisis sobre Retroalimentación Visual . . . 60

**7. Metodología en Legibilidad 61**
7.1. Clasificador/Comparador . . . 61
7.1.1. Representación de Documentos . . . 61
7.1.2. Entrenamiento del Clasificador/Comparador . . . 62
7.2. Evaluador . . . 62
7.2.1. Aleatorio . . . 63
7.2.2. Centroide . . . 64
7.2.3. Máximo . . . 64
7.2.4. Mínimo . . . 65

**8. Resultados para Legibilidad** **66**
8.1. Clasificadores/Comparadores . . . . . 66
8.1.1. Experimentos . . . . . 66
8.1.2. Resultados para Clasificadores/Comparadores Operados con Diferencia . . . . . 67
8.1.3. Resultados para Clasificadores/Comparadores Operados con División . . . . . 68
8.1.4. Resultados para Clasificador/Comparadores Operados con Concatenación . . . . . 68
8.1.5. Conclusiones para Clasificadores/Comparadores . . . . . 69
8.2. Evaluadores . . . . . 70
8.2.1. Experimentos . . . . . 70
8.2.2. Resultados para la sección de "Planteamiento del Problema . . 71
8.2.3. Justificación . . . . . 73
8.2.4. Resultados . . . . . 74
8.3. Análisis Generales . . . . . 75

**9. Conclusiones** **77**
9.1. Conclusiones en Comprensibilidad . . . . . 77
9.1.1. Evaluación del Proceso . . . . . 77
9.1.2. Conclusiones Principales . . . . . 78
9.1.3. Resultados en Contexto . . . . . 79
9.1.4. Trabajo a Futuro . . . . . 80
9.2. Conclusiones en Legibilidad . . . . . 81
9.2.1. Evaluación del Proceso . . . . . 81
9.2.2. Conclusiones Principales . . . . . 82
9.2.3. Resultados en Contexto . . . . . 82
9.2.4. Trabajo a Futuro . . . . . 83
9.3. Publicaciones Derivadas . . . . . 84

**A. Apéndice** **85**
A.1. Evaluadores . . . . . 85
A.1.1. Centroide . . . . . 85
A.1.2. Mínimo . . . . . 86
A.1.3. Máximo . . . . . 86
A.1.4. Algoritmo de Selección de Máximo y Mínimo . . . . . 87

# Índice de figuras


2.1. Ejemplo de una Nube de Términos . . . . . 17
2.2. Ejemplo de grafo no dirigido . . . . . 18
2.3. Ejemplo de grafo ponderado . . . . . 18
2.4. Ejemplo de gráfica de barras . . . . . 19
2.5. Ejemplo de clasificación a través de una SVM que aparece en [6] . . . 21
2.6. Ejemplo de perceptrón . . . . . 22

4.1. Ejemplo de corrida para obtención de la representación en un documento aleatorio del corpus de datos . . . . . 42
4.2. Nube de términos para un documento aleatorio de Doctorado . . . . 43
4.3. Grafo de relaciones entre conceptos para un documento aleatorio de Doctorado . . . . . 44
4.4. Gráfica de barras para un documento aleatorio de Doctorado . . . . . 45

6.1. Representación visual para frecuencia de conceptos y sus relaciones . 55
6.2. Relación entre conceptos y su párrafo clave . . . . . 56
6.3. Nube de términos con base en la medida de carga de comprensibilidad por concepto . . . . . 56
6.4. Grafo de Conceptos para el Documento . . . . . 57
6.5. Nube de términos con base en la medida de Conectividad por concepto 57
6.6. Representación visual para la medida de dispersión a través de gráficas de barras (TSU Y Doctorado) . . . . . 58
6.8. Grafo de Documento para una Tesis de Nivel Doctorado . . . . . 59
6.7. Nube de términos para conceptos . . . . . 59

7.1. Representación para documentos en legibilidad . . . . . 62
7.2. Diagrama de entrenamiento del Clasificador/Comparador . . . . . 63
7.3. Esquema para el evaluador Aleatorio . . . . . 64

A.1. Esquema para el evaluador basado en centroide . . . . . 85
A.2. Esquema para el evaluador basado en mínimos . . . . . 86
A.3. Esquema para el evaluador basado en máximos . . . . . 86

# Índice de tablas


5.1. Nivel académico y número de documentos . . . . . 46
5.2. Nivel académico y documentos depurados . . . . . 47
5.3. Secciones por nivel académico . . . . . 47
5.4. Estadísticas para palabras por secciones y nivel académico . . . . . 48

6.1. Estadísticas de las Medidas de Comprensibilidad para la sección de Planteamiento del Problema . . . . . 51
6.2. Estadísticas relacionadas con las Medidas de Comprensibilidad para la Sección de Justificación . . . . . 52
6.3. Estadísticas de las Medidas de Comprensibilidad para la Sección de Resultados . . . . . 53

8.1. Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a Diferencia. . . . . 67
8.2. Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a División. . . . . 68
8.3. Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a Concatenación. . . . . 69
8.4. Resultados para evaluadores con posibles referentes y modelo SVM . 72
8.5. Resultados de evaluadores con posibles referentes y modelo SVM . . . 72
8.6. Resultados de evaluadores con diferentes referentes y modelo KNN . . 72
8.7. Resultados para evaluadores con posibles referentes y modelo SVM . 73
8.8. Resultados para Evaluadores con posibles referentes y modelo SVM . 73
8.9. Resultados de evaluadores con todos los referentes con Clasificador/Comparador TSU-Maestría y modelo KNN . . . . . 74
8.10. Resultados para Evaluadores con sus posibles referentes y modelo Perceptrón . . . . . 74
8.11. Resultados para Evaluadores con sus posibles referentes y modelo SVM 75
8.12. Resultados de Evaluadores con posibles referentes . . . . . 75

# Capítulo 1

# Introducción

En este primer capítulo presentamos, en primera instancia, la temática que se aborda en este proyecto de investigación al identificar definiciones introductorias clave. Después brindamos la motivación del trabajo, donde se indica el porqué se decide realizar la investigación. A esto le siguen los beneficiarios directos del proyecto de investigación, así como a quién está dirigido. Posteriormente, planteamos la problemática y el porqué ésta es pertinente, es a través de ella que después se presentan los objetivos del trabajo, general y particulares. A continuación de los objetivos, se presenta la metodología a seguir para abordarlos y el capítulo cierra describiendo la organización del documento en general.

## 1.1. Temática

El tema principal de este trabajo de investigación cae en el área de la lingüística computacional. En específico trabajamos sobre dos aspectos del área: Comprensibilidad y Legibilidad. La primera refiriéndose a la facilidad o dificultad para entender un texto y la segunda entendida como la facilidad de lectura de un texto para un lector con un determinado nivel educativo. Con base en la Comprensibilidad son tres las medidas de nuestro interés: Carga de Comprensibilidad, Conectividad y Dispersión. Todas las medidas se basan en el análisis de los conceptos en un documento de tesis, en específico, la primera analiza la secuencialidad, la segunda la conexión y la tercera la densidad. Con estas medidas generamos parámetros de valoración para finalmente utilizar el mismo proceso de obtención de las medidas en la creación de un proceso de retroalimentación visual.

Finalizada la primera parte de nuestro proyecto de investigación, se pasa a la que involucra a la Legibilidad. Aquí se construyen clasificadores a través de aprendizaje de máquina, los cuales fungen como comparadores que son capaces de determinar entre dos textos, cuál de ellos es más o menos legible, i.e. cuál de los dos es más fácil de leer que el otro. Es con estos clasificadores que se construyen evaluadores los cuales son capaces de acotar superior e inferiormente el nivel académico de un documento nuevo o en proceso.

En resumen, nuestro proyecto utiliza Minería de Textos, Procesamiento de Lenguaje Natural, Aprendizaje de Máquina, visualización de texto y Lingüística para generar evaluaciones tanto de Comprensibilidad como de Legibilidad en documentos de tesis en proceso (borradores).

Enfocándonos al área más general, la Lingüística Computacional. Según [5], ésta

busca el desarrollar maquinaria con contenido lingüístico necesario para que un agente (ser humano o programa computacional) exhiba algún comportamiento lingüístico. Entonces, y con base en la definición, concluimos que la Lingüística Computacional es una integración de las Ciencias Computacionales y de la Lingüística. En principio, la conexión entre estas dos áreas, puede parecer extraña, sin embargo, el lenguaje involucra sistemas de símbolos y las computadoras son rápidas al procesar éstos. El desarrollo sobre estas dos áreas en conjunto, ha permitido herramientas tan poderosas como los motores de búsqueda en Internet.

Los motores de búsqueda mencionados, así como nuestra investigación, utilizan el Procesamiento de Lenguaje Natural (PLN). Según [22], PLN es un rango de técnicas computacionales motivadas teóricamente para analizar y representar texto que ocurre naturalmente en uno o más niveles de análisis lingüístico, esto para lograr un procesamiento humanoide de una gama de tareas o aplicaciones lingüísticas.

Con base en la definición, el "rango de técnicas computacionales" refiere a que encontramos múltiples métodos para lograr un análisis de lenguaje en particular. La noción de "textos que ocurren naturalmente" expresa que pueden ser de cualquier lenguaje, moda o género. El único requerimiento para los textos es que sean de un lenguaje usado por humanos y que no sea construido solo para su análisis, sino que tenga un uso natural. La noción de "niveles de análisis lingüístico" refiere a que hay múltiples tipos de procesamiento de lenguaje y teóricamente los humanos ocupamos todos al mismo tiempo. El "procesamiento humanoide de lenguaje" rebela que PLN es considerado una disciplina de la Inteligencia Artificial (IA) dado que proviene de una capacidad humana. Finalmente utilizarlo en una "gama de tareas" indica que PLN no es la meta como tal, excepto quizá para investigadores en IA. Para otras personas, PLN es el medio por el cual se realizan ciertas tareas, como podría ser sistemas de Recuperación de Información (RI), así como sistemas de evaluación de Comprensibilidad o de Legibilidad.

Íntimamente relacionado con PLN, se encuentra la Minería de Textos [15], que trata sobre el análisis de texto con apoyo de máquinas. Minería de Textos utiliza técnicas de RI, extracción de la información, así como PLN y las conecta con los algoritmos para minería de datos, aprendizaje de máquina y estadística. En nuestro proyecto de investigación, utilizamos Minería de Textos para la extracción de información e identificación de patrones en los textos, esto tanto para Comprensibilidad como para Legbilidad.

### 1.1.1. Legibilidad

Según [3], la Legibilidad implica referirnos a la suma de todos los elementos en un material textual que afectan el entendimiento del lector, la velocidad de lectura y el nivel de interés del material. Estos elementos pueden incluir características como lo serían complejidad de la sintaxis de la oración; la familiaridad de la semántica hacia los conceptos discutidos, si hay apoyos visuales; la sofisticación de los argumentos lógicos o la inferencia utilizada para conectar las ideas; y muchas otras dimensiones de contenido. Así mismo, la legibilidad también depende del lector, sus antecedentes educativos y sociales, intereses y motivación para aprender.

El aprender se traduce hacia una necesidad de información y para que el aprendizaje sea efectivo se necesita una información de calidad. Evaluar eficazmente documentos de texto es entonces imprescindible y esto es posible a través de medidas

de Legibilidad, que toman de entrada un texto y estiman un valor numérico, así mismo hay otro tipos de formas de evaluación. La evaluación predictiva indica el nivel o grado de legibilidad, siendo de nuestro interés ya que nos interesa evaluar un documento de texto dentro de un cierto nivel académico.

Las medidas de legibilidad han venido evolucionando fuertemente en años recientes, a través de nuevos métodos computacionales, así como de nuevas fuentes de datos y aplicaciones en línea y redes sociales. Por un lado se tienen las medidas de legibilidad tradicionales, e.g. [33], que estima la legibilidad de textos basándose en funciones simples de dos o tres variables lingüísticas como el conteo de sílabas y palabras. Actualmente vemos un cambio de estas medidas tradicionales y superficiales en favor de algoritmos de evaluación basados en datos. Éstos se basan en representaciones derivadas a través lingüística computacional y se combinan con sofisticados modelos de predicción basados en aprendizaje de máquina, para así lograr análisis más profundos, precisos y orientados a textos más diversos.

Nuestra evaluación en legibilidad, utiliza aprendizaje de máquina y, en general, un enfoque para predicción de la legibilidad, consiste de los siguientes pasos:

1. Un conjunto de datos con un etiquetado generado a través de un panel de expertos.

2. Un conjunto de características que se definen y deben ser obtenidas de un texto. Éstas capturan atributos semánticos, sintácticos o de otro tipo.

3. Un modelo de aprendizaje de máquina que aprende a predecir las etiquetas proveídas por nuestro panel de expertos y que son parte del conjunto de datos de entrenamiento. Esto se logra a través de los valores de las características extraídas, de aquí que para cada texto de entrenamiento se extraen características específicas para formar un vector representativo a ese texto. El modelo se entrena con base en estos vectores representativos y su respectivo etiquetado.

Con base en el primer punto, la colección utilizada en nuestro proyecto de investigación se compone por tesis de cuatro diferentes niveles académicos: Técnico Superior Universitario, Licenciatura, Maestría y Doctorado. Esta colección ha sido sometida a un panel de expertos que aceptó y valoró los proyectos de investigación (instituciones educativas, sinodales y directores de tesis) y fue construida por [9]. Con base en el segundo punto, las características que hemos elegido para representar a los documentos son las léxico-semánticas. A continuación presentamos las exploradas en la literatura:

- Léxico-semánticas: palabras raras, no familiares o ambiguas.

- Morfológicas: La morfología, en lingüística, estudia la relación entre el significado y la forma tanto en las palabras como entre ellas. Un ejemplo del estudio de estas características podrían ser las partículas morfológicas, éstas son características basadas en inflexión y derivación del idioma Alemán [12].

- Sintaxis: estructuras gramaticales

- De discurso:

    - Organización micro-estructural del texto: utilización de conectivos y otras características de cohesión para clarificar relaciones o transiciones.

    - Organización macro-estructural: partes explícitas que caracterizan una estructura argumental

- Semánticas de alto nivel: utilización de inusual de sentidos, idiomas, o connotaciones sutiles; dominio o conocimiento del mundo requerido para comprender el texto

- Pragmáticas: contextuales o sujetas a influencia de género, e.g. sarcasmo.

Finalmente, en cuanto al punto tres, los ejemplos de modelos para la predicción de legibilidad, la mayoría caen en regresión o clasificación y una excelente comparación de varios modelos de clasificación y regresión para predicción de Legibilidad, la encontramos en [13]. Otros estudios como el de [27], [32], [37] tratan a la predicción de Legibilidad como un problema de preferencias entre dos textos, de tal forma que predicen la dificultad relativa de un par de documentos. Nuestro proyecto de investigación construye Clasificadores/Comparadores de este tipo según el método reportado en [35].

### 1.1.2. Comprensibilidad

El segundo tema central en nuestra investigación es la Comprensibilidad. Definir comprensibilidad de un texto no es fácil y cae más sobre qué elementos necesita un texto para estar bien escrito y según [18] son los siguientes:

1. Los conceptos en el documento están conectados correctamente unos con otros.

2. Los conceptos en el texto están organizados de tal forma que múltiples ocurrencias de un concepto están posicionadas cercanamente entre si (característica de dispersión).

3. Los conceptos en un documento están organizados de tal forma que un concepto es primero explicado y luego este concepto es usado para explicar otro concepto (propiedad de secuencialidad).

Adoptamos estas propiedades cualitativas para considerar qué tan comprensible es un texto. En específico, a través de tres medidas ya propuestas en la literatura que son: Conectividad, Dispersión y Carga de Comprensibilidad.

## 1.2. Motivación

Una vez ya introducidos los conceptos y factores básicos de la investigación, pasaremos a presentar las motivaciones que llevan a realizar este proyecto de investigación. En primer lugar motivaremos la investigación en el área de lingüística computacional, continuando con minería de datos, evaluación de Legibilidad y finalmente en Comprensibilidad.

El éxito de la lingüística computacional se debe en parte a su capacidad de procesar rápidamente una gran cantidad de datos que provienen de la web. Esto ha llevado a la utilización de enfoques estadísticos que desplazan los enfoques tradicionales que involucraban modelos “hechos a mano” para su uso lingüístico. El éxito de estos nuevos modelos y la vasta cantidad de datos empíricos que éstos proveen, nos brinda una cantidad abundante de información para plantear nuevas preguntas.

Mientras que los algoritmos computacionales y modelos estadísticos están bien estudiados, el hacer que las computadoras adquieran una vasta información tanto lingüística como de conocimiento todavía continúa siendo un reto. Dicho de otra manera, lograr que una computadora aprenda de datos no es fácil y tiende a ser costoso. Los métodos de extracción de la información pueden adquirir información valiosa acerca del mundo y los retos actuales que enfrenta la lingüística computacional están íntimamente relacionados con el problema de representar el conocimiento humano del mundo y esto solamente se logra a través del estudio de cómo los humanos representan y comunican este conocimiento a través de lenguaje natural.

De aquí nace la motivación principal de estudiar escritos de tesis, ya que representan una parte fundamental de cómo comunicamos y aprendemos ciencia como estudiantes e investigadores. Si somos capaces de generar herramientas computacionales que logren facilitar esta comunicación, o a la tarea de favorecer la creación de este tipo de textos, estamos entonces desarrollando una investigación de importancia en un área que apoya a lectores, estudiantes, asesores/editores e investigadores en el área. Ahora, si nos interesa el análisis cualitativo de tesis de grado, la siguiente pregunta es cómo generar una evaluación de las mismas. De aquí que investigando en la literatura encontramos medidas lingüísticas para la Comprensibilidad y Legibilidad.

La elección de evaluar la Legibilidad está motivada con base en la naturaleza de su definición. Nos interesa, para una tesis de grado, la evaluación de su nivel académico a través de la dificultad de su lectura para así mejorarlo. Como siguiente punto, se elige esta característica debido a la información abundante que encontramos sobre ella, posibilitando realizar evaluaciones exitosas de la legibilidad en tesis a través de una metodología en el estado del arte. Al revisar la literatura, en general se reportan modelos que se basan tanto en regresión como en clasificación, nuestra investigación en cambio, utiliza un modelo de ranqueo de legibilidad. Esto es motivante debido a que no es un modelo que se utilice comúnmente pero ha demostrado buenos resultados, e.g. en [35]), así mismo y debido a que nuestro corpus de datos es limitado, utilizar modelos profundos de aprendizaje de máquina se vuelve complicado y este ranqueador se puede construir con los datos disponibles.

Con base en Comprensibilidad, las principales motivaciones para realizar nuestra investigación a través de este atributo lingüístico son tres. En primer lugar, es un área de estudio importante donde no hay tanta literatura, de aquí que realizar investigación en ella es relevante. En segundo lugar, gracias al trabajo de [29] y [18], pudimos identificar representaciones que con ciertas modificaciones se podían adaptar a nuestro tipo de documentos para así poder extraer información cuantitativa sobre la dificultad de estos textos. Así mismo, en la revisión del estado del arte, observamos que no se han obtenido ni representaciones ni extracciones de estas métricas de comprensibilidad para tesis de grado, de aquí que hacerlo se vuelve pertinente para la investigación en lingüística-computacional. Como última motivación, el mismo proceso de obtención de estas medidas puede ser utilizado para retroalimentar visualmente al autor del texto, son este tipo de herramientas que en un futuro pueden asistir tanto a estudiantes como a revisores en la corrección de una tesis en proceso.

## 1.3. A quién está dirigido

A continuación presentamos lo que nosotros consideramos son los principales beneficiarios de nuestro trabajo de investigación.

Como beneficiarios directos de esta investigación están estudiantes que se encuentren en el proceso de realización de una tesis de grado. Nuestro trabajo sienta bases importantes para que en un futuro, un estudiante pueda introducir su trabajo de tesis en una herramienta y ésta le regrese una evaluación del nivel académico de su tesis. Esta evaluación se realizaría con base en la Comprensibilidad y Legibilidad de la misma, así mismo se incluiría una retroalimentación que se basaría específicamente en la Comprensibilidad del texto y que abordaría posibles problemas o fortalezas que se observen en las medidas de Conectividad, Dispersión y Carga de Comprensibilidad en el texto.

Otro beneficiario directo es el asesor o editor de la tesis. Investigaciones como la presentada en este trabajo, sientan las bases para que en un futuro el asesor o editor pueda introducir el borrador de tesis de un estudiante en la herramienta y esta indique puntos relevantes que ameriten edición o trabajo en un borrador de tesis. El software señalaría problemáticas o fortalezas de Conectividad, Dispersión y Carga de Comprensibilidad para que el editor aborde éstas con el autor del texto. Todo esto ciertamente le ahorraría trabajo al asesor y lo apoyaría a concentrarse en avanzar la investigación.

Debido a que nuestro trabajo se centra sobre el área de Lingüística Computacional, investigadores relacionados con la temática pueden beneficiarse de nuestras contribuciones. Así mismo puede beneficiar a investigadores en Minería de Textos, Procesamiento de Lenguaje Natural y Aprendizaje de Máquina enfocado en clasificación de texto.

Finalmente, consideramos que los lectores de ciencia en general se pueden ver beneficiados de esta investigación. En este trabajo en específico, la experimentación se realiza sobre secciones de documentos de tesis pero el procesamiento del texto no está limitado a este tipo de documentos. Si bien es cierto que se tendrían que realizar ciertas modificaciones para incluir diferentes secciones, de diferentes tipos de texto, el corazón del proyecto yace en el análisis de medidas de Comprensibilidad y evaluaciones de Legibilidad sobre éstas. Las modificaciones no serían cuantiosas para poder incorporar diferentes documentos, como, por ejemplo, podrían ser artículos de investigación. Es por esto entonces que consideramos que investigación como la nuestra, genera apoyos en el área de comunicación de ciencia en general, esto debido a que un texto legible y comprensible es un texto que comunica de manera más efectiva información.

## 1.4. Problemática y Justificación

### 1.4.1. Problemática

La comunicación de conocimiento en la ciencia se basa en el intercambio de ideas a través de textos académicos, sin embargo, esta comunicación se puede ver afectada por diferentes factores que se reflejan en una baja Comprensibilidad y Legibilidad. Así mismo, la creación de estos textos no es una tarea sencilla para los autores debido a que se enfrentan a diversos obstáculos, que pueden llevar a la creación de un texto

vago o, en el peor de los casos, abandonar la escritura del mismo. Para evitar esto, los autores reciben apoyo de evaluadores, editores y asesores, sin embargo, este no es suficiente y hasta éstos se podrían beneficiar de recibir un texto de mejor calidad; permitiendo que los revisores se concentren en el contenido de la investigación en vez de aspectos técnicos del texto en proceso.

Los problemas mencionados anteriormente nos llevan a nuestra problemática técnica y que se abordó en este proyecto de investigación: El diseño de métodos y técnicas lingüístico-computacionales adecuadas para el análisis y mejora de un borrador de tesis.

### 1.4.2. Justificación

La problemática mencionada está, en esencia, relacionada con el lenguaje, así que es a través del procesamiento y análisis de lenguaje natural que debe y puede ser abordada. La Comprensibilidad, ya sea evaluarla o, en nuestro caso crear procesos para mejorarla, es de gran relevancia para poder impulsar el texto analizado. Relacionado con esto están las herramientas necesarias para poder mejorar la Comprensibilidad de un texto, éstas pudiendo ser aprovechadas tanto por revisores como por autores. Investigación como la encontrada en [29] y [30] es lo que permite la creación de este tipo de herramientas y es así como nuestra investigación apunta hacia la creación de técnicas lingüístico-computacionales que contribuyan a la realización de textos más comprensibles y legibles.

Aunado a las técnicas y herramientas que se crearán basados en Comprensibilidad, la Legibilidad es sustancial en nuestra investigación y en general para el desarrollo de herramientas de escritura para estudiantes. Particularmente investigamos un método de clasificación de Legibilidad que, contrario a literatura en general para evaluación de legibilidad, se basa en generar comparadores. Esto no se ha hecho antes para tesis en español, así que el simple hecho de crear y reportar estos resultados aporta al área de la Lingüística Computacional.

Finalmente, nuestro método es una combinación de dos evaluaciones, una para Legibilidad y otra para Comprensibilidad, incluyendo también una retroalimentación visual basada en el proceso de obtención de medidas de Comprensibilidad. Son estos tres factores en conjunto que crean un innovador proceso de ataque a un problema que es el de mejorar la comunicación de conocimiento en documentos científicos.

## 1.5. Objetivos

### 1.5.1. Objetivo General

El objetivo general de nuestro proyecto de investigación es la realización y valoración de métodos y técnicas lingüístico computacionales basadas en Comprensibilidad y Legibilidad para la evaluación y retroalimentación de documentos académicos en proceso.

### 1.5.2. Objetivos Específicos

Los objetivos en específico para lograr alcanzar nuestro objetivo general, son los siguientes:

**Comprensibilidad**

1. Generación y validación de resultados para las medidas de Comprensibilidad sobre los cuatro niveles académicos.

2. Generación de un proceso de retroalimentación basado en medidas de Comprensibilidad para un documento académico.

**Legibilidad**

1. Creación de clasificadores/comparadores a través de modelos de aprendizaje de máquina que sean capaces de diferenciar, entre dos documentos de texto, cuál es más legible que el otro.

2. Creación y selección de evaluadores capaces de evaluar el nivel académico de un documento en terminos de Legibilidad.

## 1.6. Contribuciones

Las principales contribuciones generadas por nuestro proyecto de investigación son las siguientes:

### 1.6.1. Comprensibilidad

1. Un proceso para evaluar la comprensibilidad basado en tres medidas validadas sobre una colección de textos escritos en español. Las medidas que evalúa este proceso son la Carga de Comprensibilidad, Conectividad y Dispersión y se realiza a través de los siguientes pasos para cada uno de los documentos:

   *a*) Extracción de las secciones

   *b*) División de secciones en párrafos

   *c*) División de párrafos en oraciones

   *d*) Etiquetado de partes de discurso

   *e*) Extracción de conceptos

   *f*) Creación del grafo de conceptos que incluye ocurrencias por párrafo, asignación del párrafo clave y relaciones entre conceptos.

2. Retroalimentación visual para la evaluación de escritos de tesis en proceso. Este proceso de retroalimentación utiliza el proceso de evaluación descrito anteriormente en conjunto con herramientas de visualización como lo son nubes de términos, grafos y gráficas de barras para proveer retroalimentación visual a un escrito en desarrollo. A través de este procedimiento somos capaces de observar las siguientes propiedades de un texto:

   *a*) Conceptos recabados y su frecuencia

   *b*) Las relaciones entre los conceptos que aparecen en el documento

   *c*) Relación entre conceptos y su párrafo clave

   *d*) Conceptos con mayor carga de comprensibilidad

*e*) Conceptos con mayor valor de conectividad

*f*) Párrafos con alta dispersión

*g*) El grafo de conceptos del documento

### 1.6.2. Legibilidad

1. La construcción y comparación de Clasificadores/Comparadores de legibilidad basados en documentos de tesis escritos en español para cuatro niveles académicos diferentes: Técnico Superior Universitario, Licenciatura, Maestría y Doctorado. Estos Clasificadores/Comparadores se construyeron con la utilización de aprendizaje de máquina y son capaces de evaluar, entre dos secciones de tesis diferentes, cuál de ellas es más o menos legible que la otra. Las secciones utilizadas para la construcción de estos Clasificadores/Comparadores son: Planteamiento del Problema, Resultados y Justificación.

2. La construcción y evaluación de evaluadores basados en los Clasificadores/Comparadores. Se construyeron cuatros diferentes: Aleatorio, que se basa en la selección de un referente aleatorio para comparación; Centroide, el cual utiliza un centroide como referente de evaluación; Máximo, que hace uso de un referente máximo para comparar y Mínimo, que compara a través del referente mínimo entre un conjunto de documentos. Estos evaluadores son capaces de recibir un documento de entrada y acotar tanto superior como inferiormente el nivel académico del mismo.

## 1.7. Organización del Documento

El presente trabajo de tesis se encuentra organizado de la siguiente forma:

- El capítulo dos incluye una revisión de los conceptos teóricos fundamentales para lenguaje, computación lingüística, Matemáticas Discretas, Estadística, Tesis de Grado que se consideran pertinentes para entender este trabajo de investigación.

- En el capítulo tres se presenta una descripción de los trabajos realizados en el área de comprensibilidad y legibilidad. Para cada uno de estos trabajos se analiza el problema abordado, el enfoque del mismo, resultados y metodología. Finalmente cada artículo se compara con nuestro trabajo de investigación y se señalan similitudes y diferencias.

- En el capítulo cuatro describimos el conjunto de datos con el que se trabajará, primero en aspectos generales, después a través del proceso de depuración que se llevó a cabo con el mismo, las secciones en específico con las que se trabajó y finalmente estadísticas obtenidas sobre las palabras en los documentos.

- El capítulo cinco presenta la metodología utilizada para el estudio de Comprensibilidad. En primer lugar se detalla la representación usada para los documentos y el segundo apartado se da la metodología para la obtención de las tres medidas de comprensibilidad: Carga de Comprensibilidad, Conectividad y Dispersión.

- El capítulo seis aborda los resultados para las medidas de Comprensibilidad. Primero se presentan las tablas de resultados para las tres secciones recabadas de cada tesis (Planteamiento del Problema, Justificación y Resultados) con su respectivo análisis por tabla así como análisis de los resultados en general. La segunda parte del capítulo muestra el sistema de retroalimentación visual basado en el proceso de obtención de las medidas de comprensibilidad. Se incluyen ejemplos así como análisis y explicación de éstos.

- El capítulo siete aborda la metodología para Legibilidad. Éste se divide en dos partes, la relacionada con el Comparador y la del Evaluador. En la primer parte se explica la representación de documentos y la forma en que se entrena al clasificador/comparador. La segunda parte explica la metodología para la construcción de los diferentes posibles evaluadores.

- El capítulo ocho muestra los resultados relacionados con el área de Legibilidad. El capítulo se divide en dos partes, la primera muestra los resultados de los clasificadores/comparadores en las tres posibles operaciones (Diferencia, División y Concatenación) así como el análisis de los mismos. La segunda parte del capítulo muestra los resultados para el Evaluador así como su análisis.

- Finalmente, el capítulo nueve incluye las conclusiones del trabajo de investigación y se divide en dos partes. La primera parte aborda nuestras conclusiones y trabajo a futuro en Comprensiblidad y la segunda parte aborda las conclusiones y trabajo a futuro en Legibilidad.

