

|  |
| --- |
| **Universidad Internacional de La Rioja**  **Escuela Superior de Ingeniería y Tecnología**  **Máster Universitario en Análisis y Visualización de Datos Masivos** |
|  |
| Análisis de la Cronología y relaciones entre personajes bíblicos y eventos a través del tiempo |

**Trabajo Fin de Máster**

**Tipo de trabajo:** Piloto Experimental

**Presentado por:** González Cano, David

**Director:** Fernández García, Diego

**Resumen**

A través del uso de diferentes técnicas de Big Data y Visualización del Procesamiento de Lenguaje Natural (NPL), usando técnicas dirigidas al análisis de modernos análisis de texto, usando Python y herramientas de software libre como Google BERT.

La propuesta de trabajo de Fin de Master (TFM) tiene el propósito central de utilizar estas técnicas para analizar la cronología de los personajes bíblicos, los eventos y las relaciones que estos personajes tienen en los libros de la Biblia.

Para ello se utilizan diferentes datasets públicos que contienen variables como “fecha, “pasaje de la biblia”, mediante técnicas de NPL se analizan las variables y finalmente se crea una plataforma de visualización que sea accesible a cualquier usuario que lo requiera.

Para la recopilación, limpieza, depuración, análisis, clasificación y visualización de datos se utilizará Python con sus respectivas librerías.

Para analizar y trabajar con textos relacionados con la cronología bíblica se va a utilizar **NPL** (Procesamiento de Lenguaje Natural)

Para la visualización se está pensando en herramientas colaborativas como Google sheet que permita obtener tanto visualizaciones estáticas como dinámicas.

Lo que se espera tener de esta investigación es poder dejar un prototipo que utilizando el texto de la biblia permita usar las herramientas para hacer preguntas sobre la cronología tanto de eventos como de personajes y presente visualizaciones claras de lo requerido.

**Palabras Clave:** Procesamiento Natural de Lenguaje, NLP, Cronología Bíblica, Google BERT, GitHub. Google Colab, Visualización, Power Bi, Biblia,

**Abstract**

The Master's Thesis proposal (TFM) has the central purpose of using different Big Data and Data Visualization techniques to analyze the chronology of biblical characters, the events and the relationships that these characters have in the books of the Bible.

Using the different datasets that are public, with different variables and different formats.

Python with its respective libraries will be used to collect, clean, debug, analyze, classify and visualize data.

To analyze and work with texts related to biblical chronology, NPL (Natural Language Processing) will be used.

For visualization, collaborative tools such as Google Sheet are being considered that allow obtaining both static and dynamic visualizations.

What is expected from this research is to be able to leave a prototype that, using the text of the Bible, allows the use of tools to ask questions about the chronology of both events and characters and presents clear visualizations of what is required.

**Keywords:** Natural Language Processing, NLP, Biblical Chronology, Google BERT, GitHub. Google Colab, Visualization, Power Bi, Bible

Índice de Contenido

[1. Introducción 7](#_Toc155220244)

[1.1 Justificación 7](#_Toc155220245)

[1.2 Planteamiento del trabajo 8](#_Toc155220246)

[1.3 Estructura de la memoria 8](#_Toc155220247)

[2. Contexto y estado del arte 9](#_Toc155220248)

[2.1 Introducción al tema 9](#_Toc155220249)

[2.2. Desarrollo (base teórica): antecedentes, estudios actuales, autores de referencia 9](#_Toc155220250)

[2.2.1 Análisis de Texto 11](#_Toc155220251)

[2.2.2 NLP (Natural Lenguaje Processing) 24](#_Toc155220252)

[2.2.2.1 Word Embedding 25](#_Toc155220253)

[2.2.2.2 Pruebas 33](#_Toc155220254)

[2.2.2.3 Haciendo Predicciones 33](#_Toc155220255)

[2.2.3 Google BERT (Bidirectional Encoder Representations from Transformers) to 34](#_Toc155220256)

[2.2.4. Herramientas Colaborativas 34](#_Toc155220257)

[2.2.4.1 GitHub / GitLab / Bitbucket / Gitkraten / SourceGorge 35](#_Toc155220258)

[*2.2.4.2* Google *Colab* 36](#_Toc155220259)

[2.2.5. Python 36](#_Toc155220260)

[2.2.5.1 Análisis de Texto 37](#_Toc155220261)

[2.2.5.2 Bibliotecas Python 39](#_Toc155220262)

[2.2.6. La Biblia 39](#_Toc155220263)

[2.2.7. Dataset bíblicas 52](#_Toc155220264)

[2.3 Conclusiones (nexo de unión de lo investigado con el trabajo a realizar) 53](#_Toc155220265)

[3. Objetivos concretos y metodología de trabajo 53](#_Toc155220266)

[3.1. Objetivo general 53](#_Toc155220267)

[3.2. Objetivos específicos 53](#_Toc155220268)

[4. Desarrollo específico de la contribución 54](#_Toc155220269)

[4.1. Descripción detallada del Piloto experimental 54](#_Toc155220270)

[4.1.1 Qué tecnologías se utilizaron (incluyendo justificación de por qué se emplearon y descripciones detalladas de las mismas). 54](#_Toc155220271)

[4.1.2 Cómo se organizó el piloto 55](#_Toc155220272)

[4.1.3 Qué personas participaron (con datos demográficos) o qué técnicas de evaluación automática se emplearon. 55](#_Toc155220273)

[4.1.4. Cómo transcurrió el experimento. 55](#_Toc155220274)

[4.1.5 Qué instrumentos de seguimiento y evaluación se utilizaron. 55](#_Toc155220275)

[4.2 Descripción de los resultados 55](#_Toc155220276)

[4.3. Discusión 55](#_Toc155220277)

[5. Conclusiones y trabajo futuro 56](#_Toc155220278)

[5.1. Conclusiones 56](#_Toc155220279)

[5.2. Líneas de trabajo futuro 56](#_Toc155220280)

[Bibliografía 58](#_Toc155220281)

**Índice de figuras**

[Figura 1 Taxonomía del machine Learning dentro de la Inteligencia Artificial. Basado en Panesar, 2019 10](#_Toc155216600)

[Figura 2 Tasa de crecimiento de los datos entre 2006-2020 (Cheesem, n.d.) 11](#_Toc155216601)

[Figura 3 Un ejemplo de un modelo de traducción neuronal, trabajando del francés al inglés. (Introduction to Neural Machine Translation with GPUs (Part 2), 2015) 14](#_Toc155216602)

[Figura 4 Técnicas como el modelado de temas utilizan métodos de modelado probabilístico para identificar temas clave del texto. (Thiyagarajan, 2018) 15](#_Toc155216603)

[Figura 6 Generando un modelo estadístico con el algoritmo de aprendizaje automático 32](#_Toc155216604)

[Figura 7 Ejemplo Predicción una estructura de árbol de dependencia usando un modelo estadístico 34](#_Toc155216605)

[Figura 8 Parte de la visualización de cronología de la Biblia (Aschmann, 2022) 52](#_Toc155216606)

**Tabla de Contenido**

[Tabla 1 Simplificado Espacio vectorial de palabras 27](#_Toc155216610)

[Tabla 2 Operación matemática vectorial en un espacio vectorial de palabras 27](#_Toc155216611)

[Tabla 3 Libros de la Biblia Hebrea o Antiguo testamento 41](#_Toc155216612)

[Tabla 4 Libros Antiguo testamento mayoría biblias cristianas 43](#_Toc155216613)

[Tabla 5 Antiguo testamento (Antiguo Testamento, n.d.) 48](#_Toc155216614)

[Tabla 6 Libros del nuevo testamento más aceptaods (Nuevo Testamento, n.d.) 51](#_Toc155216615)

# 1. Introducción

## 1.1 Justificación

Meditando sobre los proyectos que dan gasolina en el día a día y que se vuelven retos ultraístas, el leer la Biblia es uno de ellos, desde su inicio hasta el final contando con más de cuarenta escritores y aproximadamente 773.746 palabras en un libro que nos lleva desde el inicio del mundo, va recorriendo con eventos, personajes los diferentes tiempos de la humanidad y ahí, en el mismo texto lleno de aventuras nos sumergimos en cual será el final de la humanidad con señales que se conectan en los 66 libros divididos en 1189 capítulos y 31.102 versículos y 3.5 35.666.480 letras, Biblia Reina Valero 1960. (Concepto y definicion Net, 2023)

La secuencia de los tiempos, los personajes, los eventos en los escritos no son tan fáciles de visualizar debido a que las fechas no están implícitas en el texto y se necesitan estudios teológicos y contextos históricos para poder entender la cronología de los eventos con sus respectivos personajes.

Para llevar a cabo esta investigación se van a utilizar diferentes técnicas, métodos y herramientas para encontrar una forma de analizar y presentar cronológicamente eventos, personajes utilizando textos que están ya colocados en la web y también si es posible desarrollar un prototipo de base de datos que nos permita ir actualizando las fechas en una línea de tiempo actualizable por los investigadores de este tema.