# Capítulo 2

# Marco Teórico

## 2.1. Aspectos del Lenguaje

### 2.1.1. Lenguaje

Según [5], el lenguaje se define como un sistema finito de elementos y principios que hacen posible a los hablantes construir oraciones para hacer trabajos comunicativos particulares. La parte del sistema que se encarga de producir e interpretar oraciones gramaticales se llama competencia gramatical, e incluye el conocimiento de los sonidos del habla que son parte de un lenguaje y cómo podrían o no encadenarse unos con otros. Así mismo, la competencia gramatical incluye también el saber los significados que se generan al usar diferentes secuencias de sonido en un lenguaje y cómo combinar esas unidades de significado en palabras, frases y oraciones. La competencia gramatical contribuye de forma similar a la comprensión en todos los lenguajes humanos.

En cuanto a la segunda parte de la definición, la que se refiere a la realización de trabajos comunicativos en particular, esta caería en ejemplos como lo serían la comunicación con otras personas. La competencia gramatical es, de hecho, inútil sin la competencia comunicativa. Actualmente el uso de lenguaje no está en las oraciones como tal, sino en unidades más largas o pequeñas de discurso. Para ser efectivos, los comunicadores tienen que combinar la competencia gramatical con el conocimiento de cómo "ocupar" las oraciones y otras piezas de estructura lingüística apropiadamente al propósito y contexto en mano. Estas dos, en conjunto, componen a la competencia comunicativa. De aquí que el conocimiento incluido en la competencia gramatical más la habilidad de utilizar ese conocimiento para lograr una amplia variedad de trabajos comunicativos, constituyen al lenguaje.

Existen diversos lenguajes y, cada uno de ellos difiere en diferentes aspectos, sin embargo, cada uno, en general, comparte las mismas características de diseño. El entender las características y propiedades que son universales para los lenguajes, así como las que no lo son, es el trabajo fundamental del lingüista. A continuación abordamos las propiedades universales del lenguaje, de acuerdo a [5]:

- **Modularidad**: Se cree y es aceptado en general que el lenguaje es un sistema modular, esto queriendo decir que cada individuo produce e interpreta el lenguaje utilizando un conjunto de subsistemas (o módulos) de forma coordinada. Cada módulo es responsable de una parte total del trabajo; éste toma la salida de otros módulos como entrada y distribuye su propia salida a otros

módulos. Hay módulos que son centrales para la lingüística: La fonética, que habla de la producción e interpretación de los sonidos del habla; la fonología estudia la organización de la fonética; la morfología estudia la estructura dentro de las palabras; la sintaxis nos habla de la estructura de las oraciones; el lexicón, es un repositorio de elementos lingüísticos con sus significados y propiedades estructurales; la semántica, recién introducida y definida como el análisis detallado del significado literal; y finalmente, el discurso, que se refiere a la organización del lenguaje más allá de la oración.

- **Circunscripción y Recursividad**: Todos los lenguajes están organizados a través de divisiones y esto permite que unidades más complejas surjan de unidades más simples. De tal forma que podemos decir "Ella se sentó", "La inteligente mujer se sentó", "La alta, de pelo negro e inteligente mujer de suéter rojo y collar negro se sentó." Esta capacidad del lenguaje da entrada también a la recursividad, la cual permite la aplicación de procesos gramaticales repetidamente. Por ejemplo, la recursión permite la expansión de una oración corta, en una larga como lo sería "Él era alto" en lugar "Él era alto y fuerte y atractivo y pensativo y..."

- **Discretización:** La discretización del lenguaje se puede observar en el significado, e.g. el espectro de color es continuo, sin embargo, todos los lenguajes dividen el color en categorías discretas que son representadas por colores que estos significan. Cada lenguaje está compuesto de sonidos separados, palabras, oraciones y otras unidades, el hecho de que escuchemos esta separación es un hito ya que, por ejemplo, al aprender un idioma completamente nuevo, en un inicio no logramos distinguir separaciones y conforme avanzamos en nuestro nivel logramos percibir estas separaciones y, por ende, su discretización.

- **Productividad**: La productividad se refiere a la capacidad de los lenguajes de "producir" nuevas palabras. Los lenguajes pueden, sistemáticamente, combinar las unidades mínimas de significado (morfemas), y formar palabras que originalmente no existían en el lenguaje. La mayoría de estas invenciones no llegarán a un diccionario, sin embargo, algunas sí logran introducirse a largo plazo.

- **Arbitrariedad**: La arbitrariedad se refiere a que, a pesar de pocas excepciones, las palabras no requieren un principio o conexión sistemática con lo que significan. Un ejemplo de esto lo encontramos en los números; donde en inglés son "one, two, three" y en mandarín "yi, er, san". En ninguno de los casos, existe una manera "correcta" de haberlo colocado de esta forma, de aquí que encontramos arbitrariedad al colocarle nombre a los números. Similarmente pasa para los sonidos, los patrones en los que los sonidos y las palabras son acomodados son arbitrarios.

- **Dependencia sobre el Contexto**: El hablar de la palabra "uno", refiriéndonos al número o el hablar de la palabra "uno" refiriéndonos al pronombre, puede prestarse a confusión. Este no es el caso, si, por ejemplo colocamos contexto, de tal forma que si yo escribo "El uno es uno de los símbolos más exquisitos.", aquí somos capaces de entender, debido al contexto, que el primer "uno" se

refiere a un símbolo que probablemente es un número y el segundo "uno"es un pronombre. El contexto es una propiedad crucial de los lenguajes.

- **Variabilidad**: Aunque todos los lenguajes comparten características universales, los lenguajes también son diferentes en distintas formas. El lenguaje varía desde aspectos tan pequeños como lo serían la persona que lo está comunicando hasta por zonas geográficas. Para tener una perspectiva de lo importante que es la variabilidad del lenguaje, sin ésta no habría lenguaje, de aquí que la variabilidad en el lenguaje permite a la gente comunicarse de una manera más intrínseca que el contenido semántico de las palabras y las oraciones que procesan. Es en esta variabilidad que los hablantes se basan para señalizar su identidad social, geográfica, estatus, etnicidad y hasta género.

### 2.1.2. Palabras

Las palabras son los bloques fundamentales que construyen un lenguaje, según [10] una palabra es la unidad mínima independiente del lenguaje, y que puede ser separada de otras unidades en un enunciado. Así mismo, las palabras usualmente son separadas por espacios cuando se escriben y cada una suele tener una distinción fonética. Esto es parcialmente preciso debido a que tanto hay palabras que pueden compartir fonética como hay lenguajes que no separan palabras, como el mandarín, sin embargo, en ese lenguaje uno es capaz de distinguir entre palabras. Las palabras pueden entrar en construcciones gramaticales, como lo serían frases y oraciones.

Otra posible definición para palabra, que considera aspectos gramaticales es la de [5] que la definen como un signo abstracto que es la unidad independiente gramatical más pequeña del lenguaje.

### 2.1.3. Oración

Para definir una oración partimos del hecho de que no encontramos una definición única aceptada en general. En nuestro trabajo de investigación adoptamos la definición de [28]. Ellos definen la oración como una estructura morfosintáctica, entendido como un conjunto de elementos y reglas para construir frases con sentido y carentes de ambigüedad, mediante la utilización de relaciones gramaticales, indexaciones y estructuras jerárquicas. En esta estructura es que en una lengua determinada se expresa un evento completo estructuralmente independiente. Su forma es de núcleo más elementos independientes, pudiendo estos últimos tener una estructura similar de forma recursiva.

### 2.1.4. Párrafo

Al igual que las concepciones lingüísticas ya mencionadas, definir el párrafo es complicado debido a sus características y a la falta de investigación en el tema. En este trabajo utilizamos una de las definiciones de [8], donde el párrafo corresponde al conjunto de enunciados que tienen una unidad tópica y expresan un sub-tópico del tema general del texto, ya que éste se desarrolla de forma secuencial. Así mismo, los párrafos son las divisiones que hace el hablante en su texto siguiendo criterios informativos, partes en las que se divide el tema sobre el que versa el texto y también según la superestructura, es decir, el tipo de texto.

## 2.2. Lingüística

La Lingüística es el estudio del lenguaje, según [20] . Esto abarca el análisis de las partes que lo integran y los métodos para modelarlo. La lingüística, en su mayoría, estudia signos y los signos en el lenguaje están constituidos de cuatro niveles diferentes: fonéticos, morfológicos, sintácticos y semánticos. La semántica se encarga del significado, la sintaxis de cómo las palabras se organizan para formar oraciones, la fonética establece cómo los lenguajes organizan sus sonidos y la morfología estudia la formación de palabras y su relación con otras palabras en el mismo lenguaje. Finalmente, la lingüística se estudia desde diferentes perspectivas, como lo podrían ser la psicológica, neuronal, histórica, etc. En nuestro trabajo de investigación, el paradigma que utilizaremos es el de lingüística computacional, que implementa construcciones teóricas para analizar gramaticalmente o producir lenguaje natural.

### 2.2.1. Lingüística Computacional

La lingüística computacional, de acuerdo con [5], es el estudio de sistemas computacionales para entender y generar el lenguaje. Los objetivos de investigación relacionados con lingüística computacional se desarrollaron primeramente con fines prácticos y estos podrían ser la traducción de texto a través de máquinas, la recuperación de información y las interfaces de hombre máquina. La última refiriéndose a sistemas que permitan comunicación entre un humano y la máquina, donde lo más natural sería comunicarnos con éstas a través de sistemas interactivos que se basen en lingüística computacional.

Así mismo, la lingüística computacional busca desarrollar la maquinaria computacional necesaria para que un agente exhiba varias formas de comportamiento lingüístico. Por "agente" nos referimos tanto a seres humanos como a agentes artificiales como lo podrían ser programas computacionales. Y por "maquinaria" nos referimos a los programas computacionales como tal, así como al conocimiento lingüístico que éstos deben poseer.

De lo anterior, entonces concluimos que la lingüística computacional involucra el diseño y desarrollo de programas que realicen tarea lingüísticas. Estos programas entonces deben de basarse en métodos desarrollados por científicos computacionales pero que utilizan el conocimiento lingüístico desarrollado por lingüistas. De aquí que la lingüística computacional integra ideas tanto de lingüística como de ciencias computacionales

### 2.2.2. Procesamiento de Lenguaje Natural (PLN)

La definición que utilizaremos en este proyecto de investigación para PLN la encontramos en [22].

**Definición 2.2.1.** PLN es una gama de técnicas computacionales motivada teóricamente, para analizar y representar textos que ocurren naturalmente en uno o más niveles de análisis lingüístico. Esto para lograr procesos de lenguaje parecidos a los de los humanos en una gama de tareas o aplicaciones.

Ciertamente hay elementos en la definición anterior que podrían ahondarse, como lo sería la idea de "gama de técnicas computacionales". Ésta parte del hecho

de que hay múltiples métodos o técnicas de dónde se escoge para lograr un tipo particular de análisis del lenguaje. Al hablar de "textos que ocurren naturalmente" nos referimos que pueden ser orales o escritos, el único requerimiento es que sea en un lenguaje utilizado por humanos y que el texto no sea específicamente para propósitos de la investigación como tal. La noción de "niveles de análisis lingüístico" se refiere al hecho de que hay múltiples tipos de procesos de lenguaje en acción cuando los humanos producen o comprenden lenguaje y, diferentes sistemas de PLN utilizan diferentes niveles. Cuando hablamos de "parecidos a los humanos", esencialmente estamos hablando de que PLN es considerado una disciplina de Inteligencia Artificial. Finalmente, "en una gama de tareas o aplicaciones" se refiere a que PLN normalmente no es la meta por si sola, para la mayoría es el medio para lograr una tarea en específico, como lo podría ser Recuperación de la Información.

### 2.2.3. Etiquetado de Partes del Discurso

De acuerdo con [19] el etiquetado de partes del discurso es el proceso de asignación de una parte del discurso (verbo, pronombre, sustantivo, etc...) a cada palabra en un texto. La entrada es una secuencia $x_1, x_2, ..., x_n$ de palabras (toquenizadas), y la salida es una secuencia $y_1, y_2, ..., y_n$ de etiquetas; cada salida $y_i$ corresponde exactamente a cada entrada $x_i$. Etiquetar es una tarea para desambiguar y como las palabras son ambiguas (tienen más de una posible parte del discurso), la tarea es encontrar la etiqueta correcta para cada respectiva situación. La exactitud de los algoritmos de etiquetado de partes del discurso (esto refiriéndose al porcentaje de etiquetas del conjunto prueba que coincide con las etiquetas valoradas por el panel de expertos) es alto, llegando a un 97 %, según [19].

En nuestro trabajo de investigación, utilizamos un etiquetado de partes del discurso NER ("Named Entity Recognition", por sus siglas en inglés) a través de las librerías de libre acceso de [14]. Este tipo de etiquetador, lleva su nombre, en primer lugar, debido a que usa entidades nombradas. Esto quiere decir que si en una oración aparece un nombre propio, el etiquetado aparte de indicar que es un sustantivo, indicaría que es un nombre propio. El término "nombrado de entidad" se puede extender a términos que no son necesariamente entidades, como lo podrían ser fechas, organizaciones, temporalidades, etc. Es importante aclarar que en nuestro proyecto de investigación no se utiliza esta característica del nombrado de entidades del algoritmo de etiquetado de partes del discurso.

## 2.3. Documentos Académicos

En este trabajo de investigación, analizamos tesis de grado de diferentes niveles académicos, a continuación incluimos las definiciones pertinentes para este tipo de documentos.

1. Tesis o Tesina Según [4] la **tesina** o **tesis de grado** de licenciatura, como suele denominarse en el ámbito de estudios superiores de nuestro país, del mismo modo que los trabajos finales en los ciclos de especialización, constituye el requisito académico final de la formación en esos niveles. Por ello, en la elaboración de ese documento el estudiante deberá demostrar la capacidad de

integración de los conocimientos adquiridos mediante la aplicación de instrumentos metodológicos probados en un asunto o tema en específico, este, claro está, vinculado con la carrera o especialización que ha seguido. Este documento puede desarrollarse como estudios exploratorios acerca de algún tema, o bien como la compilación ordenada de una problemática o su actualización en el ámbito de la disciplina.

2. La **Tesis de Maestría** exige al aspirante, por lo menos, que aplique, profundice y en lo posible amplíe los conocimientos propios del área de la mención de maestría; examine y plantee un problema teórico o práctico con cierto grado de creatividad y sepa responder a ese problema con un diseño de investigación apropiado y habitualmente consolidado por los paradigmas vigentes en el campo de referencia.

3. La **Tesis de Doctorado** reúne todos los requerimiento formales de una investigación científica, aunque acompañada por las exigencias de carácter académico propias de la institución en la cual el candidato al título de doctor debe defenderla. Contribuye al enriquecimiento significativo del saber y, especialmente, debe demostrar en lo posible la independencia de criterio del autor en relación con los paradigmas en uso, tanto metodológicos como teóricos. Es decir, conviene que reúna los requisitos normales de todo proyecto de investigación científica.

El trabajo de tesis es, a final de cuentas, un informe de investigación. Las exigencias para su realización varían dependiendo del tipo de titulación que se pretende.

### 2.3.1. Secciones Seleccionadas de la Tesis

En esta investigación analizamos tres secciones de tesis: Planteamiento del Problema, Justificación y Resultados. A continuación se detallan cada una de acuerdo a [5].

1. La sección de **Planteamiento del Problema** tiene el objetivo de la presentación del problema de una forma no-ambigua, tanto a nivel superficial como a detalle. Describir el problema en detalle engloba declarar las miras y objetivos, así como describir las limitaciones existentes y los supuestos que se han hecho respecto al problema.

2. La sección de tesis referente a **Justificación** debe mostrar por qué el problema de estudio es importante, argumentando el porqué debe ser estudiado.

3. Finalmente la sección de **Resultados** se refiere a la presentación de los datos obtenidos durante la investigación. Hay diversas formas de presentar estos datos y en [5] encontramos información relevante a cómo hacerlo. Una vez presentados los datos, la sección termina por incluir un breve análisis de los datos evaluados en contra de los objetivos del proyecto de investigación. El objetivo de este análisis es presentar los hallazgos de la investigación y la evaluación de la hipótesis con base en los datos obtenidos.

## 2.4. Entropía de Shannon

Este trabajo de investigación utiliza una definición del área conocida como teoría de la información. Según [2], Shannon asegura que una medida de la cantidad de información $H(p)$ contenida en una serie de eventos $p_1$ ... $p_N$ debe satisfacer tres requerimientos:

- $H$ debe ser continua en $p_i$;
- Si todos los $p_i$ son igualmente probables, esto es $p_i = 1/N$, entonces $H$ debe ser una función monotónicamente creciente de $N$;
- $H$ debe ser aditiva (preserva la operación de suma), es decir $H(x + y) = H(x) + H(y)$.

Shannon después demostró que la única $H$ que satisface estos tres requerimientos es

$$H(P) = -K \sum_{i=1}^{N} p_i \ln p_i \tag{2.1}$$

donde $K$ es una constante positiva. Esta cantidad se ha vuelto desde entonces conocida como la entropía de Shannon y ha sido utilizada en una variedad de aplicaciones, en nuestro caso, la hemos utilizado para obtener la Dispersión de un documento, la cual es una medida de Comprensibilidad. Esta medida nos otorga información relacionada con la densidad de información contenida en un párrafo.

## 2.5. Retroalimentación Visual

### 2.5.1. Nube de Términos

Las nubes de términos son representaciones gráficas de la frecuencia de un término. Estas representaciones dan mayor prominencia a términos que aparecen de forma más frecuente en un texto, de tal forma que si la palabra es más frecuente en el texto, ésta tiene mayor tamaño en la imagen producida en la nube de términos. Este tipo de asistencia visual puede ayudar en el análisis exploratorio al identificar términos que aparecen de manera más frecuente a lo largo de un conjunto de documentos. En 2.1 se ilustra una nube de términos.

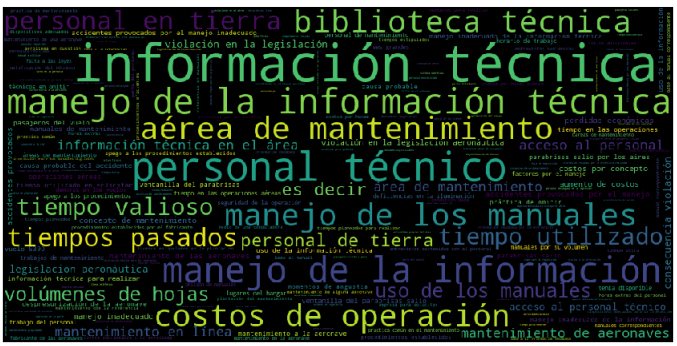



Figura 2.1: Ejemplo de una Nube de Términos

### 2.5.2. Teoría de Grafos

La Teoría de Grafos fue primeramente estudiada en 1735 por el matemático Leonhard Euler y hoy día es una herramienta poderosa y útil en varias áreas de la ciencia, que van desde la química, computación, lingüística, genética, ingeniería eléctrica, geografía, sociología, arquitectura y, claro está, las ciencias computacionales. Al mismo tiempo, la Teoría de Grafos ha emergido como una disciplina matemática por sí sola. En nuestra investigación utilizamos las definiciones de Teoría de Grafos dadas en [21].

**Definición 2.5.1.** Un **grafo** es un par ordenado $G = (V, E)$ que consiste de un conjunto no-vacío $V$ (que se llaman vértices o nodos) y un conjunto $E$ (que se llaman aristas) de subconjuntos de bi-elementos de $V$.

Con la definición de un grafo dada anteriormente, ahora compete explicar cómo se representa un grafo visualmente, ya que se utiliza en el proceso de retroalimentación visual basado en Comprensibilidad. Supongamos que se tiene el siguiente grafo:

$$g = (\{a, b, c, d\}, \{\{a, b\}, \{a, c\}, \{b, c\}, \{b, d\}, \{c, d\}\}) \tag{2.2}$$

Otra forma de representar dicho grafo de forma visual sería de la siguiente manera:

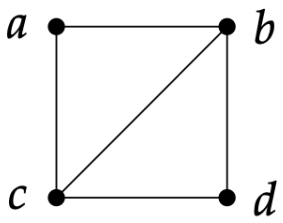



Figura 2.2: Ejemplo de grafo no dirigido

Así mismo, en esta investigación se utiliza un Grafo Ponderado, con la siguiente definición formal:

**Definición 2.5.2.** Un grafo ponderado (o grafo con pesos), es un grafo en el cual un número (el peso) es asignado en cada arista.

Los grafos ponderados aparecen en diversas circunstancias, e.g. en problemas donde se busca el camino más corto o en el famoso problema del agente viajero. Un ejemplo visual de un grafo ponderado es el siguiente:

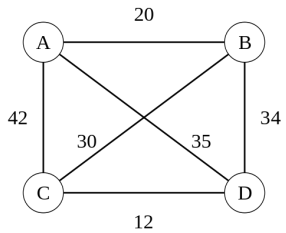



Figura 2.3: Ejemplo de grafo ponderado

### 2.5.3. Gráfica de Barras

Finalmente en el proceso de retroalimentación visual utilizamos una gráfica de barras, la cual es una forma de representar gráficamente un conjunto de datos o valores mediante barras rectangulares de longitud proporcional a los valores representados. A continuación incluimos una gráfica de barras.

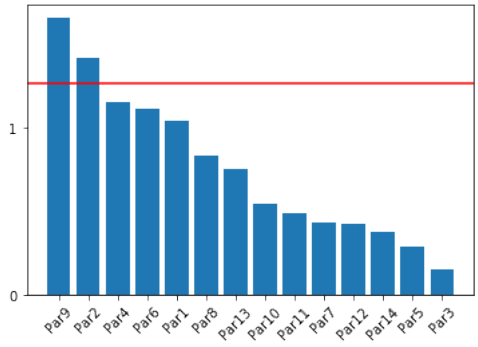



Figura 2.4: Ejemplo de gráfica de barras

## 2.6. Aprendizaje de Máquina

Este trabajo de investigación hace uso de aprendizaje de máquina para la realización de un clasificador/comparador capaz de identificar, entre dos textos de entrada, cuál de ellos es más o menos legible que el otro. De acuerdo con [34], la noción de **Aprendizaje de Máquina** refiere a un programa o sistema que entrena un modelo predictivo a partir de datos de entrada. El sistema usa el modelo aprendido para realizar predicciones útiles a partir de datos nuevos (nunca antes vistos) obtenidos de la misma distribución que la que se usó para entrenar al modelo. El aprendizaje automático o de máquina, también se conoce como el campo de estudio relacionado con estos programas o sistemas.

La primer aplicación de aprendizaje de máquina que se utilizó a gran escala, según [6] es el filtro para "spam" y si quisiéramos construir uno con técnicas de programación tradicionales lo haríamos de la siguiente forma:

1. Encontraríamos patrones en palabras específicas que se usan en correos de este tipo, nombre de destinatarios, cuerpo del correo y así.

2. Realizaríamos un algoritmo de detección de estos patrones y nuestro programa calificaría los correos electrónicos como "spam" si detectase alguno de ellos.

3. Probaríamos nuestro programa y repetiríamos los pasos anteriores hasta estar satisfechos.

Debido a que nuestro problema no es trivial, nuestro programa podría volverse una larga lista de reglas complejas y difíciles de mantener. Si en cambio, hubiésemos utilizado un filtro de "spam" basado en técnicas de aprendizaje de máquina, este hubiese aprendido cuáles palabras y frases son buenos para predecir el "spam" al detectar patrones frecuentes inusuales en los ejemplos de "spam" proveídos. Este programa sería corto, fácil de mantener y probablemente sería más preciso. En general, el aprendizaje de máquina es excelente para:

- Problemas donde las soluciones existentes requieren una amplia cantidad de afinación manual o una larga lista de reglas.
- Problemas complejos para los cuales no hay una buena solución al utilizar enfoques tradicionales.
- Problemas en los cuales el sistema tenga que adaptarse a nuevos datos.
- Para la obtención de perspectivas sobre problemas complejos y con una basta cantidad de datos.

Dentro del área de aprendizaje de máquina hay diferentes tipos de sistemas y es útil clasificarlos en categorías. Estas categorías no son exclusivas y puede haber sistemas donde hasta se combinen. Las principales categorías que se pueden encontrar son: aprendizaje supervisado, donde los datos de entrenamiento con el que se alimenta al sistema incluye las soluciones; aprendizaje no-supervisado, donde los datos de entrenamiento no están etiquetados; aprendizaje semi-supervisado, donde los datos de entrenamiento están parcialmente etiquetados; y aprendizaje de refuerzo, donde el sistema de entrenamiento (llamado agente) puede observar el ambiente, seleccionar y realizar acciones para obtener "recompensas" a cambio, es con base en éstas y una "política", que el sistema debe aprender por sí solo cuál es la mejor estrategia para obtener mayores "recompensas". En nuestro proyecto de investigación utilizamos aprendizaje de máquina supervisado con los siguiente modelos: Máquina de Soporte Vectorial Lineal, $K$-vecinos más cercanos y Perceptrón.

### 2.6.1. Máquina de Soporte Vectorial Lineal

Las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) se utilizan para detectar y explotar patrones complejos en datos al utilizar el agrupamiento. Son máquinas que aprenden y son usadas para realizar clasificaciones binarias y estimaciones de regresión. Comúnmente utilizan métodos basados en "kernels" para aplicar técnicas de clasificación lineal a problemas de clasificación no-lineales. Hay diferentes tipos de SVMs como lo podrían ser las lineales, polinomiales y sigmoidales; en este trabajo de investigación se ocupa la lineal.

La idea fundamental detrás de una SVM lineal es más sencillamente explicada a través de la Figura 2.5. En ésta podemos observar que contamos con dos clases de datos que son linealmente separables. La gráfica del lado izquierdo nos muestra posibles límites de decisión que podrían actuar como clasificadores lineales. El modelo que obtiene la línea de decisión punteada, es claramente el peor de los tres ya que no separa las clases apropiadamente, y los otros dos funcionan para separar las clases exitosamente. Ahora, pasando a la gráfica del lado derecho, podemos observar un límite de decisión de un clasificador SVM, la línea no solamente separa las dos clases pero también se coloca lo más lejos posible de la instancia de entrenamiento más cercana. Se puede pensar a un clasificador SVM como el encontrar la "calle" más ancha posible (ésta representada por las líneas punteadas) entre las clases: a esto se le llama el margen de clasificación amplio. Es importante notar aquí que el añadir más instancias de entrenamiento fuera de la "calle" no afectaría el límite de decisión, ya que éste está completamente determinado por las instancias localizadas en el borde de la "calle". Éstas instancias se llaman vectores de apoyo (están circuladas en rojo en la Figura 2.5) y de ahí viene el nombre de las SVMs.

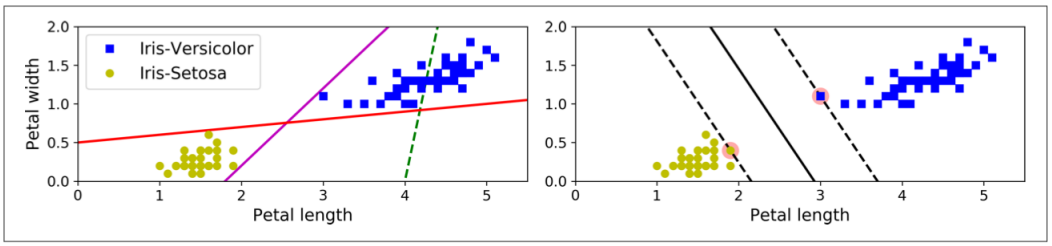



Figura 2.5: Ejemplo de clasificación a través de una SVM que aparece en [6]

Más información respecto a la formulación matemática de las SVMs se puede encontrar en [34] y no se incluye aquí debido a que está fuera del alcance de este trabajo de investigación.

### 2.6.2. $K$-vecinos más cercanos (KNN)

El algoritmo de los vecinos más cercanos pertenece a la clase de métodos estadísticos de reconocimiento de patrones. El método no supone algo a priori acerca de la distribución de la cual el modelado de la muestra de datos es tomado y requiere un entrenamiento con datos, tanto positivos como negativos de la muestra. Una nueva muestra es clasificada al calcular la distancia al vecino más cercano del caso de entrenamiento. El signo de ese punto determinara la clasificación de la muestra. En KNN los $K$ puntos más cercanos son considerados y el signo de la mayoría es utilizado para clasificar a la muestra.

El costo computacional de KNN se ve influenciado por tres factores:

1. La medida de distancia utilizada para localizar el vecino más cercano.

2. La regla de decisión utilizada para derivar la clasificación de los $k$ vecinos más cercanos.

3. El número de vecinos utilizados para clasificar la nueva muestra.

Es posible demostrar, al contrario de otros métodos, que este método es universalmente asintótico, i.e. mientras el número de datos de entrenamiento incrementa, si las observaciones son independientes y distribuidas idénticamente, no importa la distribución de dónde la muestra es tomada, la clase predicha convergerá con la asignación de clase que minimiza el error de clasificación errónea. En [34] se encuentran los detalles matemáticos.

### 2.6.3. Perceptrón

El perceptrón es un sistema ya sea de hardware o software, que toma de entrada uno o más valores, aplica una función en la suma de las entradas, y computa un solo valor de salida. En aprendizaje de máquina, la función típicamente es lineal, como podrían ser ReLU, Sigmoide o Tanh. Por ejemplo el siguiente perceptrón utiliza la función sigmoidal para procesar tres entradas.

$$f(x_1, x_2, x_3) = sigmoide(w_1x_1 + w_2x_2 + w_3x_3) \tag{2.3}$$

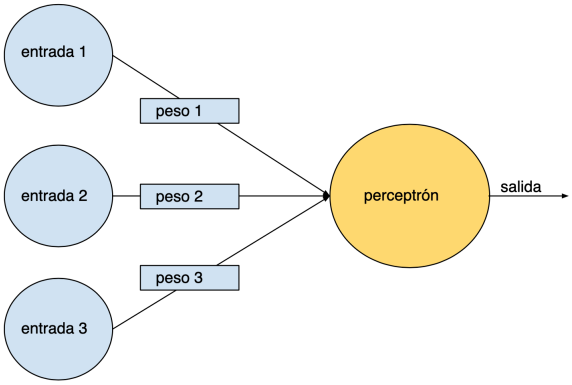



Figura 2.6: Ejemplo de perceptrón

La Figura 2.6, representa al perceptrón tomando tres entradas y cómo éstas se modifican por un peso antes de entrar al perceptrón como tal y generar una salida.

Para entrenar a un perceptrón se utilizan dos ideas: La primera consiste en aumentar el peso entre la conexión entre dos neuronas siempre y cuando tengan la misma salida. La segunda, es tomar en cuenta el error que se realiza por la red, de tal forma que se refuerzan las conexiones que ayudan a reducir el error. Específicamente, el perceptrón es alimentado una instancia a la vez y para cada instancia realiza predicciones. Para cada una de las neuronas de salida que se predijeron incorrectamente, se refuerzan los pesos que conectan las entradas que hubiesen contribuido a una predicción correcta. La regla la podemos observar en la siguiente ecuación,

$$w_{i,j}^{siguientepaso} = w_{i,j} + \eta(y_j - \vec{y}_j)x_i. \tag{2.4}$$

Donde $w_{i,j}$ es la conexión entre el peso $i$ y la neurona de salida $j$, $x_i$ es el $i$-ésimo valor de la correspondiente instancia de entrenamiento, $\vec{y}_j$ es la $j$-ésima neurona de salida para la instancia de entrenamiento correspondiente, $y_i$ es la salida requerida para la instancia de entrenamiento correspondiente y $\eta$ es la razón de entrenamiento. Es importante señalar que el límite de decisión para cada neurona es lineal, de aquí que los perceptrones son incapaces de aprender patrones complejos.

### 2.6.4. Exactitud

Las métricas de eficacia, son parte de cualquier modelos de aprendizaje de máquina. A través de ellas somos capaces de saber si se está logrando un progreso y, poner un número en éste. La exactitud o "accuracy" en inglés, es quizá la más sencilla de las métricas y se define como el número correcto de predicciones, dividido por el número total de predicciones, esto multiplicado por 100. También se le conoce como la fracción de predicciones que un modelo de clasificación tuvo correctamente.

En clasificación de múltiples clases la exactitud se define de la siguiente forma:

$$Exactitud = \frac{\text{predicciones correctas}}{\text{número total de ejemplos}} \tag{2.5}$$

En clasificación binaria, la exactitud tiene la siguiente definición:

$$Exactitud = \frac{\text{verdadero positivo} + \text{falso positivo}}{\text{número total de ejemplos}} \tag{2.6}$$

### 2.6.5. Evaluación Automática de Texto

De acuerdo con [3] actualmente nos encontramos en lo que se conoce como el enfoque de IA (Inteligencia Artificial) hacia Legibilidad. Estos nuevos enfoques, típicamente combinan una rica representación del texto a ser evaluado que se basa en la extracción de una variedad de características lingüísticas, con la combinación de sofisticados modelos predictivos que utilizan técnicas de aprendizaje de máquina. Estos avanzados modelos llegan a utilizar de docenas, a miles de características y son capaces de expresar sofisticados "espacios de decisión" que son excelentes para capturar las complejas interacciones entre las muchas variables que pueden caracterizar la dificultad de un documento para diferentes niveles de lectura o de lectores. Estos modelos, también han aumentado la exactitud y seguridad para las tareas específicas que fueron entrenados, a continuación hablamos en específico de cómo exactamente se trabaja la evaluación de la legibilidad como un problema de aprendizaje de máquina.

#### Evaluación de la Legibilidad a través de Aprendizaje de Máquina

Un problema típico de evaluación de la legibilidad, basado en un enfoque de aprendizaje de máquina consiste en tres pasos: En primer lugar se requiere de un conjunto de datos de entrenamiento con un etiquetado realizado través de un panel de expertos. En segundo lugar, se habla de definir un conjunto de características que serán extraídas de un texto; éstas deberán capturar características semánticas, sintácticas y otro atributos del texto que sobresalgan para la tarea de predicción de Legibilidad. Finalmente, se necesita un modelo de aprendizaje de máquina que aprenda a predecir el etiquetado generado por nuestro panel de expertos a través de las características extraídas de nuestro texto. Esto, en conjunto, es cómo se realiza la predicción de legibilidad a través de técnicas basadas en aprendizaje de máquina.

Concentremos nuestra atención un momento en la selección de un conjunto de datos de entrenamiento. Este conjunto de datos, en primer lugar debe estar constituido por textos individuales que sean representativos del género objetivo, lenguaje u otro aspecto de texto para el cual la tarea de evaluación de legibilidad sea necesaria. Cada texto en este conjunto de datos, debe tener un etiquetado de legibilidad realizado a través de un panel de expertos, aunque ha habido casos donde utilizar tecnologías como lo serían "crowdsourcing" han resultado útiles. Este etiquetado debe ser discreto o continuo, un etiquetado discreto puede ser la representación numérica de diferentes niveles académicos, como se maneja en nuestro proyecto de investigación. En específico, nuestro conjunto de datos se puede encontrar en el Capítulo 5 de nuestra tesis.

Una vez con el conjunto de datos, pasamos a la selección de características. Las categorías son amplias y se habló de ellas en el Capítulo 1, éstas son: léxico-semánticas, morfológicas, sintácticas, de dicurso, semánticas de alto nivel y pragmáticas. Uno tiene que seleccionar cuál o cuáles de estas categorías se seleccionarán para representar un documento de texto, y lo que se busca es que a través de estas características de representación se obtengan los atributos necesarios para la tarea en puerta. Como ejemplo, un modelo sencillo de predicción de legibilidad para lectores aprendiendo un segundo idioma, pudiese escoger la proporción de palabras no-familiares en un texto relativo a una lista de referencia. De esta forma, para cada texto se generarán vectores del tamaño de esta lista y si el texto contiene palabras

en esta lista, se tendrá un uno en la representación y cero en el caso contrario. Intuitivamente, podemos hablar de que vectores con una mayor cantidad de unos tienden a ser textos con una mayor dificultad de lectura para el aprendiz de un segundo idioma.

Una vez que las características que representarán a nuestros textos sean seleccionadas, pasamos al modelo de aprendizaje de máquina. En la mayoría de los casos, este modelo será simplemente una función que mapee texto a un valor numérico que representa la dificultad o el nivel académico. Dependiendo de la escala de medida de la variable de salida, la evaluación de legibilidad puede ser tomada como un problema de clasificación (valores discretos) o un problema de regresión (valores continuos). Aún así encontramos casos como lo sería [27], donde el problema de predicción es uno de comparación entre pares, de tal forma que entre dos textos el modelo predice cuál de ellos es más o menos legible que el otro. Nuestro proyecto de investigación utiliza este tipo de modelo al crear Clasificadores/Comparadores. Es entonces, nuestro conjunto de datos, las características seleccionadas para representar el texto y un modelo de aprendizaje máquina que se agrupan para lograr la evaluación automática de la Legibilidad de un texto.

# Capítulo 3

# Trabajo Relacionado

## 3.1. Comprensibilidad

A continuación presentaremos el trabajo relacionado para el área de Comprensibilidad. Específicamente, para cada uno de los trabajos que fueron base en este proyecto de investigación, se incluirá el problema que abordaron, cómo lo abordaron, sus resultados y su relación con nuestro proyecto. Los trabajos que se presentan a continuación tienen un factor común, en cada uno de ellos, pero de distintas formas, se aborda la problemática de evaluar la Comprensibilidad de un texto. En [23], esto se realiza para calificar traducciones; en [30], para evaluar traducciones intralinguales; en [17], para disminuir o aumentar valores en medidas de Comprensibilidad, en [29] se valoran "ebooks"; y en [38], se evalúan ensayos. Algunos de estos métodos de evaluación de Comprensibilidad sí fueron utilizados en nuestro trabajo de investigación, como sería el caso de [29] y [17], donde a través de medidas como la Conectividad, Dispersión y Carga de Comprensibilidad, se evalúan textos o libros digitales. Así mismo, trabajos como el de [30] y [29], nos invitaron a la generación de ideas para la creación de un proceso de evaluación basado en comprensibilidad que, al mismo tiempo de valorar un texto en proceso, también provee perspectivas sobre cómo mejorarlo. En los trabajos restantes ([38] y [23], no utilizamos los métodos en específico que ellos desarrollan, pero fuimos capaces tanto de validar la relevancia del estudio de la Comprensibilidad en textos, así como aterrizar la idea de la necesidad, como trabajo a futuro, de un procedimiento de evaluación para nuestro proceso de retroalimentación visual. A continuación, presentamos los artículos en lo que se basa nuestro trabajo de investigación.

La primera parte de este trabajo de investigación se centra en la creación de métodos y técnicas para obtener parámetros de evaluación basados en medidas de Comprensibilidad de textos académicos. Una vez con este proceso de obtención de parámetros de evaluación la idea es generar retroalimentación visual basada en el mismo. Finalmente tanto la evaluación como la retroalimentación proveerá un apoyo para los autores de escritos en desarrollo (o borradores).

Si deseamos mejorar la comprensibilidad de un texto, lo primero que podríamos pensar en realizar, sería tener una evaluación de su nivel. En el artículo [23], se crea un proyecto llamado ArisToCAT que aborda el problema de evaluar la Comprensibilidad de texto en "crudo" a través de sistemas de traducción automática. Este es un problema relevante ya que los sistemas de traducción son cada vez más frecuentes y evaluar su comprensibilidad de forma automática es de suma importancia. Entonces,

al igual que nosotros, en [23] se enfrentan al problema de generar una evaluación de comprensibilidad de textos y, a diferencia de nosotros, ellos buscan hacerlo por la evaluación en sí, mientras que nosotros buscamos utilizar tanto el proceso de evaluación como los valores generados por el proceso para la creación de parámetros de evaluación.

Para evaluar la calidad de estas traducciones, ellos comparan tres diferentes traductores que traducen Inglés-Holandés: un sistema neural comercial, uno basado en frases y otro basado en reglas. Para evaluar el texto traducido utilizan un proceso que anota los errores de los textos basado en dos pasos: el primer paso marca los errores de fluidez y en el segundo paso se marcan los errores de precisión; esto tanto para el texto fuente como para el texto traducido. Su proceso de evaluación de comprensibilidad difiere del nuestro ya que nosotros evaluamos con base a tres medidas de Comprensibilidad: Carga de Comprensibilidad (Secuencialidad), Dispersión y Conectividad; se podría decir que nuestro proceso más que basarse en errores en el texto para la evaluación de comprensibilidad, se basa en conexiones en el contenido conceptual.

Los resultados para su evaluación de comprensibilidad refieren al sistema neuronal comercial como el mejor entre los tres sistemas ya que presentó menos errores gramaticales. En cuanto a precisión, los resultados no fueron concluyentes. En comparación con nuestros resultados, nuestro proceso de Comprensibilidad obtiene parámetros de valoración a través de medidas, donde claramente algunas serán de mayor utilidad que otras, sin embargo, la investigación en [23] sienta las bases de cómo podríamos evaluar nuestro proceso de Comprensibilidad. Finalmente, en cuanto a las metodologías, claramente, tanto nuestro proyecto de investigación como [23] utilizan metodologías que no son equiparables, ya que nuestro interés yace más en el sentido de evaluación de la Comprensibilidad.

Otro trabajo de investigación [30] aborda la utilización de un revisor de lenguaje (RL) para semi-automatizar el proceso de una traducción intralingual, refiriéndose lo último al proceso de simplificar texto médico. Este proceso tiene un efecto sobre la comprensibilidad y legibilidad del texto, de aquí que una segunda tarea de este trabajo de investigación, es el medir el impacto del mismo.

El enfoque para atacar esta problemática se realizó, en primer lugar a través de la utilización de una herramienta que lleva por nombre Acrolinx, la cual se encarga de asistir a los realizadores de traducciones intralinguales; ellos tienen la oportunidad de utilizar las sugerencias de la herramienta o no, dependiendo de su criterio. La herramienta provee dos tipos de sugerencias: la primera es de tipo lingüista y la segunda de contenido didáctico. Así mismo se cuenta con tres diferentes conjuntos de datos, los que utilizan la herramienta, los que utilizan el viejo esquema y textos sin pasar por un proceso de traducción intralingual.

Para poder realizar una comparación de la legibilidad se utiliza una herramienta llamada Coh-Metrix, la cual es una herramienta computacional que automáticamente analiza la dificultad de un textoa lo largo de una variedad de dimensiones que incluyen: vocabulario, sintaxis y cohesión. Las medidas Coh-Metrix utilizadas por los investigadores son: narrativa, simplicidad sintáctica, qué tan concretas son las palabras, cohesión referencial, cohesión profunda y Flesch-Kincaid. Ahora, para medir la comprensibilidad, esto se realiza a través de un proceso de recuento de texto que consiste en leer un texto y luego escribir lo que uno recuerda del mismo. Esto se realizó al reclutar participantes de la Universidad de Arizona ajenos al experimento.

Los resultados obtenidos del proyecto de [30] indican que la utilización del RL influencia en algunas características de legibilidad pero no como tal en la parte de comprensibilidad, en específico se tiene una mejora en la simplicidad sintáctica. Aún así y aunque hubo mejoras significativas en legibilidad, no se encontraron indicios de mejoras en la comprensibilidad del texto. Esto puede ser debido a que aunque hay mejoras a nivel de palabras y aspectos sintácticos, el RL no genera una mejor cohesión, una característica que facilita el desarrollo de conexiones entre las ideas.

Sobre la relación con nuestra investigación, la problemática que abordamos ambos no dista mucho, ya que pretendemos disminuir la dificultad de un texto a través de la creación de herramientas lingüístico-computacionales que se basen en las medidas de Legibilidad y Comprensibilidad. Este artículo nos provee de ideas para, en un futuro, crear un software que genere tanto la evaluación como la retroalimentación para un borrador de tesis, de tal forma que el autor o revisor de este texto cuente con la opción de implementar o no los cambios sugeridos. Una vez implementado este software, la metodología que utilizan en este artículo, podría ser un excelente punto de partida hacia la realización de experimentos y obtención de resultados para medir la efectividad de nuestro esquema de retroalimentación. Un aspecto negativo de este artículo y que encontramos a lo largo de evaluaciones de Comprensibilidad, es que no tiene resultados significativos para ésta y, de hecho son las evaluaciones de Comprensibilidad donde menos encontramos resultados útiles o equiparables.

Otro artículo relacionado a nuestro trabajo es [38]. El problema que atacan es el utilizar los sistemas de puntuación automática de ensayos (AES por sus siglas en inglés) entrenados de forma neuronal para apoyar a los sistemas de evaluación automática de escritos (AWE por sus siglas en inglés), ya que los que son entrenados de forma neuronal no proveen representaciones de características útiles a los AWE. Nuestras problemáticas presentan similitudes dado que ambos debemos evaluar textos, ellos a través de AES y nosotros en términos de Comprensibilidad; así mismo la idea de los AWE para crear retroalimentación formativa va muy relacionada con uno de los principales objetivos de nuestra investigación, que es el crear retroalimentación visual, sin embargo, nosotros atacamos dicho problema a través del proceso de obtención de medidas.

[38] se centra en una tarea en específico de escritura, que es la de evaluación a respuestas de texto (RTA por sus siglas en inglés). Recientemente se construyó un sistema RTA AWE al extraer características relacionadas con el uso de Componentes Tópicos (CT) de un ensayo. Estos, sin embargo, fueron extraídos manualmente. Se realiza un método que utiliza la salida interpretable de las capas de atención del AES neuronal con la meta de extraer los CTs y utilizarlos para crear la RTA AWE. El enfoque de este trabajo va más hacia la utilización de estas capas intermedias para hacer la evaluación, sin embargo, no podemos ignorar la similitud entre la extracción de partes del texto para la realización de una evaluación con nuestro trabajo, esto debido a que extraemos conceptos de los documentos para con ellos calcular medidas de comprensibilidad. En un futuro nuestra investigación podría realizar la extracción de conceptos con otras ideas diferentes, como ésta.

Los resultados obtenidos en [38] son sumamente interesantes ya que, en primer lugar los CTs obtenidas por humanos y los obtenidos a través del método neuronal son iguales. Y en segundo lugar, los valores para las características basadas en rúbricas de los ensayos y en las CTs automáticas están correlacionados altamente con los resultados humanos. De aquí que logran generar la automatización completa para a

través de un AES, obtener un AWE y así crear retroalimentación formativa del texto, similar a lo que nuestro proyecto de investigación realiza; la retroalimentación, sin embargo con un método que difiere del nuestro.

Otro trabajo importante para el nuestro es [29]. La problemática que aborda está relacionada con "eBooks", específicamente se busca producir sugerencias para mejorarlos y una evaluación de su calidad. La presencia de este tipo de libros está en crecimiento y en un futuro cercano dominará el mercado; de aquí que "eBooks" de calidad son de suma importancia.

Para lograr su propósito, los autores primero extraen partes de los "eBooks", para despúes pasar a la extracción de conceptos. Con los conceptos obtienen una medida de cobertura que básicamente revisa si un concepto es abordado en el libro, i.e. si un libro tiene por título "Teoría de la Información", el concepto de información debe darse en el libro. Otra medida de importancia es la de Carga de Comprensibilidad, que se refiere a una medida de secuencialidad; la cual busca presentar conceptos en orden, de tal forma que para presentar ideas que requieren de conceptos base, primero se den los conceptos base y si no es así, se habla de que hay una Carga de Comprensibilidad sobre el lector. Finalmente, usan una medida de Legibilidad (Flesch-Kincaid) que estima la facilidad de comprensión del texto.

Los resultados que obtienen muestran que su método es capaz de evaluar cualitativamente los libros y esta evaluación puede ser utilizada para otro tipo de contenido digital. Como segunda idea, enfocada hacia la retroalimentación, extienden la posibilidad de un sistema de recomendación de aprendizaje, esto basado en el proceso de obtención de la carga de comprensibilidad.

Este trabajo y el nuestro tienen similitudes, la problemática que abordamos es semejante, sin embargo, nosotros buscamos la mejoría de textos académicos en proceso (borradores). En cuanto al acercamiento, compartimos la extracción de la Carga de Comprensibilidad, sin embargo, nosotros añadimos otras dos medidas de comprensibilidad como son la Conectividad y la Dispersión. Nuestro proyecto de investigación también realiza una evaluación basada en legibilidad, pero contrario al de ellos, la hacemos con base en la construcción de un comparador que después evalúa un borrador para obtener su Legibilidad. Otro punto donde compartimos ideas es en la realización de un proceso de retroalimentación visual, ellos no presentan el proceso pero sí proponen algunas posibles métricas para hacerlo. Probamos algunas de estas propuestas visuales y aunque algunas no son apropiadas para nuestra colección de documentos, otras sí fueron efectivas y se incorporaron a nuestro proceso final.

El último trabajo sobre Comprensibilidad, relacionado con nuestra investigación es [18]. Este trabajo se enfoca a la tarea de la creación de documentos de texto falsos para engañar a atacantes y así guiarlos hacia información inadecuada. Este documento tiene que ser creíble y menos comprensible para el atacante, de tal manera que llame su atención para buscar información relevante. Mientras más tiempo el atacante dedique a este documento, más recursos y esfuerzo gasta. La problemática entonces es la generación de estos documentos automáticamente, de tal forma que sean atractivos y creíblemente falsos.

El enfoque para la creación de estos documentos, ha sido el principal interés para explorar esta investigación. Los autores generan los documentos al modificar sus características de texto y así disminuir su comprensibilidad. Así, para poder realizar esta tarea, los autores deben ser capaces de medir exitosamente la comprensibilidad de los textos, así como manipular el texto de tal forma que ésta disminuya.

Los autores adoptan tres medidas para estimar la comprensibilidad de un documento: Conectividad (basada en teoría de grafos), Dispersión (utilizada en teoría de la información) y Secuencialidad (basada en principios psicolingüísticos). La creación de los documentos falsos se realiza a través de un algoritmo genético que en cada generación va disminuyendo la Secuencialidad y la Conectividad, y aumentando la Dispersión.

Los resultados para el trabajo de investigación son favorables ya que los autores son capaces de generar documentos falsos para una amplia clase de documentos legítimos, con una variedad de características de texto. Dado que el algoritmo genético funciona, el método desarrollado puede ayudar a mejorar los sistemas de engaño cibernético al proveer documentos de texto falsos difíciles de comprender que puedan engañar a atacantes.

La investigación realizada en [18] es de suma importancia para nuestro proyecto de investigación. Es a través de la misma que se seleccionan dos de las medidas de comprensibilidad por extraer (Conectividad y Dispersión), sin embargo, nuestro documentos son diferentes. Para poder llevar nuestros documentos a una representación adecuada se tuvo que realizar un PLN como el que aparece en el ejemplo de corrida del Capítulo 5. Con base en los resultados, éstos no son comparables con los nuestros debido a que ellos buscan como tal la disminución o el aumento de ciertas medidas de comprensibilidad, mientras que nosotros buscamos la obtención de rangos numéricos para poder evaluar borradores según su nivel académico. Así mismo, el artículo no presenta los rangos de los resultados obtenidos sino el porcentaje de cambio de los mismos y esto tiene sentido, ya que lo que buscaban era un cambio en la medida, ya sea positivo o negativo.

## 3.2. Legibilidad

A continuación presentamos el trabajo relacionado con nuestro proyecto de investigación para el área de Legibilidad. Al igual que con Comprensibilidad, para cada artículo que presentamos, se incluye el problema que éste abordó, cómo se abordó, sus resultados y una comparación con nuestro proyecto de investigación. El primer artículo que se presenta, es [3], y para nosotros funcionó como una introducción al área de evaluación automática de Legibilidad; a través de su investigación logramos generar fuertes puntos de partida para iniciar el proyecto de investigación. Después de esto, se estudió a [7], el cual es un trabajo híbrido entre evaluación de Legibilidad y Comprensibilidad, el trabajo sentó bases importantes sobre qué necesitaba poseer nuestro proceso de evaluación de Legibilidad y Comprensibilidad. Finalmente, y debido a que nuestro trabajo de investigación utiliza técnicas de ordenamiento de texto donde para pares de documentos se obtiene el más legible entre los dos, tanto [27], [35], [37] y [32] nos presentan trabajos donde la evaluación de Legibilidad se centra en esta técnica. [27] analiza características relevantes para generar ranqueadores de Legibilidad efectivos y, a través de ese trabajo validamos la utilización de la representación seleccionada para nuestros documentos de texto. Tanto [37] como [32], incorporan a nuestro proyecto una validez necesaria y actual sobre la utilización de ranqueadores como una técnica de evaluación en Legibilidad para el ordenamiento de textos. Finalmente, en [35], encontramos la base metodológica para crear estos ranqueadores, para así después construir evaluadores con ellos.

El primer artículo a discutir se utiliza en el área de Legibilidad es [3]. Es un

artículo central, ya que provee antecedentes sobre la evaluación automática de legibilidad, revisa el estado del arte de los algoritmos para modelado automático y predicción de legibilidad y finalmente propone retos y oportunidades para futura exploración en el área.

El enfoque del artículo es completamente introductorio hacia el área de la evaluación de la Legibilidad de textos. De tal forma que comienza por presentarnos la problemática así como la relevancia y primeros intentos de ataque hacia este problema. Después visita superficialmente los métodos automáticos para evaluar la legibilidad de textos, partiendo del paradigma de la clasificación de texto como un problema aprendizaje máquina, las características posibles en la representación del texto, los modelos de aprendizaje máquina y las evaluaciones, medidas y resultados que éstos obtienen.

El artículo concluye convenciéndonos que la evaluación de la legibilidad del texto es un área importante y de continuo crecimiento. Así mismo, nos indica hacia dónde se dirige la investigación en el área y nos invita a enfocarnos en ciertas áreas de oportunidad actuales.

En cuanto a la relación de este artículo con nuestro trabajo de investigación es claro que se tomó como una introducción a lo que es el área de evaluación de la legibilidad de textos. Se nos presentó un amplio panorama del cual, nuestra investigación en específico tomó partes pequeñas para a partir de ellas, construir el proyecto de investigación. Ejemplo de éstas podrían ser la utilización de aprendizaje máquina para clasificar, así como la variabilidad de posibilidades que existe para representar nuestros documentos de texto y las características lingüísticas para hacerlo; esto lo podemos ver en los resultados de nuestros comparadores así como los experimentos realizados para los mismos.

El siguiente artículo que tocaremos es [7], en él se aborda la problemática de generar una validación automática de los requerimientos en la creación de un documento de información del paciente (IP). Estos documentos contienen toda la información relevante del proceso al que se someterán los pacientes, en relación con un ensayo clínico. Es fácil ver que es de suma importancia que este tipo de documento sea entendible, ya que así el paciente es capaz de tomar una decisión más informada y posee la capacidad de tomar una decisión veraz sobre lo que acontecerá en su cuerpo. El crear entonces una validación en automático del IP, es una ayuda al paciente pero también a los que aplican el documento, ya que estas herramientas inteligentes proveen una asistencia en la escritura del documento y en el proceso de la aplicación del IP.

El enfoque que proponen los autores del artículo para abordar el problema anterior, es la creación de un sistema automático que revise tanto la legibilidad como la comprensibilidad del IP. Una vez con este sistema montado, se invita a los creadores de documentos IP a realizar documentos más entendibles a través de sugerencias que teóricamente mejoran la calidad de éste.

La metodología que ellos siguen para cumplir con esta tarea comienza por un pre-procesamiento del lenguaje natural del documento. A esto le sigue el análisis de componentes métricos de legibilidad como lo serían Flesch-Reading-Ease, gSMOG, etc. Una vez con esta información, continúan a la tarea de detectar términos técnicos y analizar si han sido o no explicados claramente; para realizar esto se utiliza un sistema de reconocimiento de entidades nombradas. Debido a que no todas las entidades nombradas son ajenas a los pacientes, un análisis de familiaridad de las

mismas es necesario y esto se logra a través del análisis de la ocurrencia de la palabra en un corpus de noticias. Es con base en el resultado de qué tan común o no es la palabra que se decide si se queda en el documento IP o no.

El detectar si un término técnico ha sido o no abordado se realiza siguiendo reglas de expresiones regulares. Esto quiere decir que las explicaciones para ciertos términos técnicos tienden a estar, por ejemplo, entre paréntesis, de aquí que podemos identificar a través de expresiones regulares si un término ha sido o no explicado.

El artículo presentado, como tal no tiene resultados ya que no implementan el método propuesto y, por ende, no hay una experimentación sobre el sistema. En cuanto a la problemática que abordan, es similar a la nuestra ya que a través de la evaluación de la Comprensibilidad y la Legibilidad, los autores pretenden generar un sistema automático para la mejoría de documentos de IP. Su metodología tiene similitudes con la nuestra; una de ellas es la utilización de familiaridad de palabras, ya que en nuestro trabajo de investigación utilizamos la familiariadad como característica de representación en nuestros textos. En cuanto al uso de expresiones regulares, nosotros también utilizamos esta idea pero para la identificación de conceptos en el área de Comprensibilidad del proyecto de investigación. Hubiese sido interesante observar la implementación del sistema para poder realizar comparaciones y así obtener ventajas y desventajas de su sistema y el nuestro, así mismo para obtener ideas de posibles implementaciones de nuestro proyecto de investigación ya que como trabajo a futuro nos interesa generar un software que evalúe y retroalimente un documento de tesis en proceso basado en medidas de Comprensibilidad y Legibilidad.

Pasamos ahora a analizar el artículo que mayor relación tiene sobre nuestro trabajo de investigación en el área de Legibilidad. La problemática que abordan en [35] es el tener que evaluar la legibilidad de textos y así clasificarlos entre 12 niveles diferentes. Otro reto al que los autores de este trabajo de investigación se enfrentan es la falta de datos de entrenamiento para la realización de esta tarea, de aquí que la utilización de métodos tradicionales como los que aparecen en [3], no brindan la solución.

Para atacar este problema ellos se enfocan en realizar la tarea a través de un ordenamiento de los documentos de texto. De tal forma que primero generan un comparador al entrenar una SVM. Este comparador es capaz de juzgar la legibilidad relativa entre dos textos, i.e. es capaz de decidir cuál de los dos textos es más difícil. Finalmente los textos son ordenados con la utilización del clasificador/comparador y un algoritmo de ordenamiento.

El clasificador/comparador que utilizan en el trabajo de investigación se construye a través de aprendizaje máquina. Dados dos textos $a, b \in S$, donde $S$ es un conjunto de textos, los vectores de características $V_a$ y $V_b$ que corresponden a cada documento se construyen. Una vez con estos vectores se aplica un operador $\circ$ y $V_{ab}$ es construido por la operación $V_a \circ V_b$. Cuando $V_{ab}$ se introduce al comparador, este regresa 1 si $a > b$ y $-1$ cuando $a < b$. En cuanto a los términos de las características que representan a los vectores, se utilizan características Locales (frecuencia) y Globales (familiaridad). Los operadores que se manejan son diferencia, división y concatenación. Finalmente, y una vez que se tiene el comparador construido, se ordenan los textos a través del algoritmo de inserción binaria, esto debido a que tiene la búsqueda más veloz.

El trabajo realizado por los autores logra construir un clasificador/comparador

con una precisión por encima del 90 % y con base en este comparador son capaces de ordenar los textos. Este tipo de modelo es capaz de resolver el problema de categorizar datos en múltiples clases, esto recurriendo solamente a dos clases de datos de entrenamiento. Así mismo, los autores presentan una aplicación construida con base en este sistema que se llama Terrace; este siendo un sistema que recupera un texto de nivel de legibilidad similar, de un conjunto de textos una vez que ha recibido un texto de entrada.

Las similitudes entre nuestro trabajo y el de ellos son bastantes. En primer lugar nuestro trabajo de investigación construye el clasificador/comparador basándose en este artículo. Esto se hizo así por dos razones: la primera es que se cuenta con pocos datos y la segunda es debido a que se pretende aprovechar el hecho de que solamente necesitamos dos clases de datos de entrenamiento para la construcción de esta máquina. Con esta base pudimos realizar experimentos con diferentes combinaciones de clases de datos de entrenamiento y así seleccionar el clasificador/comparador con mejores resultados.

Algo que difiere en nuestros trabajos de investigación es el qué se hace una vez que se tiene construido el clasificador/comparador. Ellos utilizan el clasificador/comparador para ordenar los textos académicos, nosotros, en cambio, utilizamos esta máquina para acotar el nivel académico de un texto con base en su Legbilidad. Es importante mencionar que ellos llegan a construir un comparador con una precisión por arriba del 90 % mientras que el mejor comparador que logramos construir nosotros llega apenas al 90 %, esto lo culpamos a la diferencia en el tipo de documentos, es posible ver dentro de sus resultados que mientras subimos la calidad de un escrito, es más difícil distinguir su Legibilidad. Nuestros documentos, aunque de diferentes niveles académicos, son todos de niveles académicos altos, mientras que los de ellos van desde un nivel de aprendizaje de inglés o japonés cero, hasta un nivel doce.