La necesidad de esta investigación surge de la diversidad de autores en sus escritos, las figuras literarias y estilísticas que se utilizan como metáforas, símiles, personificaciones, parábolas, hipérboles, ironías, y muchas más; la identificación de fechas en cada evento, la no existencia hasta el momento de una base de datos estructurada que coloque cada evento o personaje en una línea de tiempo.

Por todo ello se propone una herramienta a la población interesada en poder dar un contesto en el tiempo de los escritos ya sea por estudio u otro interés dando una visualización en una forma clara y utilizando herramientas de datos masivos.

En la población mundial aproximadamente de 8300 millones y con un cálculo de 2500 millones de cristianos en el mundo (EL orden Mundial EOM, 2019), tienen como base de su fe la Biblia, la cual se divide en Antiguo y nuevo testamento. Así a esta población se puede sumar el pueblo judío aproximadamente 13 millones (Embajada de Israel en España, s.f.), los cuales su base de lectura en su creencia es el Antiguo Testamento. ya que tienen una división diferente y se conoce como la biblia hebrea o Tanaj la cual está dividida en tres partes: la Torá (ley) , los Nevi’im (profetas) y los Ketuvim (escritos), además los judíos no consideran la parte de los libros del nuevo testamento como parte de las escrituras.

## 1.2 Planteamiento del trabajo

Utilizando técnicas de procesamiento de lenguaje natural (NLP) y modelos de machine Learning que permitan extraer información y procesarla cronológicamente. Creando una base de datos estructurada que manaje y permita trabajar con fechas de los personajes bíblicos y utilizar herramientas de visualización que permitan de una forma fácil presentar los eventos y personajes en forma dinámica.

## 1.3 Estructura de la memoria

La primera parte es describir las técnicas de manejo de Texto

Lo segunda parte está dirigida al análisis de la biblia como un dataset y como se logra organizar para obtener información cronológica

El tercer es el prototipo planteado de la unión de las técnicas y la estructura bíblica con herramientas de Visualización.

# 2. Contexto y estado del arte

## 2.1 Introducción al tema

Es interesante pensar que cuando hablamos de tiempo, aunque no lo sentimos material, tiene estados de pasado, presente y futuro; y cuando lo graficamos con personajes y eventos se nos convierte en una línea de tiempo, una herramienta visual para presentar cualquier información que mezcla datos y eventos formando una función de sucesión temporal.

La cronología bíblica se refiere al estudio y ordenamiento de los eventos históricos descritos en la biblia en secuencia temporal y busca determinar fechas aproximadas de eventos, así como la duración de periodos y la secuencia en la que ocurrieron.

El análisis de texto mediante el procesamiento del Lenguaje natural aplicado a el texto bíblico tiene varios temas de interés, asi como la identificación de personajes, eventos y sus relaciones en el tiempo usando diferentes técnicas.

## 2.2. Desarrollo (base teórica): antecedentes, estudios actuales, autores de referencia

El proyecto se enmarca en la Disciplina de la inteligencia artificial que, según la definición de la real academia española, Inteligencia Artificial: “es una Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico” (ASALE, n.d.)

Como un subcampo encontramos el procesamiento de lenguaje natural que trata el procesar y analizar los datos de lenguaje natural. Esto incluye enseñarles a máquinas para interactuar con humanos con un lenguaje que se desarrolla naturalmente a través del uso, creando algoritmos de Machine Learning que trabajan con grandes datasets desconocidas. Hoy en día se pueden usar algoritmos para observar cuyas reglas semánticas y gramaticales son bien conocidas. No es sorprendente que generar y comprender el lenguaje natural sea la tarea más prometedora y, al mismo tiempo, más desafiante, involucrada en NPL. (Vasiliev, 2020)

.

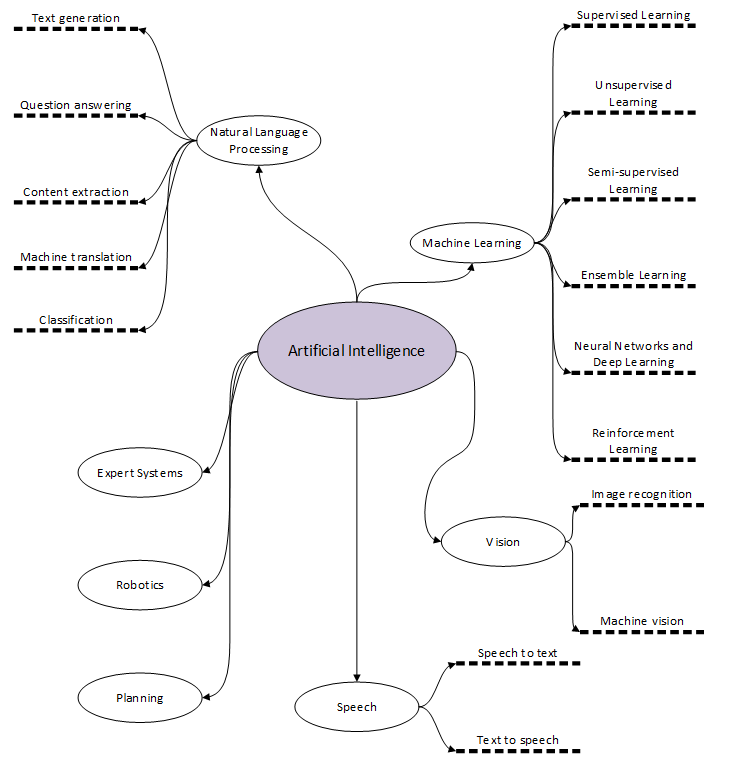


Figura 1 Taxonomía del machine Learning dentro de la Inteligencia Artificial. Basado en Panesar, 2019

.

## 2.2.1 Análisis de Texto

(Srinivasa-Desikan, 2018)Actualmente el análisis de texto se puede realizar usando Python y otras herramientas abiertas que trabajan con algoritmos que están disponibles como spaCy y Gensim.

Si hay un medio de comunicación que en mundo está inmerso, es el **texto**. Desde los periódicos, revistas, internet, email, mensaje de texto y la mayoría de información que se recibe. La cantidad de información que se maneja en compañías como Google (más de un billón de consultas por año) Twitter (1.6 billones de mensajes por día), y WhatsApp (30 mil millones de mensajes por día). Los datos textuales tienen un enorme valor comercial y las empresas pueden utilizar estos datos para ayudar a perfilar los clientes y comprender tendencias. Esto puede utilizarse para ofrecer una experiencia más personalizada a los usuarios o como información de mercadeo dirigido. Facebook, por ejemplo, utiliza una gran cantidad de datos textuales, y uno de los algoritmos fue desarrollado por el equipo de AI de Facebook. (Srinivasa-Desikan, 2018)

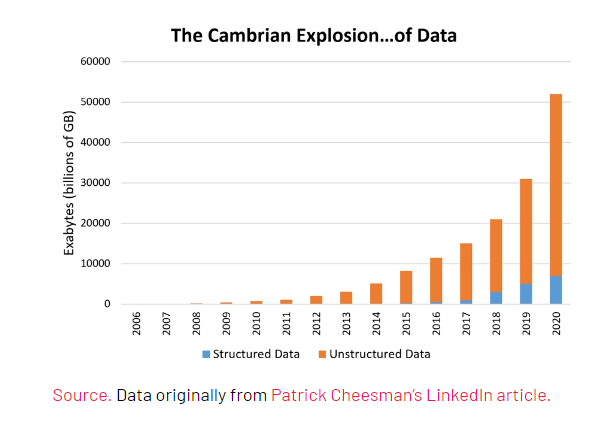


Figura 2 Tasa de crecimiento de los datos entre 2006-2020 (Cheesem, n.d.)

El análisis de texto puede entenderse como la técnica de extraer información útil del texto, la cual se puede hacer mediante varias técnicas ya sea utilizando el procesamiento del lenguaje natural (PNL), Lingüística Computacional (CL) y herramientas numéricas para obtener esta información. Estas herramientas numéricas son algoritmos de aprendizaje automático o algoritmos de recuperación de información.

El procesamiento del lenguaje natural (PNL) se refiere al uso de una computadora para procesar el lenguaje natural. Por ejemplo, eliminar todas las apariciones de la palabra de un cuerpo de texto es un ejemplo de ello, aunque sea un ejemplo básico.

La lingüística computacional (CL), como su nombre indica, es el estudio de la lingüística desde una perspectiva computacional. Esto significa utilizar computadoras y algoritmos para realizar tareas lingüísticas, como marcar el texto como parte del discurso (como sustantivo o verbo), en lugar de realizar esta tarea manualmente.

Machine Learning (ML) es el campo de estudio en el que utilizamos algoritmos estadísticos para enseñar a las máquinas a realizar una tarea particular. Este aprendizaje ocurre con datos y nuestra tarea a menudo es predecir un nuevo valor basado en datos observados previamente.

La recuperación de información (IR por sus siglas en inglés, Information Retrieval ) es la tarea de buscar o recuperar información en función de una consulta del usuario. Los algoritmos que ayudan a realizar esta tarea se denominan algoritmos de recuperación de información y incluyen temas como indexación, modelos de recuperación, relevancia y evaluación que mejoran la calidad de los resultados de las búsquedas y la interacción entre Usuario-Sistema que mejora la precisión de las búsquedas. Es importante destacar en este punto que la precisión se refiere a la proporción de los documentos que son relevantes. Este punto es muy común en motores de búsqueda, bibliotecas digitales donde la recuperación de documentos y búsqueda de información textual son fundamentales.