El siguiente artículo que llamó la atención en nuestro trabajo de investigación es [27]. La problemática que ellos abordan es el combinar características léxicas, sintácticas y de discurso para con ellas producir un modelo capaz de predecir la legibilidad de un texto. Una vez realizado esto, se dan a la tarea de comparar estas características para poder así observar cuál de ellas cumplen mejor la función de predecir la legibilidad de un texto al nivel de cómo lo juzgaría un humano. El artículo todavía va más allá y compara si algunas características son específicamente útiles para predecir la legibilidad de un texto, para ranquear o para las dos.

Para realizar esta tarea los autores utilizan textos del periódico norteamericano "Wall Street Journal". Particularmente se seleccionan treinta textos de forma aleatoria y para cada uno de estos textos se analizan factores de legibilidad que incluyen medidas base (como lo podrían ser la longitud del artículo), vocabulario, sintaxis, cohesión, coherencia de identidades y de discurso. Para cada artículo se extraen estos factores y se estudia tanto su capacidad para predecir, como para para ranquear legibilidad. Finalmente, también se analiza la asociación entre estos factores de legibilidad al realizar las combinaciones posibles entre ellos, esto para poder ver si la combinación de factores genera mejores resultados.

Entre sus resultados, los investigadores encuentran una alta correlación que los factores considerados capaces de predecir legibilidad de un texto son las relaciones de discurso, el promedio de frases verbales, características basadas en vocabulario y el número de palabras; es importante enfatizar aquí que los factores de legibilidad de

vocabulario y de relaciones de discurso son los predictores más fuertes, seguidos por el promedio de frases verbales y la longitud de texto. Con base en la combinación de factores de legibilidad como predictores de texto, los investigadores comentan que los resultados son complicados de interpretar, sin embargo, las relaciones de discurso, características de vocabulario y la longitud, son factores importantes a lo largo de combinaciones con mejores resultados. Finalmente, y con base en el ranqueo de legibilidad (Esto quiere decir que en lugar de intentar predecir legibilidad de un solo documento, se consideran dos y se selecciona cuál de ellos es mejor.), los mejores resultados se obtienen a través coherencia de entidades y características de discurso y sintaxis.

Es claro entonces cómo el trabajo de investigación presentado tiene similitudes con nuestro proyecto de investigación. En primer lugar, se analiza el problema de ranqueo de un texto, esto quiere decir que en lugar de predecir la legibilidad de un texto, se consideran pares de documentos y entre ellos se analiza el más o menos legible entre los dos. En segundo lugar, los investigadores analizan características de vocabulario y su correlación como factores predictivos efectivos en legibilidad; en nuestro trabajo de investigación utilizamos de hecho características basadas en vocabulario para representar nuestros textos. El hecho de que sus resultados muestren a estas características como factores predicitivos efectivos de ranqueo en legibilidad, valida la utilización de este tipo de características. Es un trabajo que ciertamente aportó cuerpo a nuestro proyecto de investigación.

El siguiente artículo que le compete a este trabajo de investigación es [37]. Este trabajo se enfoca a la tarea de la evaluación de la legibilidad para textos que se enfocan en la enseñanza de un segundo idioma. Esto debido a que un sistema específicamente diseñado hacia la percepción de la dificultad de lectura de textos para segundo idioma puede producir una estimación más exacta de la dificultad de lectura para un lector no-nativo de la lengua y así facilitar su aprendizaje. Este tipo de trabajos, se enfrentan también a un segundo reto: la falta de datos con un etiquetado a través de expertos.

Para abordar el problema los investigadores desarrollan un sistema que produce una estimación de legibilidad en el estado-del-arte a través de la exploración de una gama de factores de legibilidad y el análisis de su poder predictivo. En particular, comparan un modelo basado en ranqueo con uno de clasificación tradicional. Ellos creen que el problema de la dificultad de texto es una variable continua más que una discreta, de aquí que en lugar de asignarle un valor absoluto a un nivel de texto, es más natural la predicción de la dificultad de un texto a través de la comparación con otros textos.

Los resultados obtenidos con base en entrenamientos realizados con datos de textos no específicos para estudiantes de un segundo idioma son interesantes. Primero que nada observan que los mejores resultados se encuentran con la combinación de todas las características, algo que era de esperarse. En segundo lugar y basándose en la comparación entre el modelo basado en ranqueo y el modelo basado en clasificación, se hace notar que los dos modelos varían a lo largo de las diferentes características y ninguno de los dos modelos es consistentemente mejor que el otro. Finalmente, cuando se combinan todas las características, el modelo de clasificación obtiene mejores resultados que el de ranqueo, de aquí que el modelo de ranqueo no es necesariamente el mejor modelo para la tarea de predicción de legibilidad.

En el trabajo de investigación se hizo una segunda ronda de experimentos, pero

ahora con datos (textos) en específico para estudiantes de un segundo idioma. En este caso el modelo de clasificación obtuvo resultados negativos y el modelo de ranqueo se mantuvo constante. Esto mostró que el modelo de ranqueo generaliza mejor y esto se debe a que no trata de generar un empate entre los documentos hacia uno de los niveles, sino que captura la dificultad relativa entre los documentos.

El previo trabajo fue de interés para nuestro proyecto de investigación por diferentes aspectos. En primer lugar, crea un modelo de evaluación de legibilidad en el estado del arte a través de ranqueo de forma similar a nuestra investigación. Después, compara este modelo con uno tradicional de clasificación y, aunque no obtiene resultados para indicar que el de ranqueo es mejor, obtiene resultados que parecen indicar que considerar la clasificación de texto como un problema continuo generaliza de mejor manera. Así mismo los resultados cercanos entre el modelo de ranqueo y el de clasificación nos indica que los modelos de ranqueo son útiles y merecen su exploración. Finalmente los resultados para las características nos muestran que la combinación de las mismas da los mejores resultados, tanto para el modelo de ranqueo como al de clasificación tradicional, esto abre a que en algún trabajo a futuro podamos incorporar múltiples características para refinar nuestro modelo basado en ranqueo.

A continuación se presenta el último artículo de investigación que llamó la atención para nuestro proyecto de investigación. En [32], se dan a la tarea de predecir la dificultad relativa de oraciones a lo largo de un conjunto de oraciones. Este es un problema interesante y no ha recibido tanta atención ya que en general se predice la dificultad de un texto en su totalidad, pero al hablar de oraciones, la literatura es escasa. El problema, también es complicado debido a que las oraciones proveen menos datos que los que son requeridos para estimar de forma confiable la dificultad de éstas, particularmente para medidas que dependen en estadística agregada.

Para lograr resolver el problema propuesto, los investigadores primero recolectaron datos de “American National Corpus”, “New York Times” y “North American News Text Corpus”, estos datos se etiquetaron utilizando “crowdsourcing” con dos experimentos. El primer experimento consistía en proporcionar dos oraciones y un sujeto tenía que determinar cuál de ellas era más difícil que la otra, el segundo era similar, solamente que las dos oraciones estaban en su respectivo contexto. En cuanto a la generación del ranqueo, se utilizó el algoritmo “Trueskill”, el cual es un algoritmo de ranqueo a través de decisiones por parejas (como lo serían las oraciones). Finalmente, y a través de aprendizaje de máquina, se genera un proceso para predecir entre pares de oraciones o pares de oraciones con su contexto, cuál de ellas es más complicada que la otra, así mismo en medio de estos dos procesos, se incluye un proceso de selección de características a través del sistema “Oracle”.

Los resultados que se obtuvieron para la predicción de dificultad por pares encontraron que el modelo que mejor predecía la dificultad utilizaba todas las características, esto tanto para oraciones sin contexto como para oraciones con contexto. Aún así, representaciones más sencillas basadas en características léxicas, se acercaban a los resultados obtenidos por todas las características; de aquí que se logró la construcción de un modelo capaz de predecir la dificultad relativa entre oraciones con éxito. En cuanto a los resultados para el ranqueo de las oraciones, se encontraron diferencias pequeñas pero significativas entre oraciones solas o con contexto y este resultado sugiere que las propiedades que rodean a una oración pueden, de hecho afectar la dificultad de una oración.

Este artículo y nuestro proyecto de investigación ciertamente tienen similitudes. En primer lugar, una parte del proyecto construye ranqueadores de legibilidad que, entre dos oraciones o dos oraciones acompañadas de su contexto, predicen cuál de ellas es más o menos legible. Nuestro proyecto de investigación realiza lo mismo, solamente que se comparan dos secciones de tesis, de aquí que el método utilizado para comparar dos oraciones acompañadas de su contexto es de interés en nuestro trabajo. Otro aspecto importante que vemos en este proyecto de investigación, es la conclusión que las características léxicas predicen la legibilidad con efectividad, esto valida la utilización de características léxicas en nuestro trabajo. Finalmente, invita, como trabajo a futuro, utilizar los modelos de aprendizaje de máquina que se utilizan en este proyecto, ya que así podríamos comparar resultados con el modelo que utilizamos y que aparece en [35].

# Capítulo 4

# Metodología en Comprensibilidad

A continuación presentamos la metodología utilizada para la obtención de nuestras medidas de Comprensibilidad. El capítulo está organizado de tal forma que primero se presenta la representación de los documentos. A continuación se aclara la metodología precisa para la obtención de cada una de nuestras medidas de Comprensibilidad: Carga de Comprensibilidad, Conectividad y Dispersión. Finalmente, el capítulo cierra con la explicación del proceso de retroalimentación visual basado en Comprensibilidad.

## 4.1. Representación

A continuación abordaremos la representación que se utilizó para el manejo de los documentos. Una vez que se tiene dicha representación uno es capaz de extraer las medidas de Comprensibilidad.

### 4.1.1. Secuencias de Oraciones y Conceptos

En primer lugar presentaremos la dinámica de representación y extracción tanto de las secuencias de oraciones como para la obtención de los conceptos.

Un documento $d$ consiste en una secuencia finita de $K$ párrafos: $d = (p_1, ..., p_K)$. El $k$-ésimo párrafo está compuesto de una secuencia finita de oraciones $n_k$, de tal forma que $p_k = (s_1, ..., s_{n_k})$. La $j$-ésima oración del párrafo $p_k$ consiste en una secuencia finita de $m_{(j,k)}$ de palabras, tal que $s_{j,k} = (w_{(1,j,k)}, ..., w_{m_{j,k},j,k})$.

Una palabra $w_{(i,j,k)}$ o una frase (conjunto de palabras) en una oración se define como concepto $c$ si pertenece a un patrón de nuestras frases nominales [1] predominantes en el lenguaje. Identificamos un concepto al extraer el vocabulario de $d$ y etiquetar cada palabra a través de la identificación de su parte del discurso con spaCy [2].

Finalmente, un conjunto finito de conceptos en $d$ es representado por $C$ y la secuencia finita de $X$ ocurrencias de los conceptos como $O = (c_1, ..., c_X)$, donde $c_x \in C$ (Aclaramos aquí que al hablar de una ocurrencia nos referimos a la aparición del concepto a lo largo del documento, de aquí que $O$ es un conjunto de conceptos ordenado a través de la aparición de éstos en el documento.) Así mismo, un concepto

[1] Conjunto de frases nominales que de acuerdo con [26] regresa un concepto: N-N, N-A, N-[A]-Prep-N-[A], N-[A]-Prep-Art-N-[A], N-[A]-Prep-V-[N-[A]].

[2] Una librería libre de Python para Procesamiento de Lenguaje Natural avanzado [14]

$c' \in C$ se llama $concepto - relacionado$ de $c$ si co-ocurre en más de un párrafo $p_k$, al hablar de co-ocurrir nos referimos a que tanto $c$ como $c'$ ocurren en un mismo $p_k$. Para modelar esto, se realiza un diccionario donde aparecen los párrafos que se encuentran en la sección con los respectivos conceptos que aparecen en dicho párrafo; en la Figura 4.1 es posible observar este arreglo. Por otro lado, el conjunto de conceptos que se relacionan con un $c$ lo denotaremos por $R(c)$ y a la cantidad de veces que un $c$ aparece en un párrafo (frecuencia) como $f(c, p_k)$. Por último, cada concepto posee su párrafo clave $párrafo - clave$ $p_{k_c}$, i.e. el párrafo donde teóricamente dicho concepto es explicado.

### 4.1.2. Grafo de Conceptos

Las medidas de Conectividad y Dispersión son obtenidas a través de la creación de un Grafo de Conceptos. Para obtener este grafo, un documento se representa como un grafo unidireccional de conceptos $G = (C, E, w)$, donde $C$ es el conjunto finito de los nodos concepto, $E \subseteq C \times C$ es el conjunto finito de aristas entre los conceptos y $w : E \to \mathcal{N}$ es una función del peso que se le asigna a cada arista $(c, c') \in E$, de tal forma que es la frecuencia total de la co-ocurrencia de $c$ y $c'$ a lo largo de los párrafos en $d$.

## 4.2. Medidas

Habíamos mencionado anteriormente la obtención de medidas de Comprensibilidad por documento y, éstas serán tres: Carga de Comprensibilidad (Secuencialidad), Dispersión y Conectividad. A continuación explicamos cada una de las medidas así como el proceso para obtenerlas.

### 4.2.1. Carga de Comprensibilidad (Secuencialidad)

Para introducir la idea de Carga de Comprensibilidad utilizaremos el trabajo de [1]. Tres nociones son básicas en textos bien escritos: enfoque, secciones clave y secuencialidad en la presentación de conceptos. La primera nos habla de que un texto bien escrito explica pocos conceptos, de aquí que posee un enfoque. La segunda se refiere a que para cada concepto en un buen texto, se tiene una sección que se enfoca en explicar este concepto en específico. Finalmente, los conceptos en un texto bien escrito, son discutidos en una forma secuencial, esto quiere decir que un concepto es explicado de forma adecuada antes de que aparezcan ocurrencias de este concepto o de conceptos relacionados a éste a lo largo del documento. Si un concepto, llamémosle $c$, es utilizado para explicar un concepto relacionado $c'$ antes de que $c$ sea explicado adecuadamente, esto añade una Carga de Comprensibilidad al lector y es un error de secuencialidad, i.e. si queremos que nuestro lector entienda $c'$, que está relacionado con $c$, tenemos primeramente que explicar $c$. La carga de comprensibilidad como se obtiene en [18] se define a través del proceso que mostramos a continuación.

#### Significancia

Lo primero que debemos obtener es la Significancia, y antes de formularla expliquémosla. La significancia es una medida de qué tan importante es la descripción

de un concepto en una cierta sección para el entendimiento de otros conceptos en un documento. Un posible enfoque sería definir la significancia en términos de la frecuencia relativa del concepto en la respectiva sección, sin embargo, el problema con esta definición es que no diferencia entre dos conceptos $c_1$ y $c_2$ que contengan la misma frecuencia en una sección en específico. Así mismo, el concepto $c_1$ puede estar relacionado a múltiples conceptos que ocurren en otras partes del documento y pueden ser entonces más significantes para entender todo el documento, mientras que $c_2$ puede ser más relevante solamente para la sección analizada. Es por esto, que la significancia de un concepto se define con base en dos cosas: la frecuencia de un concepto en una respectiva sección, y la cantidad de conceptos relacionados con éste.

Dado un documento $d$ que contiene un concepto $c$, su conjunto de conceptos relacionados $R(c)$, un párrafo $p_k$ y la frecuencia de $c$ en $p_k$ como $f(c, p_k)$; la Significancia de $c$ en un $p_k$ se define como

$$\lambda(c, p_k) = f(c, p_k) * |R(c)|. \tag{4.1}$$

### Carga de Comprensibilidad en un Párrafo

Intuitivamente, utilizaremos que el párrafo clave para un concepto, tendrá un valor de significancia alta. Entonces, algorítmicamente obtendremos el párrafo clave para un concepto al comparar sus valores de significancia en todos los párrafos del documento, de tal forma que el párrafo clave $p_{k_c}$ de $c$ se define como el primer párrafo en $d$ donde la significancia es máxima, y la Carga de Comprensibilidad para $c$ en un párrafo $p_k$ se define como

$$\psi(c, p_k) = \begin{cases} \lambda(c, p_{k_c}), & \text{si } k < k_c \\ 0, & \text{en cualquier otro caso.} \end{cases} \tag{4.2}$$

Donde $k$ y $k_c$ se refieren a los números que indican el párrafo, de aquí que $k < k_c$ nos habla de los números de párrafo anteriores al número del párrafo clave.

Para entender la formulación anterior debemos observar con detenimiento que una ocurrencia del concepto en análisis en $p_{k_c}$ o después, no presenta una carga de comprensibilidad en el lector. Esto se debe a que el lector ya habrá entendido el concepto una vez que éste, haya sido explicado en lo que consideramos su párrafo clave.

### Carga de Comprensibilidad por Concepto

La Carga de Comprensibilidad total de $c$ es solamente la suma sobre todas las Cargas de Comprensibilidad de $c$ en cada párrafo, de tal forma que la ecuación quedaría como

$$\Psi(c) = \sum_{p_k \in D} \psi(c, p_k). \tag{4.3}$$

### Carga de Comprensibilidad del Documento

La Carga de Comprensibilidad se puede interpretar a través de lo presentado anteriormente, como una medida que cuantifica la falta de organización en las ocurrencias de los conceptos, ya que los conceptos organizados de forma secuencialmente

correcta no deben generar una Carga de Comprensibilidad en el lector. De aquí que la medida de secuencialidad de un documento es

$$Secuencilidad(d) = -\sum_{c \in C} \Psi(c). \tag{4.4}$$

Es claro también que un documento con una carga de comprensibilidad mayor, es de menor calidad (menos comprensible) que uno con una carga de comprensibilidad menor.

### 4.2.2. Conectividad

Un documento es más fácil de comprender cuando sus conceptos tienen una conexión lógica los unos con los otros, i.e. un documento debe escribirse de tal forma que cada uno de sus conceptos tenga relación con el concepto que le sigue, esto hace que podamos comprender el texto de forma correcta. Basándonos en esto es que empleamos la medida de Conectividad, que se define como el promedio de la fuerza de las relaciones entre los conceptos en un documento, ésta está basada en la colocación de los mismos y formulaciones de teoría de grafos para medir el grado (el número de nodos al cual un nodo focal está conectado) y la fuerza de un nodo (esto basado en los pesos).

#### Conectividad de un Concepto

Dado el grafo $G(C, E, w)$ de un documento $d$, conceptos $c, c' \in C$, y las conexiones de las aristas para $c$ como $E$; la conectividad de $c$ se define como

$$\theta(c) = \left( \sum_{(c,c') \in E} w(c, c') \right)^{\alpha} \tag{4.5}$$

donde $w$ es el peso de las aristas que conectan a $c$ y $\alpha$ es una constante con valor mayor a uno que proviene como un término para equilibrar tanto el grado como la fuerza de un nodo focal, todo esto de acuerdo con [25].

La fórmula proviene de formulaciones de teoría de grafos. El grado de un nodo focal es el número de adyacencias en una red, esto quiere decir el número de nodos al que un nodo focal está conectado y la formulación es la siguiente:

$$k_i = C_D(i) = \sum_{j}^{N} x_{ij}, \tag{4.6}$$

donde $i$ es el nodo focal, $j$ representa todos los otros nodos, $N$ es el número total de nodos y $x$ es la matriz de adyacencia en donde $x_{ij}$ se define como uno si $i$ está conectado con $j$ y cero de otra manera. El grado de un nodo focal se ha generalizado y extendido a la suma de lo pesos y se le refiere como la fuerza del nodo. Esta medida formalmente es, de acuerdo con [25]:

$$s_i = C_D^w = \sum_{j}^{N} w_{ij}, \tag{4.7}$$

donde $w$ es la matriz de pesos de ayacencia, en donde $w_{ij}$ es mayor a 0 si el nodo $i$ está conectado al nodo $j$ y el valor representa el peso de la unión. Aunque la medida de fuerza es un buen indicador, no es un indicador directo del número de nodos al que está conectado el nodo focal. En [25] combinan estas dos medidas (fuerza y grado) al incorporar $\alpha$, que determina la importancia relativa entre uniones y el peso de las uniones. En específico, ellos proponen una medida centralizada que es el producto del número de nodos al cual un nodo focal está conectado y el promedio de los pesos de estos nodos, ajustado con el parámetro $\alpha$. La medida es la siguiente:

$$C_D^{w\alpha} = k_i \times \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{(1-\alpha)} \times s_i^{\alpha}, \tag{4.8}$$

donde $\alpha$ es un parámetro de afinación. Es a través de este trabajo de investigación que se obtiene la fórmula para la obtención de nuestra medida de conectividad para un concepto $c$. Es a través de estas formulaciones que se obtiene la definición para la medida de Conectividad.

#### Conectividad Total

La conectividad total del documento se define como la sumatoria sobre todos los conceptos del documento, de tal manera que su ecuación es la siguiente

$$Conectividad(d) = \frac{\sum_{c \in C} \theta(c)}{|C|}. \tag{4.9}$$

Esta medida se puede interpretar como el promedio de la fuerza de las relaciones entre los conceptos en un documento, en específico de la fuerza y el grado de los conceptos vistos como nodos, en un documento. De tal forma que una mayor conectividad, nos habla de un documento que hilvana mejor las ideas entre los conceptos que se abordan en él.

### 4.2.3. Dispersión

El enfoque es una parte esencial de un documento bien escrito, de aquí que la falta del mismo puede ocasionar una carga de información sobre el lector y la necesidad de recordar más información conceptual. En [17] ellos proponen una formulación basada en teoría de la información para evaluar la cantidad de información esperada para ser recordada simultáneamente.

#### Dispersión por Concepto

Dado un documento $d$, un concepto $c \in C$, un párrafo $p_k$ como $P(c, p_k)$ (siendo la razón de $f(c, p_k)$ y el número total de conceptos en $p_k$; la dispersión de conceptos en $p_k$ basado en la entropía de Shannon se define como:

$$\theta(p_k) = -\sum_{c \in C} P(c, p_k) * log_2 P(c, p_k). \tag{4.10}$$

Más información, relacionada con el origen de esta fórmula, se abordó en el Capítulo 2 en Entropía de Shannon.

**Dispersión Total**

La medida de dispersión de los conceptos en el documento $d$ se define como

$$Dispersion(d) = \frac{\sum_{1 \leq k \leq K} \theta(p_k)}{K}. \tag{4.11}$$

Esta medida se debe interpretar inversamente, ya que que se busca que el lector no necesite recordar información conceptual en demasía, de aquí que un alto valor de la Dispersión total del documento equivale a un documento con una alta densidad de conceptos y, por ende, un documento difícil de comprender.

### 4.2.4. Ejemplo de Obtención de Representación

A continuación se incluye el ejemplo de una corrida para la obtención de la representación de un documento aleatorio del corpus de datos, el ejemplo aparece en la Figura 4.1.

1. Hasta arriba de la Figura 4.1, podemos observar que de un documento, se extrajeron sus secciones, en específico se observa la sección de objetivo con su contenido.

2. A continuación de esto, se divide el contenido de las secciones en párrafos, en la Figura 4.1 se observa la sección de problema con su contenido divido en párrafos.

3. Después, cada párrafo se divide en oraciones, se puede observar en la Figura 4.1, cómo la sección de objetivo se compone de un solo párrafo con una sola oración.

4. Le sigue el etiquetado de partes del discurso, y en la imagen se aprecia que la sección de objetivo, en el párrafo uno, en la oración uno, la palabra “Describir”, está etiquetada como un verbo.

5. Una vez con las palabras etiquetadas, extraemos los conceptos por párrafo, para este documento en particular podemos observar que en el párrafo cuatro, encontramos el concepto “monto de dinero”.

6. Ahora, obtenemos la ocurrencia de los conceptos por párrafo, de tal forma que para cada concepto observamos exactamente en qué párrafo ocurre en el documento.

7. Le sigue la asignación del párrafo clave, y esto se realiza como se indica en la Metodología para Comprensibilidad, en nuestro documento se observa que el concepto “departamento de contabilidad” tiene al párrafo doce y al cinco como posibles párrafos clave, acabaremos seleccionando al cinco como su párrafo clave debido a que este aparece antes en el documento.

8. Casi por finalizar, obtenemos los conceptos y su relación con otros a lo largo del documento. En la imagen se observa que el concepto “monto de dinero” está relacionado tanto con “seguimiento del proceso” como con “monto de dinero solicitado”.

9. Finalmente se construye el grafo de conceptos como aparece en el apartado de Grafo de Conceptos. Una vez con éste, somos capaces de obtener la Carga de Comprensibilidad, Conectividad y Dispersión del documento.

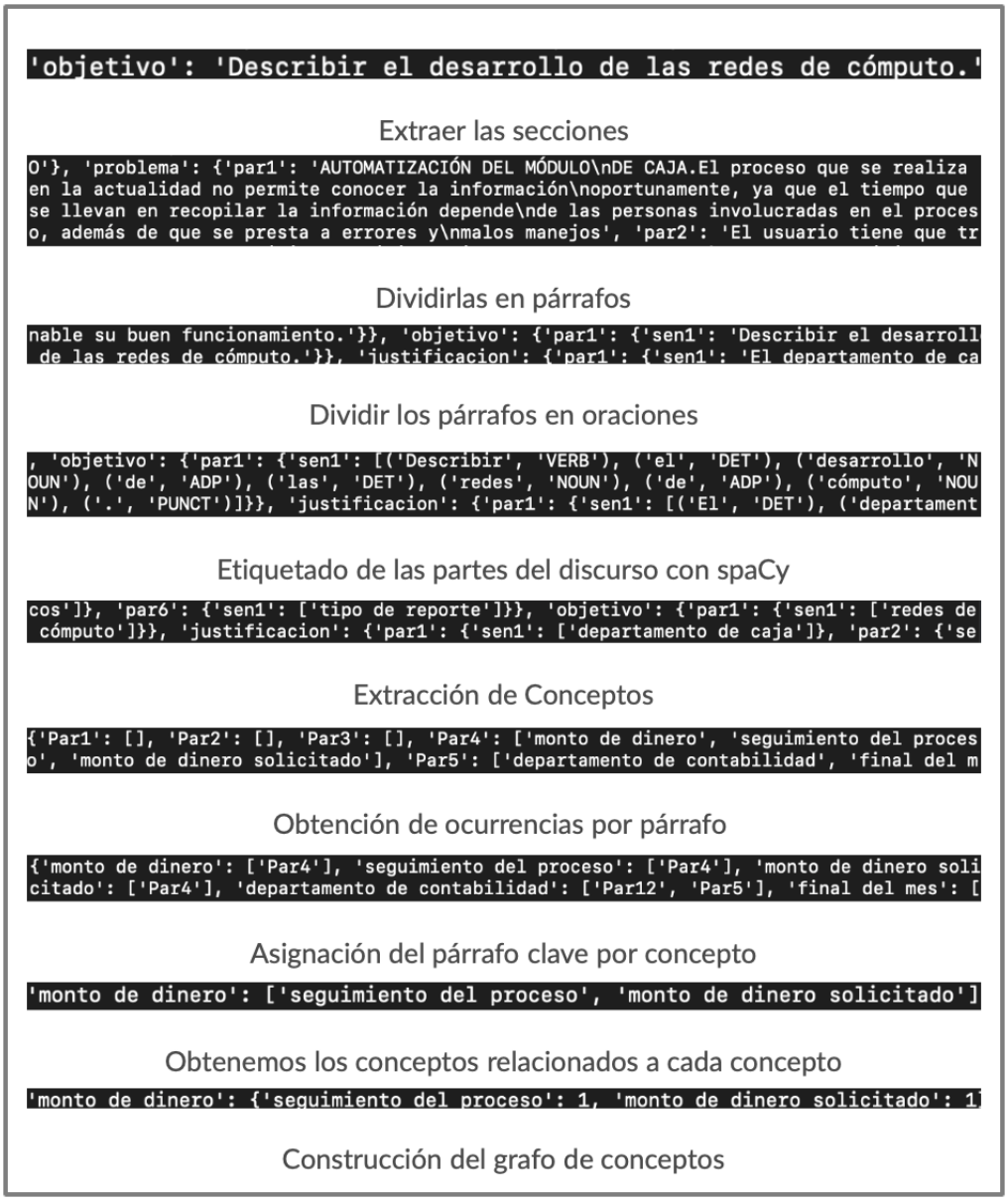



Figura 4.1: Ejemplo de corrida para obtención de la representación en un documento aleatorio del corpus de datos

## 4.3. Retroalimentación Visual

A continuación presentamos la metodología para la obtención de nuestros tres objetos visuales: Nubes de Términos, Grafos y Gráfica de Barras.

### 4.3.1. Nubes de Términos

Para la visualización de los valores por concepto que son encontrados en el documento, probamos diferentes tipos de gráficas y no nos eran útiles debido a que en algunos documentos la cantidad de conceptos podía ser sustanciosa. De aquí que decidimos utilizar Nubes de Términos para poder visualizar los conceptos y su frecuencia a lo largo del documento: la Carga de Comprensibilidad por concepto y la Conectividad por concepto. Las Nubes de Términos son una técnica para mostrar qué palabras (en nuestro caso utilizaremos conceptos) son más frecuentes a lo largo de un texto. Para crear este tipo de visualización, utilizamos la paquetería "wordcloud" que aparece en [24].

El proceso de obtención de una nube de términos consta de los siguientes pasos y se puede observar en la Figura 4.2:

1. Extracción de los conceptos (La Figura 4.2, muestra los conceptos encontrados en un respectivo documento, mientras el concepto tenga mayor tamaño esto significa que su frecuencia es mayor.)

2. Creación de un diccionario que contenga la forma concepto y valor (este puede variar en frecuencia, valor de Conectividad y Carga de Comprensibilidad)

3. Generación de la imagen de la Nube de Términos

4. Visualización de la imagen

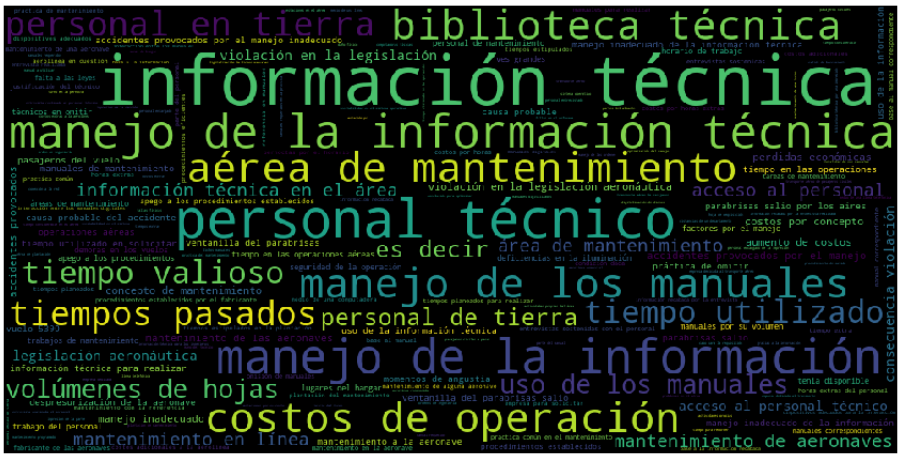



Figura 4.2: Nube de términos para un documento aleatorio de Doctorado

### 4.3.2. Grafos

Nuestro proyecto de investigación utiliza relaciones entre conceptos para la obtención de las medidas de Comprensibilidad, algunas de éstas son relevantes en el proceso de edición de un borrador de tesis de aquí que las incluimos en nuestro proceso de Retroalimentación Visual. Para mostrar estas relaciones utilizamos Grafos, y a través de ellos mostramos la relación que existe entre los conceptos y su párrafo clave y el Grafo de Conceptos del Documento, el segundo siendo base para

la construcción de las medidas de Conectividad y Dispersión. Para la construcción de este Grafo se utilizó la paquetería del lenguaje Python: NetworkX [3].