El análisis de texto en sí existe desde hace mucho tiempo: una de las primeras definiciones de Business Intelligence (BI), en un artículo del IBM Journal de octubre de 1958 escrito por H. P. Luhn, A Business Intelligence System [1], describe un sistema que hará el siguiente:

"...utilizar máquinas de procesamiento de datos para la abstracción y codificación automática de documentos y para crear perfiles de interés para cada uno de los 'puntos de acción' de una organización. Tanto los documentos entrantes como los generados internamente se abstraen automáticamente, caracterizándose por una palabra patrón y se envía automáticamente a los puntos de acción apropiados".

Es interesante hablar acerca de documentos, en lugar de números; pensar que las primeras ideas de la inteligencia empresarial fueron comprender texto y documentos es testimonio del análisis de texto a lo largo de los siglos. Pero incluso fuera del ámbito del análisis de texto para empresas, el uso de computadoras para comprender mejor el texto y el lenguaje ha existido desde el comienzo de las ideas de inteligencia artificial. La revisión de 1999 sobre el análisis de textos realizada por John Hutchins, Retrospect and prospect in computer-based Translation (Hutchins, 1999) habla de los esfuerzos por realizar traducción automática ya en la década de 1950 por parte del ejército de los Estados Unidos, con el fin de traducir revistas científicas rusas al inglés.

Los esfuerzos para crear una máquina inteligente también comenzaron con texto: el programa ELIZA desarrollado en 1966 en el MIT por Joseph Weizenbaum es un ejemplo. Aunque el programa no tenía una comprensión real del lenguaje, mediante la coincidencia de patrones básicos podía intentar mantener una conversación. Estos son sólo algunos de los primeros intentos de analizar texto: las computadoras (¡y los seres humanos!) han recorrido un largo camino desde entonces y ahora tenemos herramientas increíbles a nuestra disposición.

La traducción automática en sí ha recorrido un largo camino y ahora podemos usar nuestros teléfonos inteligentes para traducir de manera efectiva entre idiomas, y con técnicas de vanguardia como Google’s Neural Machine Translation, la brecha entre la academia y la industria se está reduciendo permitiendo experimentar de primera mano la magia del procesamiento del lenguaje natural.

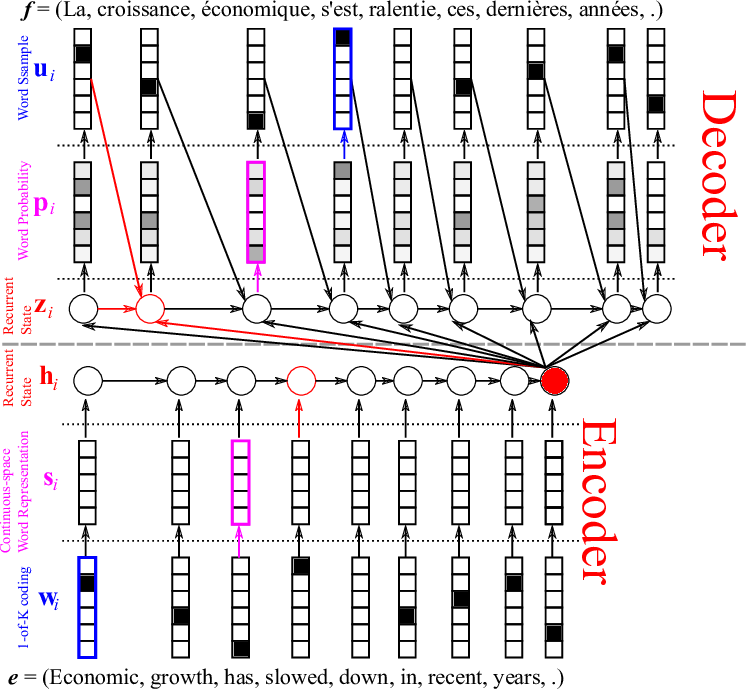


Figura 3 Un ejemplo de un modelo de traducción neuronal, trabajando del francés al inglés. (Introduction to Neural Machine Translation with GPUs (Part 2), 2015)

Los avances en este tema han ayudado a mejorar la forma en que abordamos el habla y los subtítulos en videos, y los asistentes personales como Siri de Apple o Alexa de Amazon se benefician enormemente de un procesamiento de texto superior. Comprender la estructura de las conversaciones y extraer información fueron problemas clave en los inicios de la PNL, y los frutos de la investigación están siendo muy evidentes en el siglo XXI.

¡Motores de búsqueda como Google o Bing! También se apoyan en las investigaciones realizadas en PNL y CL y afectan nuestras vidas de una manera sin precedentes. La recuperación de información (RI) se basa en enfoques estadísticos en el procesamiento de textos y nos permite clasificar, agrupar y recuperar documentos. Métodos como el modelado de temas pueden ayudarnos a identificar temas clave en cuerpos de texto grandes y no estructurados. Identificar estos temas va más allá de la búsqueda de palabras clave y utilizamos modelos estadísticos para comprender mejor la naturaleza subyacente de los cuerpos de texto. Sin el poder de las computadoras, no podríamos realizar este tipo de análisis estadístico a gran escala del texto.

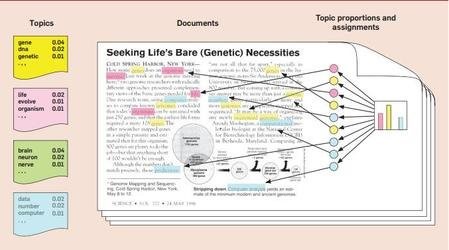


Figura 4 Técnicas como el modelado de temas utilizan métodos de modelado probabilístico para identificar temas clave del texto. (Thiyagarajan, 2018)

El poder experimentar las maravillas de la informática moderna en nuestros teléfonos móviles, los recientes desarrollos tanto en Python como en PNL significan que ahora podemos desarrollar dichos sistemas por nuestra cuenta. No sólo ha habido una evolución en las técnicas utilizadas en PNL y análisis de texto, sino que también se vuelven muy accesibles para todos: los paquetes de código abierto se están convirtiendo en herramientas de última generación y con un rendimiento tan bueno como comercial. Un ejemplo de herramienta comercial sería la API de análisis de texto de Microsoft. (Azure AI Language, n.d.)

MATLAB es otro ejemplo de una herramienta comercial popular utilizada para la informática científica. Si bien históricamente este tipo de herramientas comerciales funcionaron mejor que el software gratuito de código abierto, un aumento en el número de personas que contribuyen a las bibliotecas de código abierto, así como la financiación de la industria, han ayudado enormemente a la comunidad de código abierto. Ahora, las cosas parecen haber cambiado y muchos gigantes del software utilizan paquetes de código abierto para sus sistemas internos, como Google que utiliza TensorFlow y Apple que utiliza scikit-learn. TensorFlow y scikit-learn son dos paquetes de aprendizaje automático de Python de código abierto.

Se puede argumentar que la gran cantidad de paquetes que ofrece el ecosistema Python significa que lidera el grupo cuando se trata de realizar análisis de texto. Una comunidad de código abierto muy fuerte y activa se suma al atractivo.

**¿Dónde están los datos?**

Si bien es importante conocer las técnicas y herramientas involucradas en PNL y CL, por supuesto, es inútil sin ningún dato. Por suerte para nosotros, tenemos acceso a una gran cantidad de datos si buscamos en los lugares correctos. La forma más sencilla de encontrar datos textuales con los que trabajar es buscar un corpus.

Un corpus de texto es un conjunto grande y estructurado de textos y es una excelente manera de comenzar con el texto. análisis. Ejemplos de corpus libres son el Open American National Corpus (CORPUS, n.d.) o el British National Corpus [6]. Wikipedia tiene una lista útil de los corpus más grandes disponibles en su artículo sobre corpus de texto [7]. Estos no se limitan al idioma inglés, también existen varios corpus en idiomas europeos y asiáticos, y hay esfuerzos constantes en todo el mundo para crear corpus para la mayoría de los idiomas. Los laboratorios de investigación de las universidades son otra fuente valiosa para obtener corpus; de hecho, uno de los corpus más emblemáticos del idioma inglés, el Brown Corpus, se elaboró ​​en la Universidad de Brown.

Existen corpus que se han creado específicamente para representar textos bíblicos en el ámbito del procesamiento de lenguaje natural (NLP) y la lingüística computacional. Estos corpus son colecciones estructuradas de textos bíblicos que se utilizan para realizar análisis lingüísticos, entrenar modelos de lenguaje y llevar a cabo investigaciones relacionadas con la lingüística y la interpretación de textos religiosos.

Algunos ejemplos son Society of Biblical Literature Greek New Testament (SBLGNT, n.d.), WLC (Westminster Leningrad Codex)

Algunos ejemplos de corpora bíblicos incluyen:

1. **SBLGNT (Society of Biblical Literature Greek New Testament):**
   * Un corpus que representa el Nuevo Testamento en griego. Contiene el texto crítico del Nuevo Testamento griego junto con información adicional como lemas y etiquetas morfológicas.
2. **WLC (Westminster Leningrad Codex):**
   * Un corpus que representa el Antiguo Testamento en hebreo. Este corpus se basa en el Codex Leningradensis, uno de los manuscritos más antiguos del Antiguo Testamento.
3. **BibleCorpus:**
   * Un corpus que incluye el texto completo de la Biblia en varias versiones y traducciones. Puede ser utilizado para tareas como análisis de sentimientos en textos bíblicos y otras aplicaciones de NLP.
4. **Open Scriptures Hebrew Bible:**
   * Un corpus que representa el Antiguo Testamento en hebreo con información morfológica y sintáctica.

Los diferentes corpus tienden a tener distintos niveles de información presente, generalmente dependiendo del propósito principal de esos corpus; por ejemplo, los corpus cuya función principal es ayudar durante la traducción tendrían la misma oración presente en varios idiomas. Otra forma en que los corpus obtienen información adicional es mediante anotaciones. Los ejemplos de anotaciones en texto generalmente incluyen etiquetado de parte del discurso (POS) o reconocimiento de entidad nombrada (NER). El etiquetado POS se refiere a marcar cada palabra en una oración con su parte de la oración (sustantivo, verbo, adverbio, etc.), y un corpus anotado para NER tendría reconocidas todas las entidades nombradas, como lugares, personas y horas. Más adelante profundizaremos en los detalles del etiquetado POS y NER, en el Capítulo 5, Etiquetado POS y sus aplicaciones y en el Capítulo 6, Etiquetado NER y sus aplicaciones.

Según la estructura y los distintos niveles de información presentes en los corpus, tendría un propósito diferente. Algunos corpus también se crean para evaluar tareas de agrupación o clasificación, donde en lugar de que la anotación sea importante, lo sería la etiqueta o la clase. Esto significa que algunos corpus están diseñados para ayudar con tareas de aprendizaje automático, como agrupaciones o clasificación, proporcionando texto con etiquetas etiquetadas por humanos. La agrupación se refiere a la tarea de agrupar objetos similares, y la clasificación es el proceso de decidir qué clase predefinida e identificar para qué se utilizará exactamente su conjunto de datos es una parte crucial del análisis de texto y un primer paso importante.

Además de descargar conjuntos de datos o extraer datos de Internet, todavía existen algunas fuentes ricas para recopilar nuestros datos textuales, en particular, literatura. Un ejemplo de ello es la investigación realizada en la Universidad de Pensilvania, donde Alejandro Ribeiro, Santiago Segarra, Mark Eisen y Gabriel Egan descubrieron posibles colaboradores de Shakespeare, un problema de historia literaria que hizo tropezar a muchos investigadores [14]. Abordaron el problema identificando estilos literarios, un campo de estudio de próxima aparición en lingüística computacional llamado análisis de estilo.

El mayor uso de herramientas computacionales para realizar investigaciones en humanidades también ha llevado al crecimiento de los laboratorios de Humanidades Digitales en las universidades, donde los enfoques de investigación tradicionales se ven favorecidos o superados por la informática y, en particular, el aprendizaje automático (y por extensión), la ciencia natural. procesamiento del lenguaje. Los discursos de los políticos o los procedimientos en el parlamento, por ejemplo, son otro ejemplo de fuente de datos que se utiliza con frecuencia en esta comunidad. They WorkFor You [17] es un sistema de seguimiento del parlamento del Reino Unido, que recibe discursos y los carga y es un ejemplo de los muchos sitios disponibles que realizan este tipo de trabajo.

El Proyecto Gutenberg es probablemente el mejor recurso para descargar libros y contiene más de 50.000 libros electrónicos gratuitos y muchos clásicos literarios. Los archivos PDF y libros electrónicos personales también siguen siendo un recurso, pero nuevamente, es importante conocer la naturaleza legal de su texto antes de analizarlo.

Descargar una copia pirateada de, digamos, Harry Potter de Internet y publicar los resultados del análisis de texto puede no ser la mejor idea si no puedes explicar de dónde sacaste el texto. De manera similar, el análisis de texto en mensajes de texto privados podría no sólo molestar a tus amigos sino que también podría infringir las leyes de privacidad.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Fig. 1.4 Un ejemplo de una lista de conjuntos de datos de texto: aquí se trata de conjuntos de datos de reseñas que se encuentran en

Entonces, ¿dónde, además de descargar un conjunto de datos estructurados directamente de Internet, obtenemos nuestros datos textuales? Bueno, internet, por supuesto. Incluso si no está etiquetado, la gran cantidad de texto en Internet significa que podemos acceder a grandes partes de él: el [7] es un ejemplo de ello, y el volcado multimedia de todo el contenido en Wikipedia, después de descomprimirlo, es alrededor de 58 GB (a abril de 2018): texto más que suficiente para jugar. El popular sitio web de agregación de noticias reddit.com [9] permite un fácil raspado web y es otro gran recurso para el análisis de texto.

Python nuevamente sigue siendo una excelente opción para este tipo de web scraping, y bibliotecas como BeautifulSoup [10], urllib [11] y scrapy [12] están diseñadas especialmente para esto. Es importante tener cuidado con el aspecto legal de las cosas aquí y asegurarse de verificar los términos y condiciones del sitio web del que está extrayendo los datos; varios sitios web no le permitirán utilizar la información del sitio web para propósitos comerciales.

Twitter es otro sitio web que se está convirtiendo rápidamente en una parte muy importante del análisis de texto; incluso hay académicos que se toman este recurso muy en serio (¿Qué es Twitter, una red social o un medio de noticias? [13] tiene más de 5000 citas), con múltiples Se están escribiendo artículos sobre análisis de texto de tweets e incluso se han creado herramientas completas [15] para realizar análisis de sentimientos. La API de transmisión de Twitter también nos permite extraer fácilmente datos textuales de Twitter, y la interfaz de Python [16] es sencilla. La mayoría de los líderes mundiales son usuarios de Twitter, así como celebridades y grandes corporaciones de noticias; Twitter puede ofrecernos muchas ideas interesantes.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Fig. 1.5 Un ejemplo del recurso de texto enriquecido en el que se ha convertido Twitter, con múltiples conjuntos de datos estructurados disponibles [7]. Estos conjuntos de datos, por supuesto, tienen tareas particulares, que pueden usarse y caer en la categoría de conjuntos de datos etiquetados que discutimos antes.

Otros ejemplos de información textual que puede obtener de Internet incluyen artículos de investigación, informes médicos, reseñas de restaurantes (me viene a la mente el conjunto de datos de Yelp!) y otros sitios web de redes sociales. El análisis de sentimientos suele ser el objetivo principal en estos casos. Como sugiere el nombre, el análisis de sentimientos se refiere a la tarea de identificar sentimientos en el texto. Estos sentimientos pueden ser básicos, como sentimientos positivos o negativos, pero podríamos tener tareas de análisis de sentimientos más complejas en las que analicemos si una oración contiene sentimientos felices, tristes o enojados.

Está claro que si miramos lo suficiente, es más que fácil encontrar datos con los que jugar. Pero demos un pequeño paso atrás en la descarga de datos de Internet: ¿dónde más podemos intentar encontrar información?

Está en nuestras manos, como puede parecer: enviamos y recibimos mensajes de texto y correos electrónicos todos los días, y podemos usar este texto para el análisis de texto. La mayoría de las aplicaciones de mensajería de texto tienen interfaces para descargar chats. WhatsApp, por ejemplo, le enviará los datos por correo [18], tanto con medios como con texto. La mayoría de los clientes de correo tienen la misma opción, y la ventaja en ambos casos es que este tipo de datos suele estar bien organizado, lo que permite una fácil limpieza y preprocesamiento antes de sumergirnos en los datos.

Un aspecto que hemos ignorado hasta ahora al hablar de datos es el ruido que a menudo hay en el texto: en los tweets, por ejemplo, los formularios cortos y los emoticones que se utilizan con frecuencia y, en algunos casos, tenemos datos multilingües en los que un Un análisis simple podría fallar. Esto nos lleva al aspecto posiblemente más importante del análisis de texto: el preprocesamiento.

**Basura dentro basura fuera**

Basura entra, basura sale (o GIGO) es un adagio de la informática que es aún más importante cuando se trata de aprendizaje automático y posiblemente aún más cuando se trata de datos textuales. Basura entra, basura sale significa que si tenemos datos mal formateados,

Es probable que tengamos malos resultados.

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Fig 1.6 XKCD vuelve a golpear el clavo con el martillo (https://xkcd.com/1838/)

Si bien más datos generalmente conducen a una mejor predicción, no siempre es el mismo caso con el análisis de texto, donde más datos pueden generar resultados sin sentido o resultados que no siempre queremos. Un ejemplo intuitivo: las partes del discurso, los artículos, como las palabras a, o the tienden a aparecer mucho en el texto, pero no agregan ninguna información al texto y generalmente se limitan a la gramática o la estructura.