El proceso general para la obtención de nuestros grafos se enumera a continuación y un ejemplo de un grafo, en específico un ejemplo de un grafo que muestra relaciones entre conceptos se muestra en la Figura 4.3:

1. Creación del Grafo
2. Añadir un nodo por cada concepto al grafo (en la Figura 4.3 se observa que para cada concepto existe un nodo)
3. Generación de aristas por cada relación entre conceptos
4. Generación de posiciones para los nodos
5. Para cada una de nuestras aristas, generar el trazo de la arista entre los nodos
6. Graficación y personalización del Grafo
7. Impresión del grafo para su interacción con el usuario

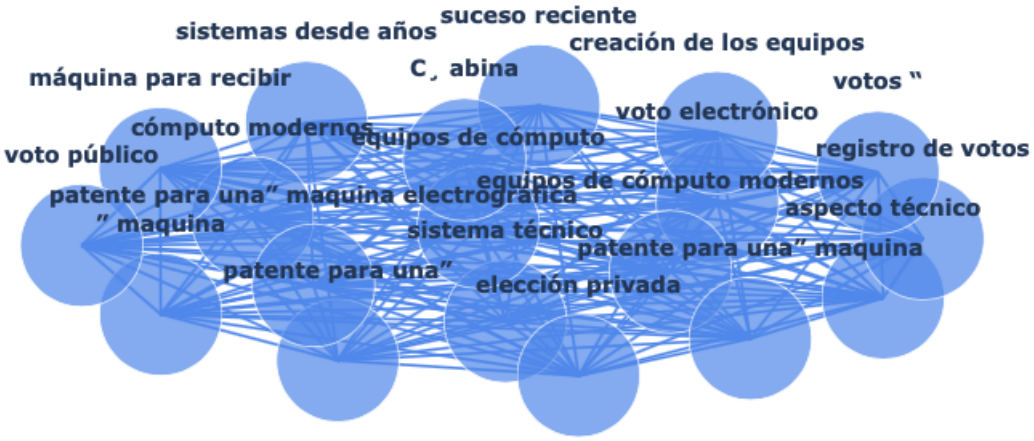



Figura 4.3: Grafo de relaciones entre conceptos para un documento aleatorio de Doctorado

### 4.3.3. Gráfica de Barras

Finalmente, para mostrar la dispersión por párrafo en nuestro documento, la metodología seleccionada fue una gráfica de barras donde señalamos con una línea roja la media del resultado para dispersión de la sección y nivel académico correspondiente. La paquetería que utilizamos es la perteneciente al lenguaje Python: Matplotlib [4] y esta aparece en [16].

[3]NetworkX es una paquete del lenguaje Python para la exploración y análisis de redes y algoritmos de redes. El paquete base ofrece estructuras de datos para representar diferentes tipos de redes o grafos, incluyendo grafos simple, dirigido y grafo con aristas paralelas y con bucles internos.

[4]Matplotlib es una paquete del lenguaje Python para desarrollo de aplicaciones, secuencias de comandos interactivos y generación de imágenes con calidad de publicación.

El proceso general para la generación de una gráfica de barras en nuestro proyecto se enumera a continuación y se muestra un ejemplo de de una gráfica de barras para una sección en específico de un documento de Doctorado en la Figura 4.4:

1. Obtención de los valores para dispersión por párrafo
2. Generación del diccionario que contenga la relación del párrafo y su valor de dispersión.
3. Creación de la figura.
4. Generación de la gráfica de barras sobre la figura.
5. Obtención del valor medio de Dispersión para la sección y nivel académico específico y creación de línea roja horizontal (ver Figura 4.4
6. Impresión en pantalla de la figura.

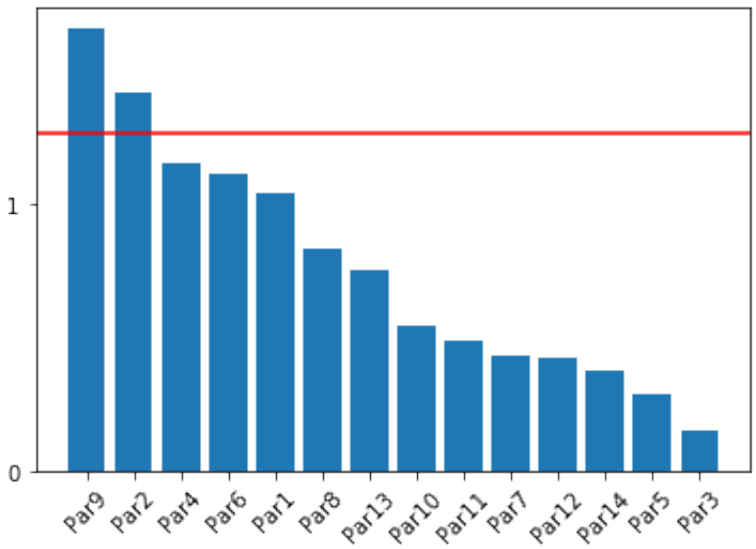



Figura 4.4: Gráfica de barras para un documento aleatorio de Doctorado

# Capítulo 5

# Colección de Datos

Los documentos que se recolectaron a partir del trabajo de [9], están compuestos por tesis y propuestas de investigación desarrolladas por estudiantes. La colección consta, por el momento, de 712 documentos que se dividen en cuatro niveles académicos: Técnico Superior Universitario (TSU), Licenciatura, Maestría y Doctorado.

Los documentos se obtuvieron de repositorios de universidades públicas y centros de investigación, el 97 % de ellos corresponde a instituciones mexicanas y el restante a universidades de otros países de habla hispana. Las secciones totales de cada uno de los documentos son: Título, Planteamiento del Problema, Justificación, Preguntas de Investigación, Hipótesis, Objetivos, Metodología y Conclusiones. Y, de todas ellas, nuestro estudio se centra en tres secciones en específico: Planteamiento del Problema, Justificación y Resultados.

## 5.1. Número de Documentos

A continuación se incluye una tabla que indica la cantidad de documentos que se tiene por nivel académico.

| Nivel Académico | Número de Documentos |
|---|---|
| Técnico Superior Universitario | 227 |
| Licenciatura | 150 |
| Maestría | 269 |
| Doctorado | 66 |
| Total | 712 |

Tabla 5.1: Nivel académico y número de documentos

En 5.1 podemos observar que Doctorado posee la menor cantidad de documentos, esto se puede explicar a través de la naturaleza del nivel académico. Debido a lo anterior, es claro que nuestros datos no están balanceados, algo que puede afectar a nuestros resultados y en las conclusiones comentamos este aspecto.

## 5.2. Documentos Depurados

Tanto en los experimentos realizados para Comprensibilidad como los realizados para Legibilidad, hemos extraído tres secciones de cada documento de tesis: Planteamiento del Problema, Justificación y Resultados. Escogimos dichas secciones debido a que contienen mayor información en relación con el documento de tesis, e.g. la sección "Título" es muy breve y por ende no fue seleccionada.

La tabla que mostramos a continuación nos muestra la cantidad total de tesis, así como la cantidad total de tesis seleccionadas después de un proceso de depuración. Por proceso de depuración nos referimos a que seleccionamos aquellas tesis que tenían al menos una sección con un párrafo, refiriéndonos a un párrafo como un conjunto de texto que al menos contiene dos oraciones.

| Nivel Académico | Número de Tesis | Depuradas |
|---|---|---|
| **Técnico Superior Universitario** | 227 | 202 |
| **Licenciatura** | 150 | 136 |
| **Maestría** | 269 | 254 |
| **Doctorado** | 66 | 64 |

Tabla 5.2: Nivel académico y documentos depurados

En la Tabla 5.2 podemos observar que todos los niveles académicos disminuyen los documentos de tesis, sin embargo, Técnico Superior Universitario es el que pierde el mayor número y Doctorado pierde el menor. Esto se da ya que los documentos de Técnico Superior Universitario presentan mayor cantidad de errores y son más cortos, en tanto que los de Doctorado son documentos con mayor cuerpo que fueron escritos por estudiantes con una mayor maduración académica y práctica en la escritura de documentos académicos.

## 5.3. Secciones

La Tabla 5.3 muestra la cantidad de secciones (Planteamiento del Problema, Justificación y Resultados) recabadas por nivel académico. Es importante recalcar que estas secciones fueron obtenidas del conjunto de documentos ya depurados.

| Nivel Académico | Sección | | |
|---|---|---|---|
| | Planteamiento | Justificación | Resultados |
| **Técnico Superior Universitario** | 80 | 104 | 102 |
| **Licenciatura** | 28 | 31 | 77 |
| **Maestría** | 92 | 81 | 179 |
| **Doctorado** | 21 | 13 | 43 |
| **Total** | 221 | 229 | 401 |

Tabla 5.3: Secciones por nivel académico

En la Tabla 5.3 observamos que la mayor cantidad de secciones recabadas corresponde a la sección de Resultados, seguida por Justificación y finalmente Plantea-

miento, aunque estas últimas dos tienen buen balance. Así mismo, podemos observar que Doctorado tiene el menor número de secciones recabadas, seguido por Licenciatura. El nivel académico que mayor número de secciones tiene es el de Maestría, seguido por Técnico Superior Universitario. Resultados es la sección con más ejemplos debido a que tiende a ser donde se presentan resultados y se puede dar más información que en la sección de Planteamiento del Problema y Justificación. Con base en los niveles académicos y su cantidad de secciones recabadas, esto está íntimamente relacionado con la cantidad de tesis que contienen los niveles académicos, de aquí que estos resultados eran esperados.

## 5.4. Estadística para Palabras

Finalmente, incluimos la Tabla 5.4 donde presentamos estadísticas para las Palabras/Tokens por sección y nivel académico. Para cada nivel académico, recabamos sus secciones (Problema, Justificación y Resultados) y por cada sección obtenemos el promedio de palabras que aparecen por documento, así como el documento con el mayor y menor número de palabras/tokens.

| Nivel Académico | Planteamiento de Problema | | | Justificación | | | Resultados | | |
|---|---|---|---|---|---|---|---|---|---|
| | Promedio | Máximo | Mínimo | Promedio | Máximo | Mínimo | Promedio | Máximo | Mínimo |
| **Técnico Superior Universitario** | 409.33 | 1708 | 75 | 385.07 | 1123 | 110 | 493.55 | 1883 | 138 |
| **Licenciatura** | 477.60 | 1253 | 114 | 342.33 | 820 | 109 | 508.92 | 2386 | 121 |
| **Maestría** | 399.39 | 1063 | 121 | 407.81 | 1612 | 113 | 656.65 | 5184 | 137 |
| **Doctorado** | 754.85 | 1643 | 310 | 433.84 | 802 | 247 | 840.27 | 3765 | 59 |

Tabla 5.4: Estadísticas para palabras por secciones y nivel académico

En la Tabla 5.4 podemos observar una tendencia a mayor nivel académico, mayor cantidad de palabras en promedio, en cada una de las secciones. Así mismo, se observa que la sección de Resultados, en promedio, posee una mayor cantidad de palabras que las otras secciones. La explicación de esto se debe a que, de nuevo, la sección de Resultados tiende a ser más extensa y elaborada que las otras dos. Lo mismo ocurre para el nivel académico, ya que las tesis de Doctorado tienden a ser más extensas que aquellas de menores niveles académicos.

# Capítulo 6

# Evaluación y Retroalimentación Visual en Comprensibilidad

## 6.1. Evaluación

Dado que uno de los propósitos de nuestra investigación es el tener una evaluación de los escritos en proceso (borradores) para así proveer una guía al autor, se decidió obtener las medidas de Comprensibilidad de la colección de textos que previamente hemos presentado. Este análisis previo, además de validar el proceso computacional en textos de diferentes niveles académicos, es capaz de proveer valores de referencia para evaluar nuevos textos. Esta evaluación se realiza sobre tres secciones (de longitud amplia) específicas por cada documento de tesis; éstas son: Planteamiento del Problema, Justificación y Resultados.

### 6.1.1. Experimentos

Los resultados que se mostrarán a continuación se basan en la serie de experimentos que explicaremos a continuación. Recordemos que por cada sección de interés (Planteamiento del Problema, Justificación y Resultados), se obtuvieron tanto promedio, desviación estándar y porcentaje de valores no-nulos (refiriéndose a documentos donde el valor de la medida fue diferente de cero) de nuestras tres medidas: Carga de Comprensibilidad, Conectividad y Dispersión. Debido a que contamos con cuatro niveles académicos, esto se hizo para TSU, Licenciatura, Maestría y Doctorado, dando lugar a las Tablas 6.1, 6.2 y 6.3.

En total entonces, se tienen tres diferentes tablas asociadas a tres diferentes experimentos. Dentro de cada medida se decidió reportar al promedio debido a que nos daría una visión general del valor de la medida a lo largo de todos los documentos correspondientes a un nivel académico, ésta, tentativamente útil para la generación de valores de referencia. La desviación estándar, se seleccionó para observar la variabilidad en este promedio, ya que está en nuestro interés observar la viabilidad de este valor como referente de evaluación de una medida para un cierto nivel académico. Finalmente, y debido a que tenemos medidas difíciles de recabar, como lo serían la carga de comprensibilidad, se decidió realizar tanto el promedio como la desviación estándar sobre los documentos que otorgaran un valor diferente de cero; esto por dos cosas: para estabilizar el resultado de la desviación estándar y debido a que los valores nulos no aportaban información relevante fuera de su

porcentaje, que ya está siendo reportado en la columna "No-nulos( %)".

### 6.1.2. Evaluación de la sección "Planteamiento del Problema"

El número de textos para esta sección no es la más abundante como podemos ver por la Tabla 6.1. Así mismo, podemos observar que los resultados para la carga de comprensibilidad con base en el nivel académico de Técnico Superior Universitario en la Tabla 6.1, son bastante altos para los pocos documentos de tesis utilizados al obtener esta medida; el nivel académico que le seguiría sería Doctorado pero con una diferencia de valor sustancial: 1,181.8. Un problema que se observa con la desviación estándar es que, en general, tiene valores superiores a los encontrados en el promedio. Finalmente y con base en el porcentaje de valores no-nulos para carga de comprensibilidad, observamos que los resultados llevan sentido, ya que en documentos más complicados, como lo serían los de Doctorado, se logra recabar un mayor porcentaje de valores.

Enfocando nuestra atención sobre la Conectividad ahora, podemos observar una mejoría en el porcentaje de datos no-nulos recabados, que va desde un 80.95 % hasta un 21.25 %, en el peor de los casos. La desviación estándar continúa por arriba de los valores del promedio, esto tanto para valores con bajo nivel de porcentaje de no-nulo, como para los altos. En cuanto al promedio, fuera del resultado para el nivel académico de Licenciatura, se observa una tendencia hacia una disminución de valor relacionada con el nivel académico, de tal forma que a menor nivel académico, menor promedio.

Refiriéndonos ahora a la medida de Dispersión, encontramos que sus valores para promedio no varían en demasía entre los diferentes niveles académicos, obteniendo la mayor diferencia entre Licenciatura y TSU con un valor de 0.38. Contrario a la Carga de comprensibilidad y Conectividad, aquí observamos que la desviación estándar está por debajo de los valores del promedio. Así mismo, somos capaces de obtener la medida de Dispersión de absolutamente todos los documentos para cada uno de los niveles académicos.

La resultados para la sección de "Planteamiento del Problema" nos invitan a pensar que contrario a una de nuestras hipótesis iniciales, no encontramos una relación clara entre nivel académico y mayor o menor valor del promedio para ninguna de nuestra tres medidas de Comprensibilidad. La desviación estándar se comporta errática tanto para la medida de Conectividad como para la de Comprensibilidad y esto sucede debido al bajo porcentaje de valores no-nulos, presencia de valores extremos y alta variabilidad en nuestros resultados. Finalmente, al encontrar un alto valor de porcentaje de valores no nulos para Dispersión, un nivel medio para Conectividad y uno bajo para Carga de Comprensibilidad, nos indican que la última, es una medida más difícil de recabar. Esto, intuitivamente lo explicamos debido a que por su naturaleza, es una medida que profundiza más el análisis de un documento de texto, recaba sus conceptos, la relación de éstos y su aparición a lo largo del documento así como su ordenamiento; de aquí que una selección cuidadosa de documentos acompañada de de un proceso óptimo para obtención de conceptos y sus relaciones, es necesaria si queremos recabarla. Un indicio que indica esto, es el hecho de que a mayor porcentaje de valores no-nulos, entonces vemos un mayor nivel académico, esto nos indica que en un documento de mayor nivel académico (mejor

escrito), recabar esta medida se vuelva más sencillo.

| | Conectividad | | | Dispersión | | | Carga de Comprensibilidad | | |
|---|---|---|---|---|---|---|---|---|---|
| | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) |
| TSU | 1.02 | 1.26 | 21.25 | 1.21 | 0.53 | 100.00 | 1340.00 | 1058.42 | 5.00 |
| Licenciatura | 5.51 | 12.45 | 25.00 | 1.59 | 0.85 | 100.00 | 54.00 | 63.69 | 10.71 |
| Maestría | 1.97 | 2.86 | 45.65 | 1.37 | 0.53 | 100.00 | 48.60 | 91.16 | 10.87 |
| Doctorado | 4.09 | 6.07 | 80.95 | 1.27 | 0.76 | 100.00 | 158.20 | 185.97 | 23.81 |

Tabla 6.1: Estadísticas de las Medidas de Comprensibilidad para la sección de Planteamiento del Problema

### 6.1.3. ón de la sección "Justificación"

Esta sección fue la más escasa, como podemos observar en la Tabla 6.2, esto claro con excepción del nivel académico de Técnico Superior Universitario. Los valores de Carga de Comprensibilidad son máximos para el nivel académico de Técnico Superior Universitario, al igual que con la sección de "Planteamiento del Problema". No se observa un decrecimiento del promedio mientras aumentamos el nivel académico y, otro aspecto clave a notar es que debido al reducido número de documentos de Doctorado, ningún documento tuvo una Carga de Comprensibilidad distinta de cero y, por ende, no se tiene información respecto a esta medida. La desviación estándar, se comporta igual que los resultados para la sección de "Planteamiento del Problema" y el porcentaje de valores no-nulos es el peor recabado para las tres secciones.

Con base en la medida de Conectividad, se observa que el porcentaje de valores no-nulos que se recaba es relativamente constante entre los cuatro diferentes niveles académicos y ronda entre un 30 % y 40 %. La desviación estándar, se comporta mejor que para la sección de "Planteamiento del Problema" y tanto para Licenciatura como para Doctorado, se obtienen valores menores al promedio. El promedio en Conectividad, fuera del valor de Doctorado, tiene un comportamiento tal que para un mayor nivel académico, se tiene un mayor valor de Conectividad; debido a que no funciona así para todos los niveles académicos, no podemos concluir nada.

La medida de Dispersión se comporta de forma similar a la sección de "Planteamiento del Problema". Los resultados del promedio no varían en demasía y van desde 1.18 hasta 1.50. Los valores de la desviación estándar están por debajo de los valores de promedio y el porcentaje de valores no-nulos es perfecto, recabando así para cada documento, su respectivo valor de Dispersión.

La sección de "Justificación" muestra un comportamiento similar al de la sección de "Planteamiento del Problema", donde la Carga de Comprensibilidad tiene el peor porcentaje de no-nulos, seguido de Conectividad y finalmente Dispersión. La tesis de que la Carga de Comprensibilidad es mejor recabada en documentos de mayor nivel académico aquí no se hizo válida ya que incluso, no recaba ningún valor para Doctorado, esto podría deberse a la poca cantidad de documentos o debido a la sección en particular. La desviación estándar se comporta de forma errática y esto empeora en niveles donde tenemos una menor cantidad de valores no-nulos, valores extremos y una alta variabilidad en los resultados; aún así, observamos una mejoría en Conectividad en comparación con la sección de "Planteamiento del Problema". Esto no se explica a través del porcentaje de valores no-nulos, ya que es mejor en

la sección anterior, quizá la respuesta se encuentre en el tipo de textos que provee la sección en específico. Finalmente, la dispersión se comporta de forma similar a la sección de "Planteamiento del Problema" y una de las preguntas de interés sería el porqué no encontramos valores con poca variabilidad.

| | Conectividad | | | Dispersión | | | Carga de Comprensibilidad | | |
|---|---|---|---|---|---|---|---|---|---|
| | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) |
| TSU | 0.57 | ±0,65 | 30.77 | 1.18 | ±0,54 | 100.0 | 343.50 | ±272,75 | 7.69 |
| Licenciatura | 0.91 | ±0,89 | 38.10 | 1.39 | ±0,57 | 100.0 | 33.00 | ±27,00 | 9.52 |
| Maestría | 2.97 | ±5,91 | 30.86 | 1.50 | ±0,71 | 100.0 | 90.67 | ±105,59 | 3.70 |
| Doctorado | 2.34 | ±2,22 | 38.46 | 1.36 | ±0,61 | 100.0 | 0.0 | ±0,0 | 0.0 |

Tabla 6.2: Estadísticas relacionadas con las Medidas de Comprensibilidad para la Sección de Justificación

### 6.1.4. Evaluación de la sección "Resultados"

Al observar la Tabla 6.3, el nivel académico de Técnico Superior Universitario obtiene los resultados más altos para la medida de Carga de Comprensibilidad, un resultado que lo observamos repetirse tanto para la sección de "Planteamiento del Problema" como para "Justificación". Es importante notar que los valores, en general, son más altos que para las dos secciones anteriores y fuera del nivel académico de Maestría o de Doctorado, vemos una relación tal que a mayor nivel académico, menor Carga de Comprensibilidad. Esto lleva sentido con la construcción de la medida ya que un documento de menor nivel académico debe ser menos comprensible que uno de mayor y, esto se debe ver reflejado en una alta Carga de Comprensibilidad. Contrario a las dos secciones anteriores, se logra obtener un valor de desviación estándar menor para el nivel académico de Doctorado y esto es una indicación de que documentos con una mayor extensión como lo serían los de la sección de "Resultados" recaban de una mejor manera la medida de Carga de Comprensibilidad. El porcentaje de documentos con Carga de Comprensibilidad no-nula no son los mejores de todas las secciones, ya que la sección de "Planteamiento del Problema", en general, tiene mayores porcentajes.

Con base en la Conectividad, podemos observar una tendencia hacia mayor nivel académico, mayor valor de promedio de Conectividad, sin embargo, el nivel académico de Maestría rompe con ello. Ver una tendencia de este tipo hubiese llevado sentido ya que esperaríamos mayores niveles de Conectividad en documentos de mejor calidad, como lo serían los de Doctorado. La desviación estándar otorga malos resultados en todos los niveles académicos, esto debido a que los resultados presentan alta variabilidad y la presencia de valores extremos. En cuanto al porcentaje de valores no-nulos, se observa un comportamiento similar al de la sección de "Planteamiento del Problema" con valores que van desde un 22.08 % hasta un 76.74 %, esto es de esperarse basándonos en los dos experimentos anteriores. Vemos también una tendencia donde a mayor nivel académico, mayor porcentaje de valores no-nulos, ésta se rompe por el nivel académico de Licenciatura. Esperábamos encontrar este tipo de tendencias a lo largo de nuestras tablas pero siempre encontramos un contraejemplo.

Finalmente, y con base en la Dispersión, los valores son muy similares a lo largo de las tres secciones, en específico en esta sección van desde 1.02 hasta 1.29. La desviación estándar está por debajo en todos los niveles académicos y esto tiene

que ver con una baja variabilidad de los datos y también con el hecho de que el porcentaje de valores no-nulos por nivel académico, se recaba perfectamente.

| | Conectividad | | | Dispersión | | | Carga de Comprensibilidad | | |
|---|---|---|---|---|---|---|---|---|---|
| | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) | Promedio | (+/-)SD | No-nulos(%) |
| TSU | 4.59 | 23.39 | 35.29 | 1.02 | 0.41 | 100.0 | 1433.25 | 1920.65 | 7.84 |
| Licenciatura | 5.11 | 9.22 | 22.08 | 1.29 | 0.72 | 100.0 | 336.60 | 372.63 | 6.49 |
| Maestría | 29.07 | 257.22 | 58.10 | 1.07 | 0.54 | 100.0 | 138.27 | 237.50 | 12.29 |
| Doctorado | 18.72 | 57.74 | 76.74 | 1.16 | 0.79 | 100.0 | 270.25 | 233.17 | 18.60 |

Tabla 6.3: Estadísticas de las Medidas de Comprensibilidad para la Sección de Resultados

### 6.1.5. Análisis de las Medidas en la Colección de Textos Académicos

Con base en la estadística proporcionada en las Tablas 6.1, 6.2 y 6.3, podemos concluir que la Dispersión es la medida que muestra un comportamiento más constante a lo largo de las tres secciones. Esta medida abarca valores que van desde 1.02 hasta 1.59 y debido a que es la medida que posee desviaciones estándar menores al valor promedio a lo largo de todos los niveles académicos y todas las secciones, se concluye que tiene valores con poca variabilidad y de confianza; esto permitiendo utilizar los valores obtenidos como puntos de referencia en documentos en proceso a valorizar. Así mismo, somos capaces de recabar esta medida a lo largo de todos los niveles académicos y secciones en cada uno de los documentos, siendo así una medida ideal para obtener información relacionada con la comprensibilidad de un documento en proceso. Es importante aclarar que la dispersión es una medida que se comporta así debido a la naturaleza de su construcción, para obtenerla solamente hacemos uso de los conceptos en un documento, de aquí que procesarla es sencillo ya que conceptos encontramos en cada uno de nuestros documentos. Sería interesante un análisis más profundo sobre la metodología de obtención de los conceptos si queremos experimentar más a profundidad con esta medida.

En lo que respecta a la conectividad, esta medida tiene cierta tendencia a comportarse de tal forma que un mayor nivel académico, indicaría un mayor valor de promedio, sin embargo, siempre logramos encontrar un contraejemplo que rompe con esto. La desviación estándar para esta medida es mejor que para la Carga de Comprensibilidad, logrando en algunos niveles académicos de algunas secciones estar por debajo del promedio, sin embargo, es peor que para dispersión y, en general, no se encuentra debajo del promedio. Esto implica que la medida tiene una alta variabilidad, presencia de valores extremos y que podría mejorarse el porcentaje de valores no-nulos que posee. Lograr esto, en esta medida en específico, implicaría la obtención de documentos con mejores relaciones entre conceptos tanto de grado como de fuerza, esto quiere decir la obtención de documentos con una mayor cantidad de conceptos, mayores relaciones entre los mismos y una mayor cantidad de ocurrencias de ellos a lo largo de diferentes párrafos de la sección. También sería importante analizar la metodología de obtención de conceptos y cómo los consideramos relacionados, esto seguramente aportaría experimentos interesantes. No alcanza a ser una medida utilizable, pero con las mejores anteriormente mencionadas mejoraría cuantiosamente para la valoración de comprensibilidad, tanto con los parámetros generados en las tablas como para la entrada de un documento nuevo.

La Carga de Comprensibilidad es la medida que refleja una evaluación en general del texto debido a que muestra cómo los conceptos se desarrollan a lo largo de los párrafos del texto. Y no solamente nos muestra eso, también permite una valorización de la secuencialidad de estos y si existe un enfoque en los mismos. Esta medida es computada en el menor número de textos porque requiere un análisis a los conceptos, relación de los conceptos, párrafos y documento en general. Notamos que para las tres secciones, los textos académicos pertenecientes a Técnico Superior Universitario obtenían valores más altos; posiblemente porque diversos conceptos son mencionados pero no adecuadamente introducidos en el texto o llevados en secuencia. Así mismo, podemos observar una ligera tendencia tal que a menor nivel académico del texto, mayor valor de promedio, sin embargo, no acaba por forjarse ya que siempre existe un contraejemplo. La desviación estándar para esta medida no fue un éxito, siempre ronda por arriba del promedio indicando alta variabilidad y valores extremos, de aquí que no se puede confiar en el promedio como un valor de referencia. Aunque concluimos que es una medida que no está lista para utilizarse como referente en una valoración de comprensibilidad, es la medida que más nos interesa para continuar estudiando. Esto debido a que logramos ver respuestas interesantes de la misma, por ejemplo, hay tendencias que muestran que documentos con mayor cuerpo como lo serían los de Doctorado, tienen una mayor probabilidad de obtener un valor para la medidad de Carga de Comprensibilidad. Así mismo, un estudio en profundidad sobre la metodología para la obtención de conceptos y sus relaciones sería de interés.

## 6.2. Retroalimentación Visual

Ya que queremos proveer un recurso para evaluar documentos en proceso (borradores) y así determinar la comprensibilidad en términos de las tres medidas que hemos venido mencionando, algunos métodos para reportar la evaluación al autor han sido propuestos. Estos métodos muestran resultados preliminares relevantes a lo largo de las diferentes etapas de la evaluación y son presentados e ilustrados a continuación para dos niveles específicos: Técnico Superior Universitario y Doctorado.