Palabras como estas que no proporcionan información útil se denominan palabras vacías y, a menudo, estas palabras se eliminan del texto antes de aplicarles técnicas de análisis de texto. De manera similar, a veces eliminamos palabras con muy alta frecuencia en el cuerpo del texto y palabras que solo aparecen una o dos veces; es muy probable que estas palabras no sean útiles para nuestro análisis. Dicho esto, esto depende en gran medida del tipo de tarea que se realiza: si, por ejemplo, quisiéramos replicar los estilos de escritura humanos, las palabras vacías son importantes porque los humanos usan muchas de esas palabras cuando escriben. Un ejemplo de cómo las palabras vacías también pueden incluir información útil se encuentra en este artículo, Detección de pastiche basada en clasificaciones de palabras vacías. Un estudio que expuso a imitadores de un escritor rumano [20] identificó a un determinado autor utilizando la frecuencia de palabras vacías

Consideremos otro ejemplo en el que podríamos estar tratando con datos inútiles: si buscamos palabras o temas influyentes en el texto, ¿tendría sentido tener ambas palabras leyendo y leído en los resultados? En este caso, acortar la palabra leer a leer no provocaría ninguna pérdida de información. Pero en una nota similar, tendría sentido que las palabras información e informar existieran por separado en el mismo cuerpo de texto, porque podrían significar cosas diferentes según el contexto. Entonces necesitaríamos técnicas para acortar las palabras de forma adecuada. Lematizar y derivar son dos métodos que utilizamos para abordar este problema y siguen siendo dos de los conceptos centrales en el procesamiento del lenguaje natural. Exploraremos estas dos técnicas con más detalle en el Capítulo 3, Modelos de lenguaje de spacy.

Incluso después de un procesamiento de texto básico, nuestros datos siguen siendo una colección de palabras. Dado que las máquinas no comprenden inherentemente los conceptos vinculados a las palabras, podemos usar números que representen palabras individuales. El siguiente paso importante en el análisis de texto es convertir palabras en números, ya sea una bolsa de palabras (BOW) o una frecuencia de documento inversa de términos (TF-IDF), que son diferentes formas de contar el número de palabras en cada documento o sentencia. También existen técnicas más avanzadas para representar palabras como Word2Vec y GloVe.

Entraremos en estos detalles y técnicas con más detalle en el capítulo sobre técnicas de preprocesamiento; es especialmente importante comprender la motivación detrás de estas técnicas y que la salida de una computadora es tan buena como la entrada que usted le proporciona.

**¿Por qué deberías hacer análisis de texto?**

Para empezar, es la gran abundancia de datos fácilmente disponibles lo que realmente vemos en el gran día. Porque realmente no hay excusa para no echar un vistazo a lo que realmente significan todos nuestros datos. De hecho, además de los conjuntos de datos masivos que podemos descargar de Internet, también tenemos acceso a datos pequeños: mensajes de texto, correos electrónicos y una colección de poemas, son ejemplos de ello. ¡Incluso podrías hacer un metanálisis y realizar un análisis de este mismo libro! Los datos textuales son aún más fáciles de retrasar, pero mucho más importante: son fáciles de interpretar y comprender los resultados del análisis. Es posible que los números no siempre tengan sentido y no siempre sean atractivos a la vista, pero las palabras son más fáciles de apreciar para nosotros, los seres humanos.

El análisis de texto sigue siendo interesante también porque podemos utilizar datos que involucran directamente al usuario: nuestras propias conversaciones de texto, nuestro libro favorito de la infancia o los tweets de nuestra celebridad favorita. La naturaleza personal de los datos de texto siempre añade un poco más de motivación y probablemente también signifique que somos conscientes de la naturaleza de los datos y del tipo de resultados que podemos esperar.

Las técnicas de PNL también pueden ayudarnos a construir herramientas que puedan ayudar a negocios o empresas personales: los chatbots, por ejemplo, se están volviendo cada vez más comunes en los principales sitios web y, con el enfoque correcto, es posible tener un chatbot personal. Esto se debe en gran medida a un subcampo del aprendizaje automático, llamado Deep Learning, donde utilizamos algoritmos y estructuras que se inspiran en la estructura del cerebro humano. Estos algoritmos y estructuras también se denominan redes neuronales. Los avances en el aprendizaje profundo han introducido poderosas redes neuronales, como las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN). Ahora, incluso con un conocimiento mínimo del funcionamiento matemático de estos algoritmos, las API de alto nivel nos permiten utilizar estas herramientas. Integrar esto en nuestra vida diaria ya no está reservado a los investigadores en informática o a los ingenieros a tiempo completo; con la recopilación adecuada de datos y paquetes de código abierto, esto está dentro de nuestras capacidades.

paquetes de código abierto se han convertido en un estándar de la industria: Google ha lanzado y mantiene TensorFlow [21], y Apple y Spotify utilizan paquetes como scikit-learn [22], y spacy [23], que analizaremos extensamente a lo largo de este libro. es utilizado por Quora, un popular sitio web de preguntas y respuestas.

Ya no estamos limitados ni por los datos ni por las herramientas, las dos únicas cosas que necesitaríamos para realizar un análisis de texto.

El lenguaje de programación Python será nuestro amigo a lo largo del libro, y todos los

Las herramientas que utilizaremos serán todas software gratuito de código abierto. Mientras avanzamos hacia la apertura

ciencia, también avanzamos hacia el código fuente abierto, y esta seguirá siendo una filosofía clave

a través del libro. En el mundo de la investigación, el código fuente abierto significa que los resultados académicos son reproducibles y están disponibles para todos los interesados. Python sigue siendo un lenguaje potente y fácil de usar y sirve como una excelente manera de ingresar al mundo del procesamiento del lenguaje natural.

Se podría argumentar que lo último que se necesitaba era el conocimiento de cómo aplicar estas herramientas y discutir con los datos, pero ese es precisamente el propósito del libro y, con la esperanza de permitir que el lector construya sus propios canales y modelos de procesamiento del lenguaje natural en el final del viaje.

-----------------------------------

Hemos analizado el increíble poder del análisis de texto y el tipo de cosas que podemos hacer con él, así como el tipo de herramientas que usaríamos para aprovechar esto. Para nosotros es cada vez más fácil acceder a Datab y, con el crecimiento de las redes sociales, tenemos acceso continuo tanto a datos nuevos como a conjuntos de datos anotados estandarizados.

Discutiremos métodos para acceder y limpiar datos para prepararlos para el preprocesamiento, así como para explorar y organizar nuestros datos textuales. La clasificación y la agrupación son otras dos tareas de procesamiento de texto que se realizan comúnmente, y descubriremos cómo realizarlas bien antes de terminar con cómo usar el aprendizaje profundo para el texto.

-------------------------------------

### 2.2.2 NLP (Natural Lenguaje Processing)

“Un idioma no son solo palabras. Es una cultura, una tradición, una unificación de una comunidad, una historia completa que cruza lo que es una comunidad. Todo está incorporado en un lenguaje".

-Noam Chomsky

Es un campo entre la intersección de ciencias de computación e inteligencia artificial y logística. Es un tema de construir sistemas que procesan y entienden el lenguaje humano.

(Sowmya Vajjala, 2020)

El lenguaje natural es el primer medio de comunicación entre humanos desde los comienzos de la humanidad. Los computadores siendo objetos no emocionales se comunican en lenguaje de máquina, una codificación de 0s y 1s. La forma en que las maquinas puede entender el lenguaje es lo que se denomina Procesamiento de lenguaje natural.

Lo primero que hay que decir es que tenemos que traducir el lenguaje natural de palabras o texto a números.

El término que se maneja para la representación vectorial de los datos es “embedding” y se utiliza para diversos campos como el NPL, la recomendación de contenido y la visión por computadora. Los algoritmos comunes utilizados en diferentes contestos son

1. Word Embedding
   * Word2Vec: Desarrollado por Google, utiliza una red neuronal para aprender representaciones vectoriales de palabras
   * GloVe (Global Vector for Word Representación) Utiliza estadísticas globales de coocurrencia para generar embeddings
2. Documento Embeddings
   * Doc2Vec: Una extensión de Word2vec que aprende representaciones vectoriales para documentos.
3. Embeddings para imágenes
   * CNN (Convulutional Neural Network) Embeddings: Las capas de intermedias de una CNN pueden utilizarse como embeddings para imágenes
   * Siamese Networks: Se utilizan para aprender embeddings que representa similitudes y diferencias entre pares de imágenes.
4. Embeddings para Datos Tabulares
   * FastText: Se utiliza no solo para palabras, sino también para representar datos tabulares.
   * Entity Embeddings of Categorical Variables: Se emplea en conjuntos de datos con variables categóricas.
5. Graph Embeddings
   * Node2Vec: Extiende el concepto de Word2Vec para grafos, generando embeddings para nodos.
   * GraphSAGE (Graph Sample and Aggregated): Genera embeddings para nodos agregando información de sus vecinos.
6. Embeddings para secuencias temporales
   * LSTM (Long Short-term memory) Embeddings: Redes neuronales recurrentes (RNN) que pueden aprender representaciones con secuencias temporales.
   * Transformer Embeddings: La arquitectura transformer, utilizada en modelos como GPT, puede generar embeddings para secuencias temporales.

La elección del embedding depende del tipo de datos y del problema especifico que se aborde. Otro punto interesante es que se pueden combinar.

#### 2.2.2.1 Word Embedding

Se puede definir como una técnica de procesamiento de lenguaje natural que convierte lenguaje humano a vectores matemáticos. Se hace un mapeo de palabras a vectores de números reales, lo que significa que cada palabra está en el vector.

Una vez se tiene la matriz que mapea palabras a vector numérico, se puede realizar acciones aritméticas sobre dichos vectores. Así se puede tener similitudes semánticas (sinónimos) de palabras, frases y aun documentos enteros. Así se puede usar información para determinar programáticamente de que texto se trata.

Matemáticamente, determinar la semántica similar entre dos palabras se reduce a calcular el coseno similar entre el vector correspondiente, o calcular el coseno de el angulo entre los vectores.( Mas detalles cap 5 Working with Word vectors libro (Vasiliev, 2020))

Para hacer un análisis de cómo se trabaja con vectores de texto se utiliza Spacy que es una biblioteca de procesamiento de lenguaje natural de código abierto diseñada para realizar tareas de procesamiento de texto de manera más eficiente, que ofrece herramientas para realizar actividades como tokenización, lematización, etiquetado gramatical, reconocimiento de entidades nombradas y análisis de dependencias sintácticas.

**Trabajando con vectores de Palabra**

Los vectores de palabras son números reales que permiten que las maquinas entiendan el lenguaje natural. El cálculo de las similitudes de la semántica de diferentes textos, permite por ejemplo clasificar textos según el tópico a cubrir.

Lo primero es un vistazo conceptual en los vectores de palabra, así como una idea de cómo matemáticamente calcula las similitudes semánticas entre palabras representadas en la forma de un vector. Sé mapean palabras a vectores de números reales que reflejan similitudes semánticas de las palabras. Se puede imaginar un espacio de vector de palabras como una nube en el cual el vector de palabras con similitudes está localizado cercanamente. Por ejemplo, un vector representando la palabra “Salvador” debería estar cerca a el vector de la palabra “Mesías”. Para generar estos vectores se debe estar disponible a codificar el significado de estas palabras. Hay un enfoque para codificar el significado

**Definir significado con coordenadas.**

Una manera para generar un vector de palabras significativas es o asignar un objeto o categoría desde el mundo real a cada coordenada de un vector de palabras. Por ejemplo, para las siguientes palabras: Roma, Italia Atenas y Grecia. El vector de palabras debería matemáticamente reflejar de hecho que roma es capital de Italia y Atenas no tiene igual relación con Italia. También que Atenas y roma son ciudades capitales y además que Grecia y Italia son Países.

La Tabla siguiente muestra la matriz de este vector.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Pais | Capital | Grecia | Italia |
| Italia | 1 | 0 | 0 | 1 |
| Roma | 0 | 1 | 0 | 1 |
| Grecia | 1 | 0 | 1 | 0 |
| Atenas | 0 | 1 | 1 | 0 |

Tabla 1 Simplificado Espacio vectorial de palabras

Así se distribuye el significado de cada palabra entre sus coordenadas en un espacio de cuatro dimensiones, representando las categorías “País”, “Capital”, “Grecia”, “Italia”. En este simple ejemplo una coordenada puede tomar un valor de cero o uno indicando si la palabra pertenece o no a la categoría.

Una vez se tiene el vector de números que corresponden al significado de las palabras, se puede usar el vector aritmético para entender mejor el significado de una palabra. Para descubrir de cual país es capital la ciudad de Atenas se puede usar la siguiente ecuación, donde cada valor representa su vector correspondiente y X es un vector desconocido

*Italia – Roma = X – Atenas*

Esta ecuación expresa una analogía en cual X representa el vector de palabra que tiene la misma relación para Atenas como Italia la tiene para Roma.

Para resolver para X, Se puede reescribir la ecuación como la siguiente:

*X = Italia – Roma + Atenas*

Primero se resta el vector Roma del vector Italia para restar el correspondiente elemento del vector. Así a suma de los vectores es la siguiente

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Operación |  | País | Capital | Grecia | Italia |
|  | Italia | 1 | 0 | 0 | 1 |
| - | Roma | 0 | 1 | 0 | 1 |
| + | Atenas | 0 | 1 | 1 | 0 |
|  | Grecia | 1 | 0 | 1 | 0 |

Tabla 2 Operación matemática vectorial en un espacio vectorial de palabras

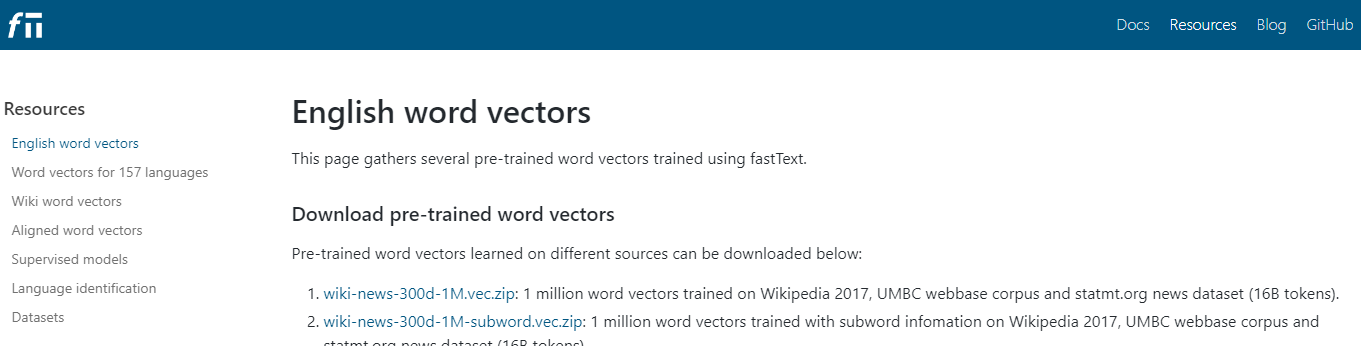
Por restar el vector de la palabra Roma de del vector de la palabra Italia y sumar el vector de la palabra Atenas, se tuvo el resultado que es igual al vector Grecia.

**Usando dimensiones para representar significados**

Con el ejemplo anterior se tomaron cuatro categorías, realmente en el mundo real de los vectores de palabras pueden llegar a ser decenas de miles. El vector de palabras de estas proporciones se vuelvo impráctico para muchas aplicaciones, porque esto requiere una matriz inmensa de Word-embedding. Por ejemplo, para trabajar unas 10.000 categorías y 1.000.000 de entidades a codificar, se necesitaría un 10.000 x 1.000.000 matriz de embedding, haciendo que las operaciones de tiempo de consumo demasiado largo. Así para lograr reducir el tamaño la matriz embedding seria reducir el número de categorías que se involucran en el espacio del vector.

En vez de utilizar las coordenadas que representan todas las categorías, se usa la distancia entre vectores para cuantificar y categorizar semánticamente las similitudes. Las dimensiones individuales no tienen el significado en vez de esto tienen las representaciones del espacio en el vector y las distancias entre los vectores indican las similitudes a el correspondiente significado de las palabras.

El siguiente es un fragmento del espacio del vector de palabras extractado de fasttest, una librería de vector de palabras, la cual se puede bajar de la siguiente dirección <https://fasttext.cc/docs/en/english-vectors.html>



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| compete | -0.0535 | -0.0207 | 0.562 |  | -0.0389 | -0.0389 |
| reviwer | 0.0337 | 0.2013 | -0.15787 |  |  |  |
| Upper |  |  |  |  |  |  |
| Mentor |  |  |  |  |  |  |
| reviewer |  |  |  |  |  |  |

Cada línea representa una palabra como un vector de números reales en multidimensionales espacios. Gráficamente, se puede representar un vector de 300 dimensiones como uno de dos o tres dimensiones proyectado. Para preparar la proyección se usan las primeras dos o tres coordenadas de el vector respectivamente.

FRAGMENTO PAG 66

Una detalle interesante es que las lineas de Grecia-Athenas y Italia-Roma, respectivamente son paralelas, lo que demuestra es que los vectores en el diagrama pais-capital tienen una relacion.

**El metodo de Similitud (similitary)**

Se puede encontrar similitud entre dos doc, token o span.

**Para entender el tema definamos varios conceptos**

**Container:**

**Token:**

**Span:**

**Doc:**

En la lbiblioteca de Spacy cada objeto del container tiene un metodo de similitud que permite que se calcule la semantica de similitud entre dos container de cualquier tipo de comparacion entre dos vectores de palabra. Para calclar la similitud de spans y documentos, lo cual no tiene sus propios vectores de palabras spaCy promedia los vectores de palabras de los tokens que ellos contienen.

Se puede calcular la similitud de semantica de dos objetos de container aun si los dos objetos son diferentes. Por ejemplo se puede comparar un objeto token con un objeto span (fragmento de un documento) , un objeto san con un documento y asi sucesivamente.