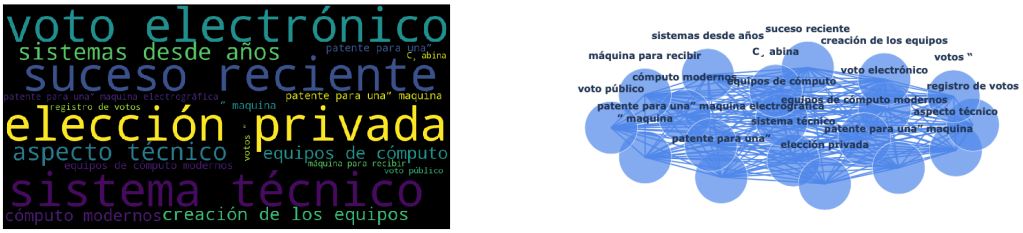
### 6.2.1. Ejemplo de Retroalimentación Visual en un Documento de TSU

#### Conceptos y Relaciones entre Conceptos

El primer paso en el análisis es la identificación de los conceptos que se abordan en el documento sometido a evaluación para así construir la representación necesaria para obtener sus medidas de Comprensibilidad. Una vez que este conjunto de conceptos es creado, una nube de términos es generada y reportada como se muestra en la Figura 6.1a, de tal forma que los conceptos que aparecen con mayor frecuencia poseen un mayor tamaño de fuente. La segunda imagen, que se muestra en la Figura 6.1b, muestra un grafo que contiene a los conceptos que aparecen en el documento y las relaciones que existen entre éstos.

Al presentarle la Figura 6.1a al autor del texto evaluado, éste es capaz, en primer lugar, de observar los conceptos que aparecen en su documento de tesis y la frecuencia de los mismos. Con esta información, el autor tiene la posibilidad de visualizar si

efectivamente, aborda con la debida prioridad los conceptos que él considera sustanciales en esta sección de su tesis. Por otro lado, el autor recibe un grafo interactivo como aparece en la Figura 6.1b, donde es capaz de observar los conceptos encontrados y su relación con los otros conceptos que aparecen en el documento. El autor puede hacer uso de este grafo para observar si conecta conceptos como se debe, de tal forma que los conceptos principales de su investigación deben estar conectados con conceptos menos sustanciales, que son utilizados para explicarlos. Si este no es el caso, el autor podría generar oraciones hilvanantes que asocien estos conceptos, y no solamente eso, si se encuentran conceptos completamente aislados, es importante generar oraciones que relacionen éstos con otros en el trabajo, ya que un documento bien escrito, tiene alta Conectividad.




(a) Nube de términos para conceptos

(b) Grafo de conceptos

Figura 6.1: Representación visual para frecuencia de conceptos y sus relaciones

### Párrafos Clave y CC para Conceptos

La Figura 6.2 nos muestra un grafo que permite la visualización entre los conceptos y sus respectivos párrafos clave (ver sección 5.2.1). Es importante recordar que estos párrafos clave son los párrafos donde teóricamente, el concepto es explicado en el documento. Una vez con este grafo, el autor debe de buscar en su documento y ver si efectivamente el párrafo en cuestión es el párrafo que explica este concepto. Si este no es el caso, el autor podría a analizar si hacer más menciones sobre el concepto explicado en el párrafo que él considera es el clave o disminuir las menciones sobre el concepto si aparece en un párrafo donde no ha sido explicado todavía, esto con miras a disminuir su Carga de Comprensibilidad.

Con base en la Figura 6.3, ésta nos habla de los conceptos y su Carga de Comprensibilidad. La Carga de Comprensibilidad se obtiene por cada uno de los conceptos en el documento y una vez con esta información, se construye una nube de términos para su visualización, de tal forma que un concepto con mayor tamaño de fuente, indica una Carga de Comprensibilidad mayor. A través de los valores de Carga de Comprensibilidad por concepto, construimos la Carga de Comprensibilidad para todo el documento (ver sección 5.2.1). Al observar la Figura 6.3, el autor es capaz de visualizar los conceptos que más peso generan al lector. Estos conceptos deben ser abordados en su trabajo de investigación para disminuir su Carga de Comprensibilidad y esto puede realizarse al analizar el párrafo clave de los mismos. La problemática aquí, se genera debido a que el concepto aparece a lo largo de diferentes párrafos antes de ser explicado en su párrafo clave. Para solucionar esto debemos introducir un párrafo donde expliquemos debidamente el concepto, este párrafo debe contener suficientes menciones al concepto para aumentar su secuencialidad y debe estar lo antes posible en el orden de párrafos donde el concepto ocurre.

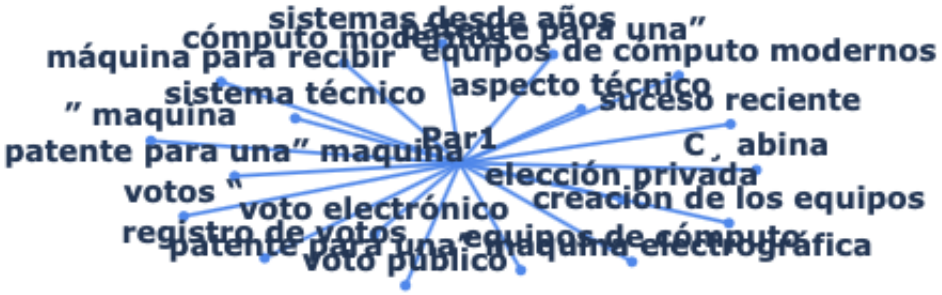



Figura 6.2: Relación entre conceptos y su párrafo clave

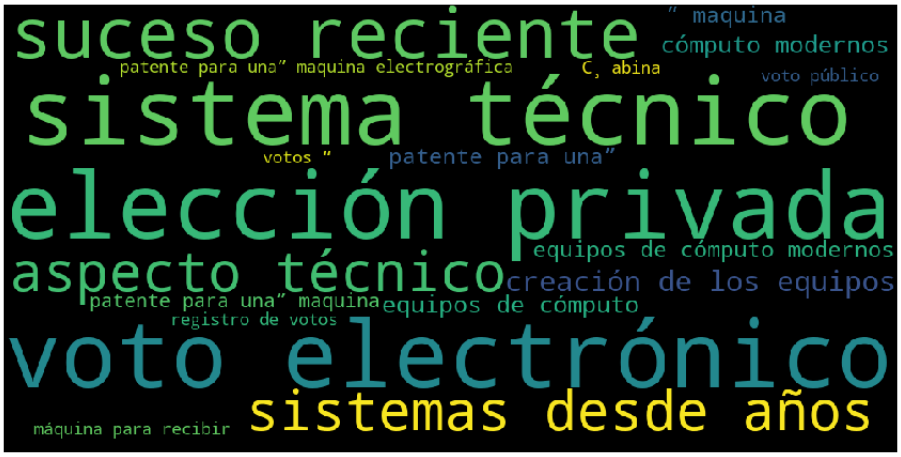



Figura 6.3: Nube de términos con base en la medida de carga de comprensibilidad por concepto

### Grafo del Documento y Valor de Conectividad por Concepto

Anteriormente mencionábamos la creación de un grafo del documento (ver Capítulo 5.2.2) y en la Figura 6.4 podemos visualizarlo; gracias a él somos capaces de visualizar los conceptos, sus relaciones y dependiendo del radio del nodo, la cantidad de relaciones que éste posee. Similar a la Figura 6.1a, el autor debe realizar lo que se mencionó para ésta, pero aunado a esto, el grafo contiene la cantidad de relaciones y esto nos sirve para obtener la medida de Conectividad y Dispersión del documento. El autor puede hacer uso de esta visualización para encontrar los conceptos más conectados en el documento y trabajar en los menos conectados a través de la creación de oraciones que los conecten con otros conceptos sustanciales en su trabajo de investigación.

Con base en la Figura 6.5, se construye una nube de términos con base en los valores de Conectividad por concepto que aparecen en el documento; finalmente, es con la suma de estos valores que se obtiene la Conectividad total del documento. La imagen debe invitar al autor del texto a actuar sobre los conceptos con tipografía pequeña, esto debido a que son valores que tienen pocas relaciones con otros conceptos y, en ocasiones, valores que no tienen conexiones. El autor podría revisar su documento y, como sugerimos previamente, conectar estos conceptos al crear ideas entre estos y conceptos sustanciales en la investigación. Recordemos que todo esto

parte de generar nodos con mayor Conectividad y de que un documento que conecta bien todas las ideas, es un documento más comprensible.

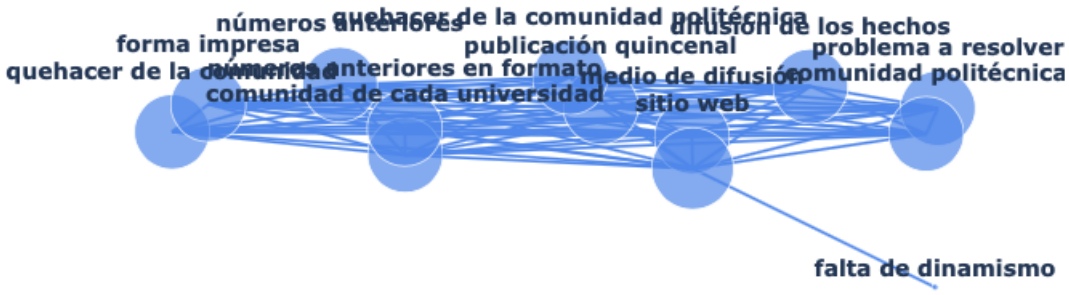



Figura 6.4: Grafo de Conceptos para el Documento

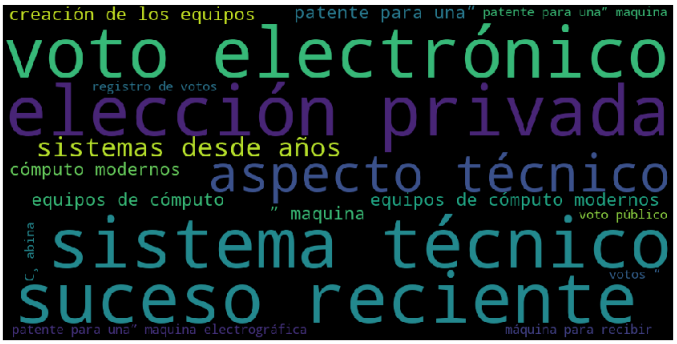
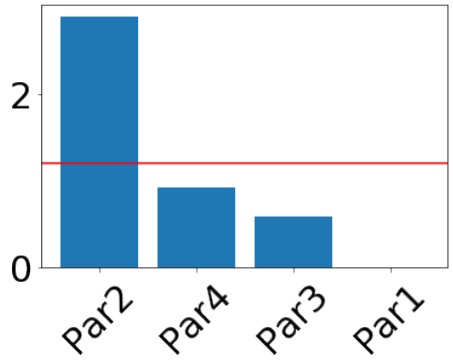


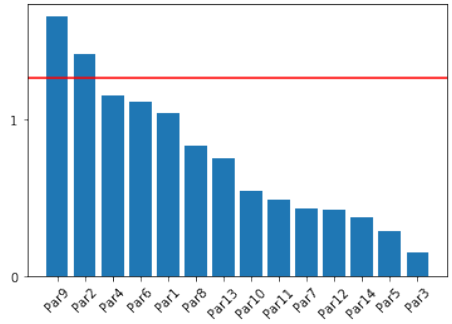

Figura 6.5: Nube de términos con base en la medida de Conectividad por concepto

### Dispersión

Los resultados de calcular la medida de Dispersión de los párrafos son reportados con una gráfica de barras como se muestra en la Figura 6.6a, esta figura en específico es para un documento de Técnico Superior Universitario. Aquí se emplea un umbral de referencia (línea roja) para indicar esos párrafos con una densidad superior al promedio obtenido en ese nivel académico. Los párrafos con una densidad superior al promedio son párrafos que intuitivamente consideramos deficientes debido a que indicarían una densidad de información conceptual alta, de aquí que estaría en nuestro interés modificar estos párrafos y quizá dividirlos en dos diferentes. Así disminuiría la densidad de información conceptual: recordemos que un texto con menor información conceptual es un texto más comprensible de acuerdo con [11]. La Figura 6.6b, ilustra los resultados de la evaluación para un documento extenso de nivel doctoral con un número mayor de párrafos. En este caso, tanto la Figura 6.6a, como la Figura 6.6b, deberían modificar el párrafo 2, ya que aparece por arriba del umbral de referencia.




(a) Gráfica de barras de la medidad de Dispersión por párrafo (documento de TSU)




(b) Gráfica de barras de la medida de Dispersión por párrafo (documento de Doctorado)

Figura 6.6: Representación visual para la medida de dispersión a través de gráficas de barras (TSU Y Doctorado)

### 6.2.2. Ejemplo de Retroalimentación Visual en un Documento de Doctorado

Incluimos aquí dos ilustraciones: una nube de términos para conceptos y un grafo que muestra la relación entre los párrafos clave y sus conceptos asociados. Las ilustraciones pertenecen a un documento de doctorado seleccionado aleatoriamente y las incluimos para que se pueda realizar una comparación antagónica entre el ejemplo de retroalimentación visual de un documento de TSU que se mostró en la Sección 6.2.1 y, el de Doctorado, que se muestra a continuación. Al observar la Figura 6.7 podemos concluir que el número de conceptos que aparece en la nube de términos, es mayor que el que aparece en la Figura 6.1a, así mismo somos capaces de ver una variabilidad marcada entre los tamaños de fuente de los diferentes conceptos que apreciamos.

Con base en la Figura 6.8, el grafo que visualizamos es uno más elaborado. Tan es así, que la cantidad de "clusters" que aparecen, nos indican diferentes núcleos de relaciones entre conceptos. Con esta imagen se busca mostrar la diferencia en cantidad relaciones entre conceptos. En la Figura 6.1b, aparecía solamente un clúster de conceptos relacionados, es claro cómo una diferencia de nivel entre dos secciones de documentos de tesis, se puede visualizar a través de nuestro proceso de retroalimentación visual. Finalmente, aclaramos que en la sección previa, pudimos apreciar un ejemplo de una gráfica de barras para un texto de nivel doctoral (Figura 6.6b). Si la comparamos con la Figura 6.6a, se observa la diferencia de párrafos para los dos secciones, ya que la de nivel doctoral cuenta con doce y la de nivel técnico con solamente dos.

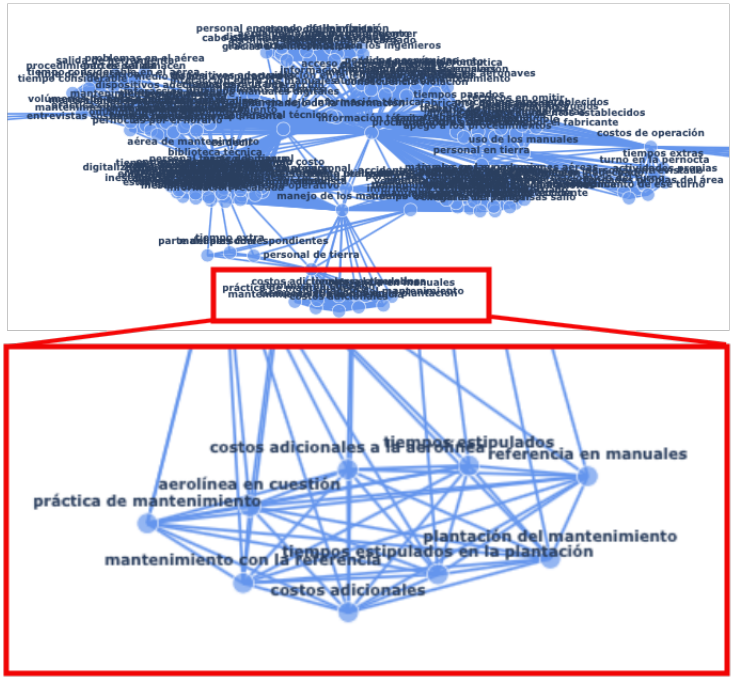



Figura 6.8: Grafo de Documento para una Tesis de Nivel Doctorado

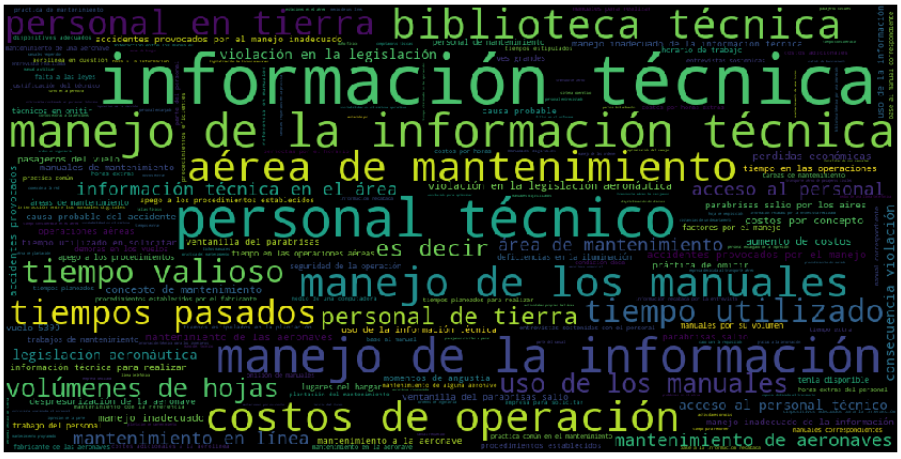



Figura 6.7: Nube de términos para conceptos

### 6.2.3. Análisis sobre Retroalimentación Visual

Las nubes de términos desplegadas en diferentes etapas del análisis permiten visualizar conceptos en variedad de perspectivas. Inicialmente, denotando aquellos conceptos extraídos que son más mencionados y luego reflejando los que mayor valor poseen tanto de Carga de Comprensibilidad como de Conectividad.

Los grafos de conceptos desplegados también muestran cómo el análisis de la Comprensibilidad progresa, yendo desde cómo los conceptos se relacionan, hasta llegar a grafos más elaborados que nos muestran conceptos clave en cada uno de los párrafos y el grafo general del documento con el que somos capaces de obtener las medidas de Conectividad y Dispersión. Es importante señalar que estos grafos se generan con una herramienta interactiva donde el usuario (autor del texto) puede acercarse y alejarse para ver en detalle los diferentes conceptos así como sus relaciones. Esto se ilustra a través de la caja que aparece en la parte inferior izquierda de la Figura 6.8, donde se hace un acercamiento hacia uno de los párrafos clave que aparece en el grafo.

Finalmente, las barras son utilizadas para desplegar la evaluación del valor de la dispersión en los párrafos, ordenadas en orden descendente para que aquellas con valores más altos se identifiquen con claridad. Esto aunado a la línea de referencia, indica inmediatamente aquellos párrafos que necesitan mejora.

# Capítulo 7

# Metodología en Legibilidad

Como se mencionó anteriormente, este trabajo de investigación sobre legibilidad consiste en la creación de un Clasificador/Comparador y, con base en éste, la creación, experimentación y selección de Evaluadores. A continuación describimos el proceso de construcción del Clasificador/Comparador, basado en [35]. Así mismo, y después de la presentación del Clasificador/Comparador se detalla cómo se utiliza el Clasificador/Comparador para construir los diferentes Evaluadores que utilizamos.

## 7.1. Clasificador/Comparador

El Clasificador/Comparador que se construye en [35] es capaz de determinar, entre un par de textos, cuál de ellos es más o menos legible. Este Clasificador/Comparador se genera a través de aprendizaje de máquina de la siguiente manera.

### 7.1.1. Representación de Documentos

Para entrenar al modelo y obtener el Clasificador/Comparador, primero se obtuvieron datos de entrenamiento representados de la siguiente manera: A partir de dos textos $a, b \in S$, donde $S$ es una conjunto de textos, la representación vectorial de los textos $V_a$, $V_b$ se construye como se muestra en en la parte superior de la Figura 7.1. Ahí podemos observar que cada vector ya sea $V_a$ ó $V_b$ posee sus características locales (frecuencia relativa) y globales (familiaridad). La frecuencia relativa corresponde a la frecuencia de cada palabra dividida por la frecuencia del número de palabras en el texto y la frecuencia global corresponde al logaritmo de la frecuencia que aparece en la lista de las 5000 (Se tuvo la opción entre 1000, 5000 y 10000; se escogieron las 5000 debido a que estaba en el medio de las opciones pero está en nuestro interés experimentar con este número y observar efectos.) palabras más comunes del Español[1]. Cabe hacer notar que la notación $V_{ab}$ corresponde al resultado de los vectores $V_a$ y $V_b$ ya operados. Manejamos tres tipos de operación: conjunción (diferencia), división y la concatenación, la primera refiriéndose a una resta vectorial (entrada a entrada), la segunda siendo la división entrada a entrada de los vectores y la tercera a una concatenación del vector $V_a$ con $V_b$ como se muestra en la parte inferior de la Figura 7.1.

[1] [31] es un corpus de datos creado por la Real Academia Española que provee una estadística de las palabras más utilizadas en el Español.

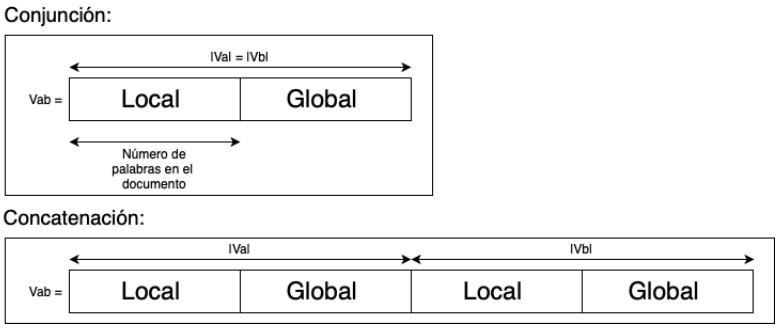



Figura 7.1: Representación para documentos en legibilidad

### 7.1.2. Entrenamiento del Clasificador/Comparador

Una vez con la representación para los textos se procede al entrenamiento del Clasificador/Comparador. En la Figura 7.2 podemos observar que los datos de entrenamiento que se alimentan a la máquina de soporte vectorial (SVM por sus siglas en inglés) corresponden a vectores del tipo $V_{ab}$, mencionados anteriormente que son el resultado de la operación del vector $V_a$ con el vector $V_b$. En específico, la imagen refiere a una operación que llamamos conjunción (diferencia).

Los datos de entrenamiento provienen del conjunto de documentos de texto de dos niveles académicos diferentes, llamémosles $A$ y $B$ con $n$ cantidad de documentos cada uno, esto para facilitar la explicación aunque no sea forzosamente así. Cada vector generado en $A$ se opera con cada uno de los vectores que genera $B$ y es clasificado como $+1$ si $A$ es más difícil que $B$ o $-1$ en el caso contrario; es importante aclarar que consideramos un vector más difícil que otro dependiendo de su nivel académico, i.e. el vector generado por un texto de TSU es más sencillo que uno de Doctorado. Realizamos este proceso con $A$ primero (generando vectores de la forma $V_{ab}$) y después con $B$ primero (generando vectores de la forma $V_{ba}$), ya que un vector $V_{ab}$ no tiene la misma clasificación que un vector $V_{ba}$. Al terminar el proceso, se observa que se termina con una matriz de vectores operados de $2 \times n \times n$, ésta servirá como nuestra matriz de entrenamiento para el modelo de aprendizaje de máquina seleccionado, en el caso específico de la Figura 7.2, una SVM.

Una vez finalizado el entrenamiento del modelo y teniendo un par de documentos de texto, obtenemos primero su representación, los operamos, se introducen al modelo predictivo y finalmente éste regresa una clasificación sobre cuál de ellos es más o menos legible que el otro. La Figura 7.2 muestra el proceso gráficamente.

## 7.2. Evaluador

La creación de los clasificadores es la primera parte del proceso, la segunda y nuestra mayor aportación en Legibilidad, consiste en la creación de Evaluadores que sean capaces de valorar el nivel del texto que le introduzcamos. Se construyeron cuatro tipos de evaluadores diferentes: Aleatorio, Centroide, Máximo y Mínimo. Estos evaluadores llevan el nombre debido a los vectores referentes que utilizan para operar el documento que queremos evaluar.

Ahora, el evaluador Aleatorio se construye para analizar si un documento aleatorio de un cierto nivel académico es capaz de acotar tanto superior, como inferiormen-

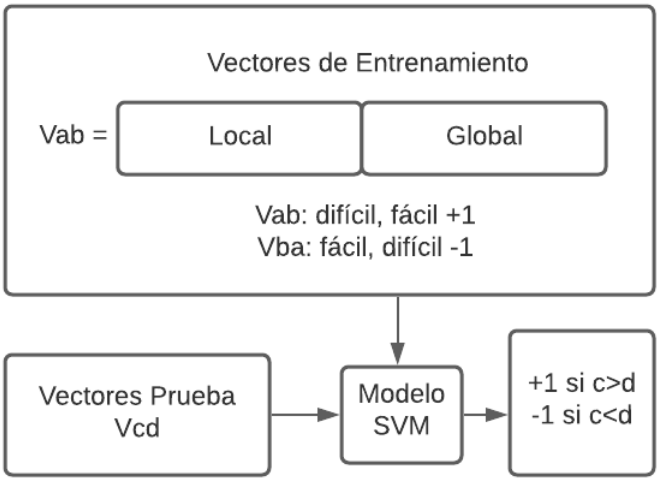



Figura 7.2: Diagrama de entrenamiento del Clasificador/Comparador

te un nivel académico contiguo. El evaluador Centroide, se construye debido a que este tipo de evaluador nos produce un referente promedio de todos los documentos, este vector debe ser más representativo del conjunto en general y nos interesa ver sus resultados, los cuales intuitivamente deben ser buenos tanto para acotar superior como inferiormente. El evaluador Máximo, se construye para experimentar que, efectivamente, es un evaluador efectivo para acotar superiormente, sin embargo, es también de nuestro interés observar si este evaluador nos sirve para acotar inferiormente. Finalmente, el evaluador mínimo se construye para observar si, en efecto, es una cota efectiva inferior, pero también nos interesa observar cómo acota superiormente. Estos dos últimos evaluadores pueden ser de gran utilidad si funcionan para las cotas inversas a las que fueron originalmente construidos. A continuación, describimos la metodología para cada uno de ellos.

### 7.2.1. Aleatorio

El evaluador Aleatorio se llama así debido a que se escoge un referente aleatorio para operar contra el documento que buscamos evaluar. El evaluador funciona de la siguiente manera, en primer lugar se tiene un documento que queremos evaluar, de aquí que primero lo llevamos a su representación vectorial. Ya con dicha representación, escogemos un documento aleatorio entre nuestro conjunto de documentos de un cierto nivel académico y también lo llevamos a su respectiva representación vectorial. Con éstos, los operamos, introducimos el vector resultante en nuestro modelo entrenado (Clasificador/Comparador) y así obtenemos la información de cuál de ellos es más legible que el otro.

Con esta información podemos acotar el nivel académico de nuestros documentos por evaluar, e.g. si introducimos un documento por evaluar de nivel licenciatura y lo operamos en contra de uno de Maestría, esperaríamos que resultara de menor nivel, contrariamente para uno de Técnico Superior Universitario donde esperaríamos que resultara de mayor nivel. La siguiente figura ilustra el evaluador Aleatorio y la utilizaremos para la explicación de los demás evaluadores ya que los cambios son sutiles, sin embargo, si hay interés por las imágenes específicas de cada evaluador,

se pueden encontrar en el apéndice del trabajo de investigación.

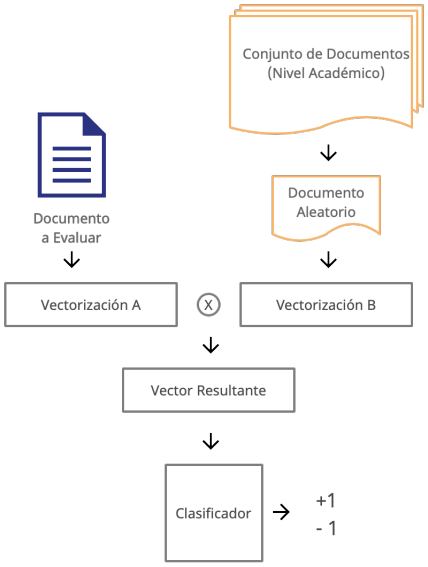



Figura 7.3: Esquema para el evaluador Aleatorio

### 7.2.2. Centroide

En el evaluador Centroide, la elección del referente para llevar a cabo la operación con un documento nuevo, es a través de la obtención del centroide del conjunto de documentos representados vectorialmente de un cierto nivel académico. Todo lo demás del evaluador funciona de la misma forma que se explicó con el evaluador Aleatorio, un esquema en específico para este evaluador aparece en la Figura A.1 en el apéndice de nuestro trabajo.

### 7.2.3. Máximo

En el evaluador basado en máximos, el vector referente para llevar a cabo la operación proviene del documento máximo o de mayor dificultad en el conjunto de documentos de un determinado nivel académico. Tanto el documento máximo como el mínimo se obtienen por medio de la utilización del Clasificador/Comparador, de tal forma que seleccionamos el primer o el segundo vector de la colección como máximo o mínimo (esto dependiendo de si el primero o el segundo es más difícil que el otro) y esta selección se compara con todos los demás vectores en la colección, si alguno resulta más difícil o menos difícil (esto a través del Clasificador/Comparador), es nuestro nuevo máximo o mínimo; el algoritmo en profundidad se puede encontrar en el apéndice como Algoritmo 1. Una vez con el documento más difícil, el evaluador funciona como lo explicado para el evaluador Aleatorio, un esquema en específico para el evaluador máximo se encuentra en la Figura A.3 que aparece en el apéndice del trabajo.

### 7.2.4. Mínimo

El último evaluador que se utilizó se basa en el documento más fácil (mínimo) de una colección de documentos. El algoritmo que se utiliza para la selección de éste es el que se explica para el evaluador Máximo y aparece en el apéndice de este trabajo como Algoritmo 1. Una vez con el mínimo el evaluador funciona de forma similar al evaluador Aleatorio; su esquema aparece en la Figura A.2 en el Apéndice.

# Capítulo 8

# Resultados para Legibilidad

## 8.1. Clasificadores/Comparadores

### 8.1.1. Experimentos

Contamos con cuatro diferentes niveles académicos: Técnico Superior Universitario, Licenciatura, Maestría y Doctorado. Como el Clasificador/Comparador se construye con dos niveles académicos tenemos la posibilidad de construir seis Clasificadores/Comparadores diferentes para cada sección éstos son: TSU-Licenciatura, TSU-Maestría, TSU-Doctorado, Licenciatura-Maestría, Licenciatura-Doctorado y Maestría-Doctorado. Estos Clasificadores/Comparadores, son capaces de decidir entre dos documentos de entrada, cuál de ellos es más o menos legible. Y una vez construidos, estos se ocuparán para la realización de evaluadores, los cuales son máquinas que reciben un documento de entrada y son capaces de acotar tanto superior como inferiormente el nivel académico de un texto.