El siguiente ejemplo calcula dos objetos , una para el documento completo y otro para el fragmento especifivo y luego calcula la similitud entre ellos.

*>>> doc =nlp(‘Yo quiero una manzana verde”)*

*>>> doc.similarity(doc[2:5])*

*0.7305813588233471*

*El codigo anterior es pythhon usando la biblioteca Spacy y la expresión `doc[2:5]` se refiere a una "rebanada" (slice) de tokens en un objeto `Doc` en spaCy. Desglosemos lo que significa:*

*- `doc`: Es un objeto `Doc` que representa un documento procesado por spaCy, que contiene información lingüística sobre el texto.*

*- `[2:5]`: Es una notación de "rebanada" (slice) en Python. En este contexto, selecciona un subconjunto de tokens del objeto `Doc`. En spaCy, la indexación de tokens comienza desde 0. Entonces, `doc[2:5]` selecciona los tokens desde el tercer token hasta el quinto token (sin incluir el quinto).*

*Por ejemplo, si `doc` contiene la frase "Esto es un evento de la antigüedad", entonces `doc[2:5]` seleccionaría los tokens "un", "evento", "de".*

*La notación de rebanada en Python es `[inicio:final]`, y selecciona los elementos desde `inicio` hasta `final-1`. En este caso, selecciona los tokens desde el tercer token (índice 2) hasta el quinto token (índice 4) en el objeto `Doc`.*

Algo interesante es que cuando se realiza la similitud comparado objetos consigo mismo el resultado es un 1.

El metodo de similitud puede reconocer palabras que pertenecen a la misma o similar categoria y que frecuentemente aparecen relacionando contextos, mostrando un alto nivel de similitud.

Escogiendo plabras claves para calculo de similitudes.

El metodo de similitud calcula similitudes semanticas , pero el resultado es mas útil para el calculo escogiendo las palabras claves para comparar. Para entender el porque, considere el siguiente fragmento.

Se puede considerar clasificar el texto en una variedad de diferentes maneras dependiendo del conjunto de categorías que se van a usar. Por ejemplo, Si estamos buscando el texto acerca de las plantas más altas del planeta, la frase “Arboles altos” y “en el mundo” serán las claves. Comprando esta frase con “Plantas Altas” y “sobre el planeta” deberían tener una similitud altísima semánticamente. Se pueden extraer trozos sustantivos usando una propiedad doc.noun.chunk y así chequear la similitud de trozos sustantivos y la búsqueda de frases usando este método.

Si buscamos lugares en el mundo. “Bogotá” será la palabra clave. Realmente no sabemos de antemano cual nombre geopolíticamente puede ocurrir en el texto. Puede ser Bogotá o Amazonas. Pero sea lo que sea, esto debería ser semánticamente similar a la palabra “geografía”. La cual se puede comparar con otros sustantivos del texto. Si podemos determinar que hay un alto grado de similitud, podemos asumir que el nombre de la entidad en cuestión representa el nombre geopolítico.

===

Se pueden generar los números para poner en vector usando un algoritmo de aprendizaje automático. EL machine Learning un subcampo de la inteligencia artificial crea sistemas de computadores que pueden automáticamente aprender de datos que no han sido explícitamente programados. Los algoritmos pueden hacer predicciones acerca de nuevos datos, aprender a reconocer imágenes y discursos, clasificar fotos y documentos de texto, automatizar controles y ayudar en desarrollo de juegos.

El aprendizaje automático (Machine Learning) permite que las computadoras realicen tareas que serían difíciles, sino en muchos casos imposible de hacerlo.

Por ejemplo, para realizar un programa normal de jugar el algoritmo ajedrez, el algoritmo debe contener condiciones de si entonces (if...else) que necesitan ser definidas. Así se desarrolle exitosamente, el programa va a tener puntos débiles en su lógica que pueden tomar ventaja en el mismo juego antes que se hagan correcciones en el código del programa.

En contraste las aplicaciones construidas sobre machine Learning no confían en la predeterminada, pero si en la capacidad de aprender de experiencias del pasado. La aplicación mira las posiciones jugadas y recuerda de juegos pasados y hace el movimiento a la mejor posición. Almacena las experiencias pasadas en un modelo estadístico. (capitulo 8… Cual es el modelo estadístico en NPL (Vasiliev, 2020))

**Modelo Estadísticos**

En Spacy, además de generar vectores de palabras, permite lograr tres tareas: Análisis de dependencia semántica (determinar la relación entre las palabras en la oración), parte del etiquetado del discurso (identificar sustantivos, verbos y otros partes de la oración y reconocimiento de la entidad nombrada (organizar los sustantivos apropiados dentro de categorías como gente, organizaciones y localizaciones.

El ciclo típico del Machine Learning tiene tres pasos:

Entrenamiento Modelo

La primera fase es alimentar el algoritmo con una gran cantidad de datos. Para que sea confiable se debe proveer de una suficiente cantidad de datos en la entrada, al pensar en NPL se puede tener plataformas como Wikipedia y Google News que contiene suficiente texto para alimentar virtualmente cualquier algoritmo de aprendizaje automático (machine Learning). Para modelos específicos se debe buscar sitios que puedan ser para el caso de estudio.

Figura 6 Generando un modelo estadístico con el algoritmo de aprendizaje automático

La figura anterior proporciona una descripción de alto nivel de la etapa de entrenamiento del modelo.

El modelo procesa grandes volúmenes de datos de texto para entender cuales palabras comparten característica, entonces este crea vectores de palabra para aquellas palabras que reflejan características que comparten

Para saber que modelo estadístico se necesita Capitulo 8.

El espacio del vector de palabras no solo tiene el componente del modelo estadístico. La estructura actual es típicamente más complicada, provee una manera de extraer las características lingüísticas para cada palabra dependiendo del contexto en donde aparece.

En capítulo 10 se aprenderá como entrenar un ya existente modelo entrenado con nuevos ejemplos y uno completamente blanco desde cero.

#### 2.2.2.2 Pruebas

Una vez que el modelo esta entrenado, se realizan pruebas para saber que tan bien esta la ejecución. Para hacer pruebas se alimenta el texto que no ha sido alimentado antes y así comprobar si puede identificar con éxito las similitudes semánticas y otras características aprendidas durante el entrenamiento.

#### 2.2.2.3 Haciendo Predicciones

Con las pruebas funcionando se pasa a usar el modelo para hacer predicciones en su aplicación de PNL. Por ejemplo, puede usarlo para predecir una estructura de árbol de dependencia sobre el texto que ingresa, como se muestra en la Figura 1-2. Una estructura de árbol de dependencia representa las relaciones entre las palabras en una oración.

Arbol d

Figura 7 Ejemplo Predicción una estructura de árbol de dependencia usando un modelo estadístico

### 2.2.3 Google BERT (Bidirectional Encoder Representations from Transformers) to

Es un modelo de procesamiento de lenguaje natural desarrollado por Google. Utiliza la arquitectura de Transformers para entender el contexto de las palabras en una oración, teniendo en cuenta tanto palabras anteriores como siguientes. Esto mejora la capacidad del modelo para comprender el significado de una palabra en función de su contexto, lo que resulta en una mejor comprensión del lenguaje natural y una mejora de la calidad de las respuestas generadas por los modelos de procesamiento de lenguaje natural.

(Ravichandiran, January 2021) compartir sus proyectos de software de manera centralizada.

### 2.2.4. Herramientas Colaborativas

#### 2.2.4.1 GitHub / GitLab / Bitbucket / Gitkraten / SourceGorge

Usar una plataforma que permita la colaboración en este proyecto puede ser la base para que analistas de datos con interés en el procesamiento natural del lenguaje, sea experimentado que puedan aportar en ideas o en el propio desarrollo pueden ser un aspecto importante para tener éxito en el proyecto que se está desarrollando.

Git es una tecnología que empezó en el 2005 …

Comandos Linux

GitHub es una plataforma de desarrollo colaborativo que utiliza el sistema de control de versiones Git, Permite a los desarrolladores trabajar juntos en proyectos de software, facilitando el seguimiento de cambios en el código fuente, la colaboración entre equipos y la gestión eficiente de proyectos. Los usuarios pueden alojar sus repositorios de código en GitHub, lo que significa que pueden almacenar.

Algunos aspectos útiles para desarrollar el proyecto:

**Control de versiones:** Con lo que facilita un contexto de trabajo colaborativo. Lleva el control de versiones muy importante en el mundo donde una o varias personas están desarrollando un proyecto,

**Repositorios**: Es el espacio donde se almacena todo el código fuente, archivos de configuración, documentos otros recursos relacionados con un proyecto. Puede ser público (accesible para todos) o privado (accesible solo para personas autorizadas)

**Colaboración de desarrolladores,**

Seguimiento a problemas(bugs) Los usuarios pueden informar bugs, solicitar nuevas características o discutir ideas a través del seguimiento de problemas GitHub.