Para cada una de las secciones (Planteamiento del Problema, Justificación y Resultados) somos capaces de construir seis Clasificadores/Comparadores. Esto lo hacemos debido a que está en nuestro interés construir diferentes Clasificadores/Comparadores para diferentes secciones del documento y así tener Clasificadores/Comparadores especializados en ciertas secciones. Así mismo, tenemos curiosidad de observar si alguna sección en específico produce mejores Clasificadores/Comparadores que otra. Entonces, por cada sección y para cada Clasificador/Comparador utilizamos diferentes modelos de aprendizaje de máquina. A través de [35], sabíamos que las SVMs eran nuestra mejor apuesta, pero se decidió comparar con KNN y Perceptrón también.

Por otro lado, sabemos que la generación de nuestros vectores de entrenamiento se puede realizar a través de tres operaciones: Diferencia, División y Concatenación. Está en nuestro interés comparar estas operaciones para poder así observar variación de resultados y rendimiento de las mismas. De hecho, los resultados de las tablas que mostraremos a continuación se dividen con base en estas tres operaciones y son los resultados de nuestros experimentos. Las tablas son 8.1, 8.3 y 8.2; y cada una de éstas muestra los resultados de los tres experimentos realizados con las diferentes operaciones.

Específicamente cada resultado que aparece en cada tabla se refiere a la exactitud (ver sección 2.7.4) de cada Clasificador/Comparador construido a través de un cierto modelo de aprendizaje de máquina. Cada Clasificador/Comparador se entrenó

con datos generados a través del nivel académico que lo nombra, de tal forma que para entrenar el Clasificador/Comparador TSU-Licenciatura primero se generó una matriz con el cojunto de documentos de TSU y Licenciatura a través de la metodología de entrenamiento que aparece en el Capítulo 7, y una vez con esta matriz los datos se divieron aleatoriamente en un 80 % para entrenamiento y un 20 % para evaluación (datos prueba). Los resultados que se muestran en las tablas refieren al resultado para un cierto modelo entrenado con un cierto par de conjuntos de datos de un cierto nivel académico evaluado con sus datos prueba.

### 8.1.2. Resultados para Clasificadores/Comparadores Operados con Diferencia

Con base en la Tabla 8.1 se tiene que el Clasificador/Comparador con peor valor de exactitud es TSU-Licenciatura, esto para las tres secciones analizadas de los documentos de tesis. El mejor resultado lo tiene el Clasificador/Comparador de Licenciatura-Doctorado, seguido así por TSU-Doctorado. En cuanto a los promedios, SVM tiene los mejores resultados para las tres secciones. En específico, para las secciones de Planteamiento del Problema y Justificación, la diferencia puede llegar hasta 6.35 puntos porcentuales mientras que para la sección de Resultados la diferencia se achica y es de 0.73 puntos porcentuales.

El hecho de que los peores resultados aparezcan para TSU-Licenciatura no es de sorprendernos. Este tipo de Clasificador/Comparador se entrena y evalúa con niveles académicos contiguos, e intuitivamente debe ser más complicado distinguir documentos académicos con niveles académicos contiguos. Uno esperaría resultados similares en Licenciatura-Maestría y en Maestría-Doctorado, y sí es el caso, ya que por cada sección, observamos que los Clasificadores/Comparadores mencionados tienen los peores resultados. Contrario a esto, esperamos obtener los mejores resultados para Clasificadores/Comparadores construidos a través de niveles académicos distantes, y así el caso a lo largo de las secciones, ya que TSU-Doctorado y Licenciatura-Doctorado tienen los primeros lugares en resultados. Finalmente, y como era de esperarse, las SVMs resultan con mejores resultados sobre los otros modelos y esto lo sabíamos a través de trabajo de [35], pero también es conocido en la teoría de aprendizaje de máquina que SVM es un modelo poderoso para clasificación de texto.

| Clasificador/Comparador | Problema | | | Justificación | | | Resultados | | |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón |
| **TSU-Licenciatura** | 36.65 | 41.59 | 37.39 | 64.70 | 53.61 | 63.87 | 46.38 | 56.26 | 42.41 |
| **TSU-Maestría** | 54.41 | 55.55 | 53.18 | 61.99 | 64.93 | 56.97 | 64.14 | 62.17 | 69.43 |
| **TSU-Doctorado** | 81.09 | 55.46 | 82.14 | 83.58 | 67.69 | 73.33 | 82.14 | 65.87 | 81.15 |
| **Licenciatura-Maestría** | 57.53 | 47.42 | 47.81 | 70.84 | 50.34 | 65.31 | 77.67 | 74.96 | 81.73 |
| **Licenciatura-Doctorado** | 90.58 | 77.64 | 87.64 | 90.58 | 77.64 | 87.64 | 86.11 | 62.64 | 85.64 |
| **Maestría-Doctorado** | 77.64 | 67.64 | 64.11 | 77.64 | 67.64 | 64.11 | 74.35 | 53.20 | 75.40 |
| **Promedio** | 66.31 | 57.55 | 62.04 | 74.88 | 63.64 | 68.53 | 71.79 | 62.51 | 72.62 |

Tabla 8.1: Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a Diferencia.

### 8.1.3. Resultados para Clasificadores/Comparadores Operados con División

Con base en la Tabla 8.2 tenemos que el Clasificador/Comparador que peores resultados posee es TSU-Licenciatura, esto para las tres secciones analizadas de los documentos de tesis. El mejor resultado lo tiene el Clasificador/Comparador de TSU-Doctorado, seguido así por Licenciatura-Doctorado. En cuanto a los promedios, SVM tiene los mejores resultados para las tres secciones, con una diferencia de hasta 9.33 puntos porcentuales para las secciones de Planteamiento del Problema y Justificación, mientras que para la sección de Resultados la diferencia es de 0.53 puntos porcentuales.

Al igual que con los resultados para el operador de diferencia, vemos un comportamiento similar de los resultados, donde los Clasificadores/Comparadores con peores resultados son los construidos con niveles académicos colindantes y los de mejores resultados los desarrollados a través de niveles académicos distantes. De nuevo, esto intuitivamente lo esperábamos debido al hecho de que es más sencillo identificar diferencias de legibilidad en textos con diferencias de nivel académico más distantes y nuestros modelos son capaces de identificar estas variaciones en el entrenamiento. Las SVMs, al igual que con el operador de diferencia, son los modelos que obtienen los mejores resultados y esto se debe a que son modelos poderosos para clasificación de texto debido a su amplia capacidad en el manejo de múltiples características. La sección con mejores resultados, en promedio, es la de Resultados y esto se podría explicar debido a que son los documentos más extensos.

| Clasificador/Comparador | Problema | | | Justificación | | | Resultados | | |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón |
| **TSU-Licenciatura** | 48.73 | 53.15 | 42.33 | 54.97 | 49.69 | 50.00 | 41.97 | 56.61 | 43.38 |
| **TSU-Maestría** | 58.61 | 51.06 | 61.31 | 59.35 | 52.79 | 58.74 | 65.45 | 64.74 | 66.59 |
| **TSU-Doctorado** | 65.33 | 56.51 | 50.00 | 71.53 | 60.00 | 50.00 | 82.14 | 64.88 | 78.67 |
| **Licenciatura-Maestría** | 57.93 | 45.73 | 55.65 | 59.25 | 49.13 | 61.24 | 72.47 | 71.40 | 71.54 |
| **Licenciatura-Doctorado** | 65.81 | 58.16 | 66.83 | 73.52 | 50.58 | 50.00 | 82.25 | 68.05 | 83.48 |
| **Maestría-Doctorado** | 62.50 | 59.72 | 50.00 | 57.35 | 44.41 | 50.00 | 69.55 | 53.20 | 73.39 |
| **Promedio** | 59.81 | 54.05 | 54.35 | 62.66 | 51.10 | 53.33 | 68.97 | 63.14 | 69.50 |

Tabla 8.2: Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a División.

### 8.1.4. Resultados para Clasificador/Comparadores Operados con Concatenación

Con base en la Tabla 8.3 se observa que el Clasificador/Comparador con los peores resultados es el de TSU-Licenciatura, con resultados de exactitud tan bajos como de 35.08. El mejor Clasificador/Comparador lo encontramos en Licenciatura-Doctorado con una exactitud de 88.23, a este le sigue TSU-Doctorado, donde su mejor resultado está en 87.69. En cuanto a los promedios, para la sección de Planteamiento del Problema el mejor modelo de entrenamiento corresponde al del Perceptrón, para Justificación corresponde a SVM y para Resultados también.

Los resultados son similares a los observados con los operadores de diferencia y división, sin embargo, encontramos diferencias sutiles e interesantes. Una es el caso

de que aquí el Clasificador/Comparador con mejores resultados no es Licenciatura-Doctorado, sino TSU-Doctorado, esto valida la hipótesis que hemos venido manejando en donde Clasificadores/Comparadores creados a través de niveles académicos distantes, generarán mejores resultados. El segundo aspecto es que el modelo de SVM no resulto ganador a lo largo de todas las secciones, sin embargo, y considerando las tablas para los operadores de diferencia y división, en general tiene los mejores resultados. Finalmente la sección con los resultados más altos para promedio es la de Justificación, y lo mismo ocurrió para el operador de diferencia, esto no es lo que se esperaría intuitivamente ya que no es la sección que contiene documentos de texto más extensos.

| Clasificador/Comparador | Problema | | | Justificación | | | Resultados | | |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón | SVM | KNN | Perceptrón |
| **TSU-Licenciatura** | 35.08 | 44.53 | 41.38 | 64.55 | 48.26 | 65.38 | 45.41 | 61.19 | 44.22 |
| **TSU-Maestría** | 52.77 | 56.29 | 61.35 | 61.76 | 53.77 | 56.90 | 65.65 | 66.84 | 62.43 |
| **TSU-Doctorado** | 79.83 | 57.14 | 86.97 | 83.58 | 68.71 | 87.69 | 80.35 | 68.25 | 50.00 |
| **Licenciatura-Maestría** | 57.34 | 52.08 | 68.45 | 71.19 | 47.75 | 72.31 | 80.74 | 67.29 | 60.23 |
| **Licenciatura-Doctorado** | 74.48 | 40.81 | 64.79 | 88.23 | 75.29 | 80.58 | 80.24 | 66.35 | 50.00 |
| **Maestría-Doctorado** | 69.04 | 50.79 | 79.16 | 78.82 | 61.17 | 64.41 | 67.78 | 56.25 | 50.00 |
| **Promedio** | 61.42 | 50.27 | 67.01 | 74.68 | 59.15 | 71.21 | 70.02 | 64.36 | 52.81 |

Tabla 8.3: Porcentaje de exactitud para Clasificadores/Comparadores con tres modelos diferentes, SVM, KNN y Perceptrón sobre las tres secciones seleccionadas. La representación de entrenamiento utilizada en los Clasificadores/Comparadores corresponde a Concatenación.

### 8.1.5. Conclusiones para Clasificadores/Comparadores

Los comportamientos que observamos para los resultados de las tres tablas 8.1, 8.2 y 8.3 son los siguientes: El mejor Clasificador/Comparador es el de Licenciatura-Doctorado y le sigue TSU-Doctorado. El peor Clasificador/Comparador lo encontramos en TSU-Licenciatura. Este tipo de comportamiento se explica a través de la idea que los Clasificadores/Comparadores entrenados con niveles académicos "distantes" clasifican con mayor exactitud que los que tienen niveles académicos contiguos, e.g. TSU-Licenciatura tiene peores resultados de clasificación que TSU-Doctorado. Intuitivamente, como humanos, somos mejores para distinguir el nivel de dificultad entre dos textos si éstos tienen niveles académicos más distantes, sin embargo, textos con niveles académicos contiguos, ciertamente nos sería más complicado tomar una decisión.

En cuanto a los promedios, podemos observar que el operador de diferencia nos arroja los mejores resultados, seguido de concatenación y por último se tiene al de división. Esto lleva sentido con lo reportado en [35] y es importante aclarar aquí que el operador de concatenación arroja vectores del doble de tamaño que los otros dos operadores y esto alargaba temporalmente el entrenamiento de los Clasificadores/Comparadores cuantiosamente. Obtuvimos los resultados para los experimentos pero no vale la pena continuar trabajando con este operador ya que no genera mejores resultados. Es claro que nuestra elección de operador apunta hacia diferencia.

En cuanto a los modelos de entrenamiento, para la operación de diferencia, SVM resultó mejor en las tres secciones. Para la operación de división, SVM resultó mínimamente debajo de Perceptrón en la sección de Resultados, mientras que arriba para las otras dos secciones. Finalmente para la operación de concatenación SVM

resulta mejor para la sección de Justificación y por debajo de Perceptrón para la sección de Planteamiento del Problema. SVM, en general resulta el mejor modelo de aprendizaje de máquina para entrenar a nuestros Clasificadores/Comparadores y esto se debe al poder que tiene para trabajar de manera más efectiva en espacios de altas dimensionalidades, esto incluso cuando son de mayor tamaño que el número de datos de entrenamiento.

## 8.2. Evaluadores

### 8.2.1. Experimentos

Los evaluadores acotan tanto superior como inferiormente el nivel académico de un documento, y esto se logra a través de la utilización de un Clasificador/Comparador, ya que como sabemos, éstos nos regresan la dificultad relativa entre pares de documentos. La idea tentativa, a largo plazo, es que una máquina de este tipo nos sirve para poder obtener una valoración del nivel académico de un texto. A continuación explicamos los experimentos realizados para la selección de evaluadores efectivos.

Nuestros experimentos se concentran sobre la evaluación de documentos de nivel licenciatura, esto por dos razones: en primer lugar porque estos documentos están en un nivel académico que se encuentra entre los otros niveles académicos de estudio y, en segundo, porque debido a que tenemos miras de utilizar estos evaluadores en una aplicación web, tendría mayor impacto, en un principio, hacerlo para personas cursando una licenciatura que una maestría. Para nuestros experimentos, apartamos aleatoriamente un 20 % de documentos de Licenciatura aleatoriamente; es importante aclarar que estos documentos no se utilizaron para el entrenamiento de los Clasificadores/Comparadores.

En nuestros experimentos se busca analizar la capacidad de los evaluadores para acotar tanto inferior como superiormente el nivel académico de los documentos de Licenciatura. Para acotarlos superiormente, se seleccionan Clasificadores/Comparadores que hayan sido entrenados con niveles académicos superiores o iguales, por ejemplo, Licenciatura-Maestría, así mismo estos Clasificadores/Comparadores deben poseer los mejores resultados de la tabla 8.1. Los Clasificadores/Comparadores seleccionados para experimentar fueron tanto Licenciatura-Maestría como Licenciatura-Doctorado, el último debido a que su clasificación sobresalía notablemente de los otros. El modelo de aprendizaje de máquina utilizado se basó en el que tuvo la mejor exactitud para un respectivo Clasificador/Comparador, así que podía variar entre SVM, KNN y Perceptrón. En la tabla de resultados del experimento se aclara cuál de estos se utilizó en específico. En cuanto a la selección de Clasificadores/Comparadores para las cotas inferiores, se seleccionaron los TSU-Maestría, esto debido a sus resultados superiores comparados con TSU-Licenciatura como se observa en la tabla 8.1. Con base en los operadores y debido a los resultados de los Clasificadores/Comparadores, se utilizó el operador de diferencia para realizar todos nuestros experimentos.

Ahora, recordemos que tenemos cuatro evaluadores: Aleatorio, Centroide, Máximo y Mínimo. Cada uno de éstos, lleva su nombre debido a que de una serie de documentos representados en forma vectorial, obtiene un vector representativo que lleva el nombre del evaluador: aleatorio, centroide, máximo o mínimo; estos son

vectores referentes a ese conjunto de vectores, y fungen como vectores que representan a ese nivel académico. Utilizaremos estos evaluadores junto con estos vectores representativos para acotar superior e inferiormente el nivel académico de un documento de entrada. Los evaluadores en conjunto con su Clasificador/Comparador; ya sea Licenciatura-Maestría o Licenciatura-Doctorado para las cotas superiores o TSU-Maestría para las cotas inferiores acotan correcta o incorrectamente el nivel académico de un documento de Licenciatura.

Para clasificar un documento como acotado correctamente, el documento (que pertenece al nivel académico de licenciatura) de entrada se vectorizará y se operará con un vector referente de un cierto nivel académico, una vez operados se recurre al Clasificador/Comparador para saber cuál de los dos es más o menos legible que el otro, y con esta información si el documento efectivamente es menos legible que el otro la cota se toma como correcta o incorrecta, dependiendo si es documento es efectivamente de menor o mayor nivel que el otro. Por ejemplo, si opero mi documento de entrada con el referente aleatorio de Maestría y obtengo que mi documento de entrada es menos legible que el de Maestría, eso se categorizará como una predicción incorrecta y en el caso contrario, como una predicción correcta. Esto se explica a través del hecho de que un documento de Licenciatura es de menor nivel que uno de Maestría. Este proceso lo realizaremos con cada uno de los niveles académicos y para cada uno de los posibles referentes (evaluadores), obteniendo así tablas como la Tabla 8.7.

A continuación se presentan los resultados de los experimentos para los evaluadores. En cada una de las tablas se subrayan los referentes de interés, i.e. si se están estudiando resultados con cotas superiores, nos interesan los resultados con referentes superiores al nivel académico del documento a evaluar. Esto debido a que estudiamos evaluadores con Clasificadores/Comparadores que intuitivamente deben ser mejores para acotar superiormente (Licenciatura-Maestría y Licenciatura-Doctorado), pero también estudiamos evaluadores para acotar inferiormente (TSU-Licenciatura). De aquí que para cotas superiores nos interesa ver cómo acotan los referentes de Doctorado y Maestría, mientras que al experimentar con cotas inferiores nos interesa ver el comportamiento los referentes de Técnico Superior Universitario. Es a través de estos experimentos que seremos capaces de seleccionar el mejor evaluador (referente en combinación con su Clasificador/Comparador) para acotar tanto superior, como inferiormente un documento de Licenciatura.

### 8.2.2. Resultados para la sección de "Planteamiento del Problema

#### Cotas Superiores con Clasificador/Comparador Licenciatura-Maestría y Licenciatura-Doctorado

A partir de las tablas 8.4 y 8.5, podemos observar lo siguiente. La tabla 8.4 muestra que el referente de maestría acota de forma esperada los documentos de licenciatura, sin embargo, para el referente de doctorado se comporta fuera de lo esperado. El referente de licenciatura muestra comportamientos erráticos, como era de esperarse y sorpresivamente (debido a que el entrenamiento del Clasificador/Comparador utilizado fue con Licenciatura-Doctorado) el referente de Técnico Superior Universitario acota de forma efectiva inferiormente.

En los resultados de la tabla 8.5, observamos que tanto el referente de Maestría como el de Doctorado se comportan de forma esperada, pero el de Doctorado acota más efectivamente. El referente de Licenciatura, como era esperado es errático y técnico superior universitario tiene resultados aceptables, sin embargo, son mejores los resultados de la tabla 8.4.

En cuanto a los evaluadores, el peor evaluador para generar una cota superior es el Mínimo y le sigue Centroide. Finalmente, tanto Aleatorio como Máximo muestran ser cotas superiores efectivas y se podría tomar cualquiera de ellos.

Tabla 8.4: Resultados para evaluadores con posibles referentes y modelo SVM

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 92 % | 76 % | 100 % | 100 % |
| Licenciatura | 0 % | 7 % | 0 % | 23 % |
| Maestría | 100 % | 100 % | 92 % | 100 % |
| Doctorado | 38 % | 23 % | 7 % | 92 % |

Tabla 8.5: Resultados de evaluadores con posibles referentes y modelo SVM

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 69 % | 69 % | 100 % | 0 % |
| Licenciatura | 15 % | 7 % | 0.0 % | 38 % |
| Maestría | 100 % | 76 % | 0 % | 100 % |
| Doctorado | 100 % | 100 % | 100 % | 100 % |

### Cotas Inferiores para Clasificador/Comparador TSU-Maestría con modelo SVM

Observaciones referentes a la tabla 8.6 indican que con el referente de Técnico Superior Universitario uno es capaz de acotar inferiormente a los documentos de licenciatura con todos los evaluadores excepto con el Máximo, esto llevando sentido ya que es el evaluador que selecciona al documento más difícil entre todo el conjunto de posibles referentes. El referente de Licenciatura es errático para acotar, Maestría obtiene los mejores resultados como cota superior y Doctorado obtiene resultados bajos, un resultado que no teníamos esperado.

Tabla 8.6: Resultados de evaluadores con diferentes referentes y modelo KNN

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 100 % | 100 % | 100 % | 15 % |
| Licenciatura | 15 % | 0 % | 0 % | 100 % |
| Maestría | 100 % | 100 % | 92 % | 100 % |
| Doctorado | 61 % | 53 % | 0 % | 100 % |

### 8.2.3. Justificación

**Experimentos para Cotas Superiores con Clasificadores/Comparadores Licenciatura-Maestría y Licenciatura-Doctorado**

A continuación analizamos las tablas 8.7 y 8.8. La tabla 8.7 nos indica que el referente de Doctorado es un acotador efectivo, siendo su mejor evaluador el Máximo y su peor el Mínimo. El referente de Maestría es un evaluador aún más efectivo que el de Doctorado, un comportamiento similar al observado en la tabla 8.4 de la sección de Planteamiento del problema; en cuanto a sus evaluadores vemos que todos son efectivos y el de peor resultado es el evaluador Mínimo. El referente de Licenciatura muestra comportamientos erráticos como esperábamos y Técnico Superior Universitario se muestra efectivo para acotar inferiormente con los evaluadores Mínimo y Máximo pero no con Aleatorio y Centroide.

En la tabla 8.8 observamos que el referente de doctorado es un excelente acotador superior con cada uno de los evaluadores. En cuanto al referente de maestría observamos que es un acotador superior excelente con el evaluador Máximo y medianamente bueno con los evaluadores Aleatorio y Centroide, sin embargo, con Mínimo se tiene una exactitud del 0 %. El referente de Licenciatura se comporta erráticamente al igual que en la sección anterior y Técnico Superior Universitario es un acotador inferior efectivo con los evaluadores Mínimo y Centroide.

Tabla 8.7: Resultados para evaluadores con posibles referentes y modelo SVM

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 23 % | 29 % | 100 % | 100 % |
| Licenciatura | 11 % | 5 % | 0 % | 23 % |
| Maestría | 100 % | 100 % | 88 % | 100 % |
| Doctorado | 88 % | 94 % | 41 % | 100 % |

Tabla 8.8: Resultados para Evaluadores con posibles referentes y modelo SVM

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 35 % | 76 % | 100 % | 47 % |
| Licenciatura | 17 % | 23 % | 0 % | 54 % |
| Maestría | 65 % | 65 % | 0 % | 100 % |
| Doctorado | 100 % | 100 % | 100 % | 100 % |

**Resultados de Experimentos para Cotas Inferiores con Clasificador/Comparador de TSU-Maestría**

La tabla 8.9 muestra los resultados que hipotetizábamos, ya que el referente de Técnico Superior Universitario es el mejor acotador inferior que tenemos para los resultados de la sección de Justificación, siendo Centroide el peor evaluador entre todos los posibles. Licenciatura, como era de esperarse, muestra un comportamiento

errático, mientras que Maestría y Doctorado muestra resultados efectivos con los evaluadores Aleatorio y Máximo.

Tabla 8.9: Resultados de evaluadores con todos los referentes con Clasificador/Comparador TSU-Maestría y modelo KNN

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 100 % | 88 % | 100 % | 100 % |
| Licenciatura | 47 % | 17 % | 0 % | 100 % |
| Maestría | 100 % | 23 % | 29 % | 100 % |
| Doctorado | 52 % | 35 % | 11 % | 94 % |

### 8.2.4. Resultados

#### Resultados Experimentales de Cotas Superiores con Clasificadores/Comparadores de Licenciatura-Maestría y Licenciatura-Doctorado

A continuación analizamos los resultados para las tablas 8.10 y 8.11. La tabla 8.10 muestra comportamientos similares a los observados en las dos secciones anteriores. En cuanto al referente de Doctorado observamos que es más efectivo con el evaluador Máximo y menos efectivo con el Mínimo, así mismo el referente de Maestría tiene mejores resultados, con sus evaluadores comportándose de forma similar a excepción entre Aleatorio y Centroide. El referente de Licenciatura es errático, al igual que en las dos secciones anteriores y Técnico Superior Universitario puede ser un referente efectivo si utilizamos al evaluador Mínimo.

Observando la tabla 8.11 se puede concluir que el referente de Doctorado es un acotador superior efectivo con todos sus evaluadores por igual, mientras que el referente de Maestría es un acotador superior efectivo con los evaluadores Mínimo y Máximo. El referente de Licenciatura se comporta de forma errática, al igual que en las dos secciones anterior y Técnico Superior Universitario es un acotador inferior efectivo solamente para el evaluador Mínimo.

Tabla 8.10: Resultados para Evaluadores con sus posibles referentes y modelo Perceptrón

| **Referente** | **Aleatorio** | **Centroide** | **Mín.** | **Máx.** |
|---|---|---|---|---|
| TSU | 37 % | 29 % | 100 % | 0 % |
| Licenciatura | 29 % | 14 % | 0 % | 55 % |
| Maestría | 92 % | 96 % | 77 % | 100 % |
| Doctorado | 81 % | 88 % | 37 % | 100 % |

Tabla 8.11: Resultados para Evaluadores con sus posibles referentes y modelo SVM

| Referente | Aleatorio | Centroide | Mín. | Máx. |
|---|---|---|---|---|
| TSU | 48 % | 55 % | 100 % | 0 % |
| Licenciatura | 29 % | 18 % | 0 % | 77 % |
| Maestría | 51 % | 74 % | 100 % | 100 % |
| Doctorado | 100 % | 100 % | 100 % | 100 % |

**Resultados Experimentales de Cotas Inferiores con Clasificador/Comparador de TSU-Maestría y Modelo SVM**

Con base en los resultados mostrados en la tabla 8.12 podemos concluir que el referente Técnico Superior Universitario es el mejor acotador inferior que tenemos para esta sección y, en particular, su evaluador Mínimo es el más efectivo. El referente de Licenciatura se comporta medianamente bien, con el Máximo siendo su mejor evaluador. En cuanto al referente de Maestría, podemos observar que es un acotador superior efectivo para los evaluadores Máximo y Mínimo. Finalmente el referente de Doctorado es un excelente acotador superior con todos evaluadores.

Tabla 8.12: Resultados de Evaluadores con posibles referentes

| Referente | Aleatorio | Centroide | Mín. | Máx. |
|---|---|---|---|---|
| TSU | 81 % | 81 % | 100 % | 33 % |
| Licenciatura | 51 % | 59 % | 0 % | 100 % |
| Maestría | 100 % | 100 % | 66 % | 100 % |
| Doctorado | 81 % | 100 % | 44 % | 100 % |

## 8.3. Análisis Generales

Con base en las tablas presentadas anteriormente podemos concluir lo siguiente. Los entrenamientos para las cotas superiores son efectivos, y con base en el Clasificador/Comparador Licenciatura-Maestría, sería conveniente usar el referente de Maestría, ya que es más estable que el de Doctorado y tiene mejores resultados en general. Ahora, si quisiéramos ocupar el clasificador Licenciatura-Doctorado, nuestro referente de preferencia sería el de Doctorado, ya que tiene mejores resultados que Maestría y sus resultados son estables. En cuanto a acotar inferiormente, en las tres secciones observamos que la elección del Clasificador/Comparador de TSU-Maestría ha sido una buena elección, es tanto estable como efectivo.

La selección de los evaluadores pueden variar dependiendo de si queremos acotar superior o inferiormente. Si la idea es acotar superiormente, observamos que podemos elegir cualquier evaluador siempre y cuando no escojamos el Mínimo, esto es debido a que selecciona como referente el documento peor escrito y este puede ser de un nivel aún más bajo que uno de un nivel académico inferior, esto generando resultados negativos. En cuanto a acotar inferiormente, observamos gracias a los resultados, que está en nuestra conveniencia utilizar el evaluador Mínimo y no utilizar el evaluador

Máximo, ya que éste selecciona el documento de mayor nivel en nuestro conjunto de datos como referente y este documento puede ser de un nivel más alto que el documento a evaluar, afectando así nuestros resultados. Es importante recalcar aquí que aunque intuitivamente sabíamos que este comportamiento era esperado, incluir a este evaluador en este análisis permitió observar si distaba bastante en resultados de los otros evaluadores, ya que si este no fuera el caso es un evaluador que conviene utilizar debido a que genera una cota inferior con más cercanía al nivel académico de Licenciatura.

Es importante señalar que en los resultados para cotas superiores e inferiores de las tres secciones el referente de Licenciatura se comporta erráticamente. Esto es de esperarse ya que como habíamos mencionado estamos evaluando documentos de Licenciatura, así que para acotar un documento de Licenciatura tanto superior como inferiormente se necesitan documentos de menor o mayor nivel. Otra posible solución sería refinar nuestro proceso y generar un experimento donde acotemos superiormente con el documento máximo del mismo nivel pero de nuestro conjunto de documentos referentes e inferiormente con el documento mínimo del mismo nivel también de nuestros conjunto de documentos referentes, aún así creemos que en los resultados encontramos procesos que logran acotar con buena exactitud. Esto pasaría a considerar si acotamos superiormente, pero qué tan superiormente, e.g. no quisiéramos poner un documento completamente por encima del nivel de Licenciatura, queremos un documento que esté arriba de Licenciatura, sin embargo, que tenga la capacidad de definirme bien si se pasa a un siguiente nivel, como lo sería Maestría. Este tipo de sutilezas nos importan y en un trabajo a futuro sería interesante lograr generar experimentos para generarlas.

# Capítulo 9

# Conclusiones

## 9.1. Conclusiones en Comprensibilidad

A continuación damos las conclusiones derivadas del proceso de investigación que se realizó para Comprensibilidad. Primero presentamos el problema que se abordó y a través de qué técnicas se atacó, y es a través de ello que podremos señalar cuáles soluciones fueron efectivas y cuáles no. Después recalcamos los resultados principales de la investigación para después colocarlos en el contexto del área de investigación. Finalmente presentamos el trabajo a futuro basándonos en nuestros objetivos, continuación del proyecto, preguntas que surgieron en el proceso de investigación y posibles aplicaciones.

### 9.1.1. Evaluación del Proceso

Con base en Comprensibilidad el presente trabajo de investigación presentó métodos y técnicas, tanto de evaluación como de retroalimentación para documentos académicos. En primer lugar se creó un proceso de extracción de medidas de Comprensibilidad, el cual es capaz de extraer la Carga de Comprensibilidad, Conectividad y Dispersión en documentos académicos, en específico para tesis de grado. Con base en este proceso de extracción de métricas de Comprensibilidad fuimos capaces de obtener resultados para la realización de un evaluador de nivel académico basado en medidas de Comprensibilidad. Así mismo, y a través del proceso de extracción, se generó un esquema de retroalimentación visual para documentos académicos nuevos (borradores de tesis de grado).

El proceso de extracción de medidas de Comprensibilidad involucra dos partes. La primera se basa en el análisis de los conceptos, sus relaciones y la secuencia de éstos a lo largo del documento para obtener la Carga de Comprensibilidad. Esta medida resultó no ser la más confiable de los resultados ya que regresaba un valor de cero para la mayoría los documentos. Dado que la Carga de Comprensibilidad es una medida que requiere "cuerpo" de los documentos académicos, si contábamos con textos con poca información, esto implica documentos con pocos conceptos, relaciones disminuidas y esto acaba por resultar en valor nulo para la Carga de Comprensibilidad. Se concluye entonces que la medida de Carga de Comprensibilidad es poco práctica para evaluar la Comprensibilidad en tesis por secciones.

La segunda parte consta del cálculo de la Conectividad y la Dispersión, la cual se realiza a través de la construcción del Grafo de Conceptos. En este grafo se re-

presentan los conceptos, las relaciones entre ellos y el peso de estas relaciones. La estimación de la Conectividad fue sustancialmente mejor que la Carga de Comprensibilidad, sin embargo, el proceso para obtener Conectividad se basa en la relación entre los conceptos, de aquí que si no existen relaciones en los textos analizados, también puede regresar valores nulos. Finalmente, la Dispersión fue la medida que mejor logramos recabar y esto es porque mide la densidad de conceptos en los párrafos, de aquí que si teníamos párrafos y conceptos dentro de ellos, tendríamos un valor no nulo para esta medida. Tanto la Conectividad como la Dispersión son medidas viables para evaluar la Carga de Comprensibilidad en una tesis de grado, siendo la primera la mejor entre éstas dos.

Refiriéndonos ahora a las relaciones entre conceptos, esto afectó nuestra medición para Carga de Comprensibilidad y, hasta cierto punto a la medida de Conectividad. Al revisar los documentos, podíamos observar que algunos de los conceptos extraídos de cada documento aparecían una vez y no volvían a aparecer más adelante en el documento. Esto causa que no se generara una relación de los conceptos mencionados con otros que podrían aparecer en el documento, terminando en valores nulos tanto para la Carga de Comprensibilidad como para Conectividad. Una posible solución para esto es la necesidad de utilizar documentos que contengan amplias relaciones entre conceptos ya que si estos no las contienen, éstas medidas pueden no brindar la información suficiente. Así mismo, se podrían incluir implementaciones en nuestro método que logren capturar relaciones más profundas entre los conceptos.

Con referencia al proceso de retroalimentación visual para tesis de grado. Las nubes de términos son efectivas pero podrían ser visualizaciones sumamente densas que, en ocasiones no permitirían observar una clara diferencia entre términos con mayor frecuencia que otros, quizá en documentos con una densidad de conceptos alta sería conveniente escoger los $n$ conceptos más frecuentes. En cuanto a los grafos, su proceso de extracción en documentos con una densidad alta de relaciones puede ser sumamente lento, lo mismo pasa con la interacción del grafo generado; sería importante explorar bibliotecas diferentes para experimentar si es posible aumentar la rapidez de generación del grafo y la interacción. Por último, la gráfica de barras es tanto efectiva como rápida de obtener.

### 9.1.2. Conclusiones Principales

Los resultados obtenidos para las medidas de Comprensibilidad nos llevan a las siguientes conclusiones. La Dispersión es la medida que muestra un comportamiento con poca variabilidad a lo largo de todos los niveles académicos y las secciones de estudio, así mismo, es una medida fácil de recabar, y esto se concluye porque tiene un porcentaje perfecto de valores no-nulos. La Dispersión es una medida de confianza que podríamos utilizar para, en un futuro, evaluar la comprensibilidad de un documento en proceso (borrador). La Conectividad no alcanza a ser una medida confiable por dos cosas, su desviación estándar siempre se encuentra por debajo del promedio y no recabamos altos porcentajes de valores no-nulos. Se concluye entonces que esta medida, si se quisiera utilizar para la evaluación de Comprensibilidad en borradores no nos es útil. Finalmente, la Carga de Comprensibilidad, con base en los resultados obtenidos, se vuelve nuestra medida más débil. Se concluye que no es una medida utilizable, al menos por ahora, debido a su alta variabilidad y al bajo porcentaje de datos no-nulos que somos capaces de recabar de la misma. Sin em-

bargo, es importante aclarar que es la medida más interesante de seguir estudiando debido al profundo análisis que somete a un documento, borrador, tan es así que el mismo proceso de extracción de información nos provee información importante del borrador que se ocupa a través del proceso de retroalimentación visual.

Con base en el proceso de Retroalimentación Visual, concluimos entonces que las nubes de términos, gráfica de barras y grafos son efectivos. Las nubes de términos nos permiten visualizar los conceptos más frecuentes a lo largo del documento, los que tienen una Carga de Comprensibilidad más alta y los que poseen una Conectividad mayor. La gráfica de barras también cumple su función, que es el indicar qué párrafos contienen una dispersión por arriba de la media, está en nuestro interés disminuir la densidad de conceptos en estos párrafos, quizá al convertirlos en dos diferentes. Finalmente, los grafos son elementos útiles en la visualización de las relaciones existentes entre los conceptos mismos y hasta entre los conceptos y sus párrafos clave. Es importante aclarar que nuestro proceso de retroalimentación visual necesita una valoración experimental y esto se incluye en el trabajo a futuro.

### 9.1.3. Resultados en Contexto

Con los resultados que presentamos para Comprensibilidad, todavía no somos capaces, con certeza, de generar segmentos numéricos donde una cierta medida, para un cierto nivel académico debería pertenecer, e.g. en un futuro buscamos la posibilidad de generar un proceso que evalúe un documento académico de un cierto nivel a través de la obtención de sus medidas de Comprensibilidad y de saber si están dentro del segmento generado por nuestra estadística de medidas de Comprensibilidad. Estamos confiados en que esto se logrará a través de continuar realizando experimentos con documentos con más cuerpo, así como con una selección cuidadosa de documentos con una clara diferenciación de nivel académico. Así mismo, con la cuidadosa selección de una metodología de extracción de conceptos con mayor análisis. En un futuro, el proceso en general puede ser utilizado, como habíamos mencionado, tanto por asesores de tesis, estudiantes y hasta sinodales. Así mismo, el tener acceso a esta estadística que generamos es de interés, ya que aunque las medidas de Comprensibilidad ya han sido utilizadas anteriormente, los artículos que las utilizan no reportan resultados específicos para algún tipo de documento, niveles y secciones; y, en la literatura abordada, nadie analiza documentos de tesis en Español en específico.

Este trabajo de investigación entonces beneficia a cualquier área de trabajo donde la comunicación de la ciencia a través de texto sea importante, esto debido a que se generó un proceso de análisis, evaluación y retroalimentación de un texto académico. Es entonces un proyecto con claras aplicaciones, de tal forma que uno podría introducir un texto y gracias a nuestro trabajo el autor podría recibir una evaluación del nivel académico para ciertas secciones de su documento, qué párrafos podrían separarse para disminuir su valor de dispersión y hacerlos más legibles, qué conceptos no se abordan en profundidad o no tienen relaciones que deberían existir y cómo está la conexión de nuestro documento en general.

### 9.1.4. Trabajo a Futuro

Ahora explicaremos cuáles de los objetivos que nos planteamos no logramos concretar y cómo podríamos hacerlo en un futuro. Las preguntas que salieron a lo largo de la investigación y cómo podríamos atacarlas. Finalmente abordamos la problemática que se refiere a qué necesitaríamos para llevar a la práctica nuestra investigación.

Nuestro proyecto de investigación ubicó herramientas para un análisis adecuado de la Comprensibilidad de un texto, sin embargo, la Carga de Comprensibilidad fue una medida que no se logró estudiar adecuadamente. Para tener esta medida más fiable es necesario contar con textos relativamente largos y en ocasiones nuestros textos a analizar podían ser tan cortos como dos párrafos, de aquí que añadir tesis con mayor extensión a la colección de datos, o hasta crear una propia, está en un posible trabajo a futuro. Con esta mejora del corpus de datos, se pueden mejorar los resultados de las medidas de Comprensibilidad.

El proceso de retroalimentación visual necesita ser evaluado en la práctica. En la revisión del estado del arte hemos visto que una posible evaluación del mismo, podría realizarse a través de la selección de dos grupos de estudiantes con tesis de licenciatura en proceso. Uno de estos grupos utilizará nuestro proceso de retroalimentación visual para edición de su tesis y el otro grupo no. Una vez realizado esto, buscaríamos un proceso que lograse comparar el antes y después de los dos documentos para los dos grupos, esto podría ser a través de un método computacional o de evaluadores humanos que lean los documentos. Así mismo, está en nuestro interés realizar un análisis a profundidad de los conceptos. Aunque se hizo uno al comenzar a recabarlos, se quedó en las reuniones con el grupo de investigación. Nos gustaría generar un reporte de los conceptos donde se valore si nuestra metodología de extracción es efectiva y si no lo es, valorar nuevos métodos de extracción de conceptos y para encontrar relaciones entre ellos.

Hay hipótesis que nuestro proyecto de investigación no logró contestar y que ciertamente nos gustaría responder. Una de ellas es que debe existir una relación entre nivel académico y las medidas de Comprensibilidad, sin embargo, no logramos ver relaciones claras aún con la medida más estable que sería la de Dispersión. Entonces quedaría atacar este problema a través de la refinación y ampliación del corpus de datos, así como la continuidad de búsqueda en el estado del arte por más medidas de comprensibilidad que sean efectivas para nuestras metas. Por otro lado, la medida de Dispersión nos indica la densidad de información (esta basada en los conceptos) por párrafo, y debido a mi formación como físico, sería sumamente interesante, en un trabajo de doctorado, investigar posibles medidas que recaben densidad de objetos en diferentes instancias; éstas se podrían incorporar al análisis de conceptos en documentos de texto. Sería también prudente analizar posibles métodos para recabar de mejor manera esta medida y uno de ellos podría ser el ser más permisivo en qué acepta y no nuestro método como un concepto.

Finalmente, hablamos de llevar el proyecto de investigación a la práctica y tentativamente vemos esto posible a través de la idea que aparece en [35]. Ellos realizaron una API que básicamente utiliza un conjunto de textos ordenados con base en su Legibilidad y, al recibir un texto con un cierto nivel de Legibilidad el programa le ofrecía opciones de textos con un nivel de Legibilidad similar. En nuestra versión, el alumno o asesor, introduciría una tesis y nosotros regresaríamos su respectiva evaluación y retroalimentación como ya hemos mencionado anteriormente. Todo esto

se debería implementar de cero, pero las bases ya se tienen.

## 9.2. Conclusiones en Legibilidad

A continuación presentaremos las conclusiones para legibilidad. Primero hablaremos del problema que se abordó así como el proceso que se utilizó para abordarlo y sus ventajas y desventajas. Seguido de esto hablaremos de los resultados principales, para así pasar a colocarlos en el contexto del área. Finalmente abordaremos el trabajo a futuro.

### 9.2.1. Evaluación del Proceso

El problema que enfrentábamos en nuestro trabajo de investigación era la generación de un proceso lingüístico-computacional que fuese capaz de evaluar el nivel académico de un texto con base en su Legibilidad. Nuestro trabajo de investigación realizó esta tarea a través de dos pasos, primeramente se construyó un Clasificador/Comparador el cual es capaz de distinguir entre dos documentos de texto, cuál de ellos es más o menos legible. El segundo paso constó en la realización de un evaluador que a través de este Clasificador/Comparador fuera capaz de acotar tanto superior como inferiormente un documento. Con este evaluador uno podría estimar el nivel académico de un documento ajeno al corpus de datos.

El proceso para llevar a cabo la construcción, experimentación y selección del Clasificador/Comparador se basó en primera instancia, en la representación de documentos, los cuales se representaron a través de su frecuencia relativa y familiaridad, elecciones efectivas debido a que en nuestro proyecto de investigación no era central el análisis de características de representación. Una vez con los documentos representados, era necesario operarlos para entrenar al Clasificador/Comparador, en este aspecto analizamos tres posibles operaciones: concatenación, división y diferencia. La operación de diferencia obtuvo los mejores resultados y le siguió concatenación, sin embargo, continuar la experimentación con la operación de concatenación no era viable debido a que generaba vectores del doble de tamaño y el entrenar SVMs con matrices de esta dimensión no era viable. Por último, el esquema de un Clasificador/Comparador en nuestro proceso tiene ventajas claras, como lo son el poder entrenar con pocos datos y con solamente dos tipos de niveles académicos, sin embargo, existe la limitante de solamente poder comparar y no hay una clasificación directa, de aquí que la necesidad de crear un proceso de evaluación fue necesaria.

Analicemos ahora el proceso de los evaluadores; en total se construyeron cuatro: Aleatorio, Centroide, Máximo y Mínimo. Los cuatro se basan en una idea similar, el documento a evaluar se opera con base en el documento representativo de un conjunto de datos de un cierto nivel académico. Una vez operados, este documento se alimenta a un cierto Clasificador/Comparador ya entrenado, y dependiendo si se busca una cota superior o inferior, este Clasificador/Comparador es capaz de decidir si el documento es más o menos legible que el documento representativo mencionado. Este proceso tiene sus ventajas, ya que somos capaces de acotar perfectamente tanto superior como inferiormente y de esta manera aproximar la evaluación de un documento nuevo, sin embargo, esto mismo es un problema ya que solamente podemos acotar y no podemos específicamente aclarar que el documento corresponde a ese nivel en específico.

### 9.2.2. Conclusiones Principales

Los resultados principales en el área de Legbilidad, en específico para nuestros Clasificadores/Comparadores son los siguientes. El mejor Clasificador/Comparador es el de Licenciatura-Doctorado y le sigue TSU-Doctorado, como peor Clasificador/Comparador tenemos a TSU-Licenciatura. En cuanto a los modelos de entrenamiento, SVM mostró, en promedio, los mejores resultados. Otra observación importante que podemos concluir es que mientras los clasificadores se entrenen con niveles académicos más distantes, su clasificación será más exacta. Esto de hecho lleva sentido, ya que documentos con mayor "distancia académica" tendrán una diferencia en Legbilidad más notoria; esto afianza nuestra confianza en nuestros resultados.

En cuanto a los resultados para los evaluadores, nuestras conclusiones para acotar superior e inferiormente a un documento de licenciatura son las siguientes: Para las cotas superiores, el referente de Maestría es tiene mejores resultados que el de Doctorado y el Clasificador/Comparador que mejor utiliza a este referente es el de Licenciatura-Maestría. En cuanto a los evaluadores, todos tienen buenos resultados con excepción de Mínimo, el cual no debería ser utilizado para acotar superiormente. Para las cotas inferiores el referente de Técnico Superior Universitario es el de mejores resultados, el Clasificador/Comparador que utilizamos es el de TSU-Maestría y no se comparó con otros debido a los bajos resultados de TSU-Licenciatura. En cuanto al evaluador para acotar inferiormente, se concluye que los tres evaluadores obtienen cotas inferiores efectivas con excepción del evaluador Máximo. Es importante recalcar que las conclusiones sobre evaluadores son exclusivamente para documentos de licenciatura.

### 9.2.3. Resultados en Contexto

A continuación expondremos cómo nuestros resultados para legibilidad se acomodan en el contexto general del área de investigación. En primer lugar, nuestros resultados benefician en directo a cualquier estudiante en medio de la escritura de una tesis y, por ende, también benefician a los asesores que revisen esa tesis.

Nuestro proceso de legibilidad tiene dos principales contribuciones, la primera siendo la creación de los Clasificadores/Comparadores, los cuales son capaces de clasificar con una exactitud, en promedio, de hasta el 74,88 %. La segunda contribución es el proceso de evaluación y generación de los cuatros evaluadores (Aleatorio, Centroide, Máximo y Mínimo), éstos son capaces de acotar un documento de tesis y obtener una evaluación del nivel académico del documento de tesis a evaluar. Es también importante señalar que otra contribución importante es el manejo del tipo de datos (documentos de tesis académicos), en el estado del arte no encontramos la construcción de Clasificadores/Comparadores con este tipo de datos ni procesos de evaluación basados en Clasificadores/Comparadores. Finalmente, este proceso puede ampliarse para así evaluar diferentes tipos de documentos de texto, no está limitado a tesis de licenciatura.

Anteriormente mencionamos que nuestros resultados podrían tener una aplicación práctica como la que aparece en [35]. Es en ese mismo sentido que proponemos la inclusión del proceso de Legibilidad sobre esta aplicación. Por otro lado, nuestros resultados concuerdan con los mostrados para los Clasificadores/Comparadores que aparecen en [35], sin embargo, ellos logran clasificaciones en el 97 % y nosotros no, aún así, los documentos de texto que ellos manejan son sumamente diferenciables y

en sus resultados se puede observar que diferenciar documentos con niveles académicos de menor nivel se logra más sencillamente. Nuestros documentos son todos de un nivel académico alto, es por eso que diferenciarlos no se vuelve tan sencillo como uno esperaría. Por último su corpus de datos ha sido trabajado con expertos en lingüística mientras que nosotros aún podemos trabajar en mejorar la base de datos de [9].

### 9.2.4. Trabajo a Futuro

A continuación presentamos el trabajo a futuro en el área de Legibilidad. Con base en los objetivos que nos planteamos y lo que cumplimos con el trabajo de investigación nos sentimos satisfechos ya que se construyeron los Clasificadores/Comparadores con buenos porcentajes de exactitud. En cuanto a los evaluadores, la situación es similar ya que son capaces de acotar eficientemente el nivel académico de un texto de Licenciatura tanto superior como inferiormente.

Si tuviésemos más tiempo para continuar con el trabajo, nos interesaría trabajar con un modelo de ordenamiento de pares como el que aparece en [36]. Este modelo de ordenamiento se basa en una LSTMR y en su primera fase se entrena con pares ranqueados de datos donde uno de ellos es más propagandístico que el otro. Es claro entonces cómo este tipo de investigación se adapta a nuestro trabajo, sin embargo, es diferente y los resultados pueden arrojar conclusiones interesantes.

Así mismo, a lo largo del proyecto hemos hablado de la extracción de características como un elemento primordial en la clasificación de texto. De continuar con el proyecto, exploraríamos la inclusión de nuevas características lingüísticas de representación, esto para mejorar nuestros valores de exactitud. Así mismo, buscaríamos la combinación de diversas características para mejorar aún más la exactitud y para analizar si, efectivamente, necesitamos la representación donde se incluyen las características locales y globales; quizá una sola característica es suficiente. La implementación de éstas, sería basándonos en [27], ya que ellos tienen el estudio de las indicadas a usar para tareas de ordenamiento de texto. Otros detalle de interés puede ser la realización de experimentos con las características globales, ya que podemos experimentar con el número de palabras más usadas en el español, de tal forma que tendríamos tres experimentos: para 1000, 5000 y 10000 palabras.

Nuestro proyecto de investigación ha generado dudas que no hemos podido responder y que nos interesaría abordar de continuar con la investigación. Una de ellas recae en los bajos resultados que obtuvimos en los Clasificadores/Comparadores de TSU-Licenciatura, aunque es verdad que son Clasificadores/Comparadores que son entrenados con niveles académicos que no son distantes, esto no es suficiente para explicarlo ya que Licenciatura-Maestría y Maestría-Doctorado no tiene resultados tan bajos. Los datos pueden tener un papel importante aquí y quizá nuestros modelos de aprendizaje de máquina no son capaces de diferenciar bien entre estos dos documentos.

Una segunda pregunta que nos interesaría contestar con base en nuestros evaluadores, es el porqué encontrábamos mayor estabilidad en los referentes de Maestría que en los de Doctorado. Uno esperaría un comportamiento contrario ya que los referentes de Doctorado son, en teoría más complicados que los de Maestría y así mismo que los de Licenciatura, de aquí que la hipótesis no se cumple y sería necesario realizar más experimentos para obtener conclusiones sobre este comportamiento.

Finalmente, habíamos mencionado que nos gustaría incluir el proceso de evaluación de Legibilidad en una API, de tal forma que para llevar esto acabo necesitaríamos conjuntar los tres procesos desarrollados, el de evaluación de Comprensibilidad, el de generación de retroalimentación basado en Comprensibilidad y finalmente el de evaluación de Legibilidad. La idea es que todo esto en conjunto funcione para que un estudiante o asesor ingrese una tesis y nuestro software sea capaz de generar estos tres análisis.

## 9.3. Publicaciones Derivadas

1. Comprehensibility Analysis and Assessment of Academic Texts, Por publicarse en Proceedings of AIET 2021, 2nd International Conference on Artificial Intelligence in Education Technology, July 2-4, 2021, Wuhan, China.

2. Training Readability Comparators for Academic Texts at Different Levels, 14th Workshop on Intelligent Learning Environments, WILE 2021. October 25th or 26th, 2021. Cd. de México. Reconocimiento como mejor trabajo.

# Apéndice A

# Apéndice

## A.1. Evaluadores

### A.1.1. Centroide

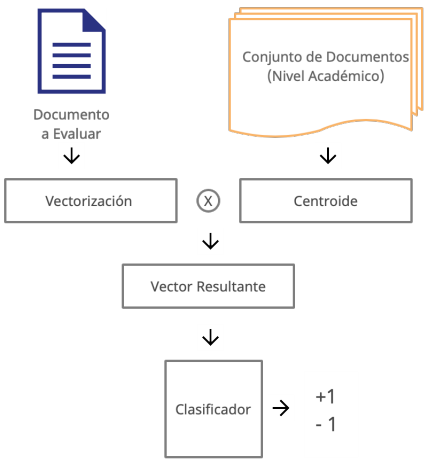



Figura A.1: Esquema para el evaluador basado en centroide

### A.1.2. Mínimo

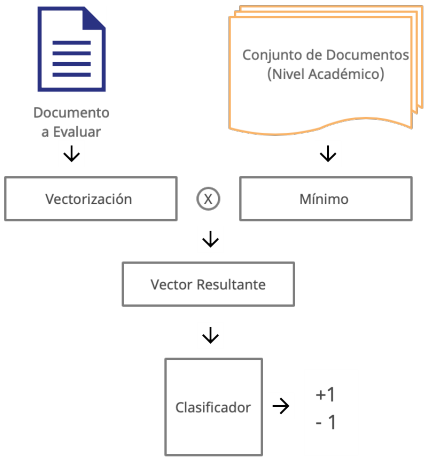



Figura A.2: Esquema para el evaluador basado en mínimos

### A.1.3. Máximo

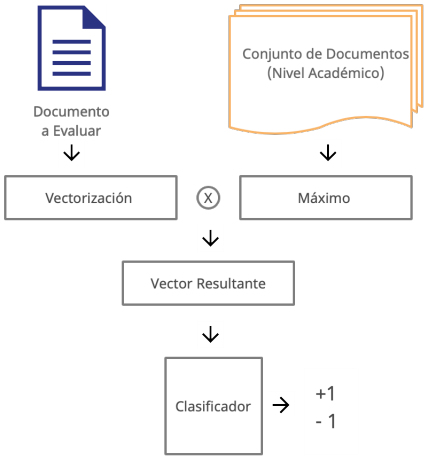



Figura A.3: Esquema para el evaluador basado en máximos

### A.1.4. Algoritmo de Selección de Máximo y Mínimo

**Algorithm 1:** Algoritmo para obtener el Máximo y Mínimo

```
Entradas: Matriz, Clasificador/Comparador
Inicialización
if Clasificador/Comparador(Matriz[0]-Matriz[1]) == 1 then
    max_vec = M[0]
    min_vec = M[1]
else
    min_vec = M[0]
    max_vec = M[1]
end
Recorremos los vectores restantes
for i desde 2 hasta tamaño(Matriz) do
    if Clasificador/Comparador(Matriz[i]-max_vec) == 1 then
        max_vec = M[i]
    end
    if Clasificador/Comparador(Matriz[i]-min_vec) == -1 then
        min_vec = M[i]
    end
end
```

# Bibliografía


[1] Rakesh Agrawal y col. "Empowering authors to diagnose comprehension burden in textbooks". En: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, págs. 967-975.

[2] PA Bromiley, NA Thacker y E Bouhova-Thacker. "Shannon entropy, Renyi entropy, and Information". En: *Statistics and Segmentation Series (2004-001)* (2010).

[3] Kevyn Collins-Thompson. "Computational assessment of text readability: A survey of current and future research". En: *ITL-International Journal of Applied Linguistics* 165.2 (2014), págs. 97-135.

[4] Daniel Dei. "La tesis". En: *Cómo orientarse en su elaboración. Buenos Aires: Prometeo. Di Stefano, M. et al.(2003). Escritura y producción de conocimiento en las carreras de posgrado (ponencia). Segundo Congreso Internacional Cátedra Unesco.* Universidad Católica de Valparaíso, Chile, 2006.

[5] Ralph W Fasold y Jeff Connor-Linton. *An introduction to language and linguistics*. 2.ª ed. Cambridge CB2 8BS, United Kingdom: Cambridge University Press, 2014.

[6] Aurélien Géron. "Hands-on machine learning with scikit-learn and tensorflow: Concepts". En: *Tools, and Techniques to build intelligent systems* (2017).

[7] Ingo Glaser y col. "Towards Computer-aided Analysis of Readability and Comprehensibility of Patient Information in the Context of Clinical Research Projects". En: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 2019, págs. 260-261.

[8] Jacinto González Cobas. "Estudio sobre el párrafo". En: *Estudios de Lingüística. N. 18* (2004), págs. 87-106.

[9] Samuel González-López y Aurelio López-López. "Colección de tesis y propuesta de investigación en tics: un recurso para su análisis y estudio". En: *XIII Congreso Nacional de Investigación Educativa*. 2015, págs. 1-15.

[10] Philip B Gove. "New International Dictionary". En: *Springfield, MA: Merriam-Webster Inc* (1986).

[11] Barbara Grosz y Candace L Sidner. "Attention, intentions, and the structure of discourse". En: *Computational linguistics* (1986).

[12] Julia Hancke, Sowmya Vajjala y Detmar Meurers. "Readability classification for German using lexical, syntactic, and morphological features". En: *Proceedings of COLING 2012*. 2012, págs. 1063-1080.

[13] Michael Heilman, Kevyn Collins-Thompson y Maxine Eskenazi. "An analysis of statistical models and features for reading difficulty prediction". En: *Proceedings of the third workshop on innovative use of NLP for building educational applications*. 2008, págs. 71-79.

[14] Matthew Honnibal y col. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303. URL: https://doi.org/10.5281/zenodo.1212303.

[15] Andreas Hotho, Andreas Nürnberger y Gerhard Paaß. "A brief survey of text mining." En: *Ldv Forum*. Vol. 20. 1. Citeseer. 2005, págs. 19-62.

[16] J. D. Hunter. "Matplotlib: A 2D graphics environment". En: *Computing in Science & Engineering* 9.3 (2007), págs. 90-95. DOI: 10.1109/MCSE.2007.55.

[17] Prakruthi Karuna. "Manipulating Comprehensibility of Text: An Automated Approach to Generate Deceptive Documents for Cyber Defense". Tesis doct. George Mason University, 2019.

[18] Prakruthi Karuna y col. "Fake document generation for cyber deception by manipulating text comprehensibility". En: *IEEE Systems Journal* (2020).

[19] Vlado Keselj. *Speech and Language Processing Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6,* 115,00. 2009.

[20] Marcus Kracht. *Introduction to linguistics*. Department of Linguistics, UCLA, 2007.

[21] Oscar Levin. *Discrete mathematics: An open introduction*. University of Northern Colorado, 2019.

[22] E. D. Liddy. "Natural language processing". En: *Encyclopedia of Library and Information Science* (2001).

[23] Lieve Macken y col. "Assessing the Comprehensibility of Automatic Translations (ArisToCAT)". En: *EAMT2020*. 2020, págs. 485-486.

[24] Kanubonu Narayan Murthy y PG Scholar. "Word Cloud in Python". En: *Complexity International* 24.01 (2020).

[25] Tore Opsahl, Filip Agneessens y John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths". En: *Social networks* 32.3 (2010), págs. 245-251.

[26] Anselmo Peñas, Felisa Verdejo, Julio Gonzalo y col. "Corpus-based terminology extraction applied to information access". En: *Proceedings of Corpus Linguistics*. Vol. 2001. 2001, pág. 458.

[27] Emily Pitler y Ani Nenkova. "Revisiting readability: A unified framework for predicting text quality". En: *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008, págs. 186-195.

[28] Manuel Alcántara Plá. "El concepto de oración para el español oral". En: *Oralia* 17 (2014), págs. 63-84.

[29] Mahima Relan, Swati Khurana y Vivek Kumar Singh. "Qualitative evaluation and improvement suggestions for ebooks using text analytics algorithms". En: *Proceedings of Second International Conference on Eco-friendly Computing and Communication Systems, Solan, India*. 2013.

[30] Alessandra Rossetti. "Intralingual translation and cascading crises: evaluating the impact of semi-automation on the readability and comprehensibility of health content". En: *Translation in Cascading Crises*. Abingdon, UK: Routledge (Taylor Francis), 2019.

[31] Mercedes Sánchez Sánchez y Carlos Domınguez Cintas. "El banco de datos de la RAE: CREA y CORDE". En: *Per Abbat: boletın filológico de actualización académica y didáctica* 2 (2007), págs. 137-148.

[32] Elliot Schumacher y col. "Predicting the relative difficulty of single sentences with and without surrounding context". En: *arXiv preprint arXiv:1606.08425* (2016).

[33] Luo Si y Jamie Callan. "A statistical model for scientific readability". En: *Proceedings of the tenth international conference on Information and knowledge management*. 2001, págs. 574-576.

[34] Alex Smola y SVN Vishwanathan. "Introduction to machine learning". En: *Cambridge University, UK* 32.34 (2008), pág. 2008.

[35] Kumiko Tanaka-Ishii, Satoshi Tezuka e Hiroshi Terada. "Sorting texts by readability". En: *Computational Linguistics* 36.2 (2010), págs. 203-227.

[36] Liqiang Wang y col. "Cross-Domain Learning for Classifying Propaganda in Online Contents". En: *arXiv preprint arXiv:2011.06844* (2020).

[37] Menglin Xia, Ekaterina Kochmar y Ted Briscoe. "Text readability assessment for second language learners". En: *arXiv preprint arXiv:1906.07580* (2019).

[38] Haoran Zhang y Diane Litman. "Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring". En: *arXiv preprint arXiv:2008.01809* (2020).