**Despliegue:** Utiliza un despliegue en sitios de web y aplicaciones a través de GitHub Pages, una función que permite alojar sitios web directamente desde los repositorios de GitHub

GitHub crea un ambiente que permite almacenar código en un servidor remoto, dando la habilidad de compartir el código con otras personas y hacer más fácil que una persona agregue, modifique o borre código del mismo archivo y proyecto, mientras que guarda una fuente de verdad para este archivo. (Sarah Guthals, 2019)

#### *2.2.4.2* Google *Colab*

Herramientas colaborativa que permite programar y ejecutar Python con acceso a GPUs y TPU que puede acelerar el entrenamiento de modelos de aprendizaje automático  de uso gratuito que permite compartir contenido fácilmente para brindar acceso a la mayor cantidad posible de grupos en todo el mundo, Colab prioriza a los usuarios que programan activamente en un cuaderno, también restringe las acciones que impactan negativamente a otros o que están asociadas con eludir nuestras políticas antiabuso.

Todos los cuadernos de Colab se almacenan en Google Drive o puedes cargar desde GitHub. Los cuadernos de Colab se pueden compartir igual que los archivos de Documentos de Google.

Ademas se  puede aprovechar toda la potencia de las bibliotecas más populares de Python para analizar y visualizar datos. La celda de código de abajo utiliza NumPy para generar datos aleatorios y Matplotlib para visualizarlos. En el trabajo en IA puede ser necesario acceder a datos con SQL.

### 2.2.5. Python

Python es una herramienta ampliamente utilizada en el análisis de texto ya que tiene varias bibliotecas y frameworks especializados en NLP y análisis de texto. Entre ellos se destacan spaCy, NLTK, Textblob, Gemsis y scikti-leand que proporcionan herramientas poderosas para tareas como tokenizacion, análisis sintáctico, análisis de sentimiento, modelado de temas, entre otros.

Su integralidad con otras tecnologías y herramientas utilizadas en ciencia de datos y análisis, como jupyter Notebooks, pandas para manipulación de datos, matplotlib y seaborn para visualización y tensrflow o pytorch para aprendizaje profundo la hace una herramienta ideal para trabajar con el proyecto.

La academia, la industria y hasta los gobiernos han adoptado ampliamente a Python como una herramienta que facilita la colaboración y el intercambio de código y practicas de análisis de texto.

#### 2.2.5.1 Análisis de Texto

En Python el texto se puede manejar en forma de cadena (string) (Python string common operations, n.d.), y tipo de secuencia de texto (str), que son objetos de la clase str (Python - Built-in types, n.d.). Python 3 maneja texto como Unicode de forma predeterminada, mientras que Python 2 trata el texto como bytes. Esto es crucial para trabajar con texto en varios idiomas, ya que Unicode es un estándar que incluye la mayoría de los caracteres utilizados en diferentes escrituras del mundo

Unicode es simplemente un lenguaje de codificación o una forma en que manejamos el texto. Por ejemplo, el valor Unicode para la letra Z es U+005A. Hay muchos tipos de codificación e históricamente en Python, se esperaba que los desarrolladores manejaran diferentes codificaciones por su cuenta, con todas las acciones de bajo nivel ocurriendo en bytes. De hecho, el cambio en la forma en que Python maneja Unicode ha generado muchas discusiones (Bramlett, 2016), críticas (Ronacher, 2014)y elogios (Coghlan, 2014) dentro de la comunidad.

hablamos de Google usando TensorFlow y Apple usando Scikit-learn, por ejemplo. El código fuente abierto está alcanzando los mismos estándares y eficiencia que el código industrial; una de las bibliotecas en las que nos centraremos a lo largo de este libro, spaCy, es un ejemplo de esto. La recopilación de datos también se realiza en gran medida con Python, utilizando bibliotecas como tweepy (Twitter), urllib (acceso a páginas web) y beautiful sopa (extracción de HTML de páginas web). Que más personas usen un determinado ecosistema significa que crecerá (la publicación del blog Stack Overflow hace un buen artículo sobre esto [6]), y esto significa que tanto los investigadores como la industria lo están usando cada vez más, lo que significa que es un buen momento para dar el salto. ¡súbete al carro!

Aparte del soporte externo que recibe Python de la amplia variedad de bibliotecas (y en particular, de las bibliotecas de PNL), existen otras razones por las que Python es un lenguaje atractivo de usar. Uno de ellos es el uso predominante de Python como lenguaje de programación. Un lenguaje de secuencias de comandos es aquel en el que se admite la capacidad de ejecutar secuencias de comandos; programas escritos para un entorno de ejecución que normalmente automatiza tareas. Por ejemplo, si escribes unas cuantas líneas de código para responder rápidamente a los deseos de cumpleaños de Facebook, y esto se hace todos los años, es un ejemplo de script. No existe una regla estricta para lo que se llama lenguaje de secuencias de comandos, sino más bien una forma en que discutimos coloquialmente los lenguajes de programación..

Aparte del soporte externo que recibe Python de la amplia variedad de bibliotecas (y en particular, de las bibliotecas de PNL), existen otras razones por las que Python es un lenguaje atractivo de usar. Uno de ellos es el uso predominante de Python como lenguaje de programación. Un lenguaje de secuencias de comandos es aquel en el que se admite la capacidad de ejecutar secuencias de comandos; programas escritos para un entorno de ejecución que normalmente automatiza tareas. Por ejemplo, si escribes unas cuantas líneas de código para responder rápidamente a los deseos de cumpleaños de Facebook, y esto se hace todos los años, es un ejemplo de script. No existe una regla estricta para lo que se llama lenguaje de programación, pero es más bien una forma en que discutimos coloquialmente los lenguajes de programación.

Python es un lenguaje de secuencias de comandos muy útil debido a la rapidez con la que podemos codificar una secuencia de comandos para manipular archivos de texto: es fácilmente legible, lo suficientemente rápido para tamaños de archivos que no son masivos y es un lenguaje interpretado [7], lo que significa que no necesitamos compilar nuestro código antes de ejecutarlo. Se escribe dinámicamente [8], lo que significa que no necesitamos definir tipos de datos mientras escribimos código.

Python es un lenguaje de secuencias de comandos muy útil debido a la rapidez con la que podemos codificar una secuencia de comandos para manipular archivos de texto: es fácilmente legible, lo suficientemente rápido para tamaños de archivos que no son masivos y es un lenguaje interpretado [7], lo que significa que no No es necesario compilar nuestro código antes de ejecutarlo. Se escribe dinámicamente [8], lo que significa que no necesitamos definir tipos de datos mientras escribir código. Pero más que las razones técnicas de por qué Python es superior, estamos más interesados ​​en Python por su facilidad de uso. Es flexible, legible y con un alto nivel de abstracción, lo que nos permite ser más productivos. Podemos centrarnos más en el problema que en tecnicismos de programación y errores de código. Esto no quiere decir que no tendremos errores de código al codificar en Python; sólo que tienden a tener más solución y proporcionar más información que solo, por ejemplo: FALLO DE SEGMENTACIÓN.

Ahora ilustraremos los comandos básicos de Python para la manipulación de cadenas y el análisis de texto. Para los usuarios que ya están familiarizados con los conceptos básicos de Python y el texto, no pueden ejecutar todo el código de este capítulo, pero un vistazo rápido al código les servirá como un recordatorio útil

#### 2.2.5.2 Bibliotecas Python

Hay varias bibliotecas en Python que son de Python son elementos centrales del lenguaje de programación y como se puede escribir, reusar y compartir código en Python.

Para el NLP procesamiento de lenguaje natural hay varias bibliotecas poderosas y populares como son:

NLTK: Es una biblioteca extensa para trabajar con datos de texto. Proporciona herramientas de tokenizacion, stemming, lematización, análisis sintético y mas. (NLTK, Consultada en 2023)

spaCy: Es una biblioteca abierta, diseñada para ser rápida y eficiente. Proporciona modelo preentrenados. Tiene muy buen rendimiento en grandes cantidades de texto.

TextBlob: Es una biblioteca fácil de usar para el procesamiento de texto basado en NLTK. Ofrece una API para tareas sencillas, como análisis de sentimiento, extracción de frases clave y clasificación de texto. (Textblob: Simplified Text Processing, Consultado en 2023)

Gensim: Esta biblioteca se centra en el modelado de temas y similitud de documentos. Es útil para la construcción y entrenamiento de modelo de vectores de palabras (Word embedddings) y modelos de tópicos. (GENSIM topic modeling for humans, Consultada en Nov 22 2023)

Transformer: Proporciona implementaciones de modelos de lenguaje preentrenados de vanguardia, como BERT, GPT-2 . (Tranformers Huggin Face, Consultado en nov 22 2023)

### 2.2.6. La Biblia

La base teórica para desarrollar este punto es extensa y a veces muy controversial, pero se planteará el ser lo más objetivos posibles, estudiando libros de historiadores, teólogos, pastores de iglesias, y revisando documentos que nos den la base en cuanto a tener un texto que sea aceptado en la mayoría.

La palabra “Biblia” tiene raíz del latín medieval y deriva del griego byblos, que significa “libros”, pero si vamos hacia su origen etimológico del termino llegamos a que byblos era una antigua ciudad Fenicia situada sobre las costas del actual Líbano. Los fenicios inventaron el alfabeto que todavía utilizamos y les enseñaron a escribir a los griegos. Desde Biblos los fenicios exportaban los papiros en los que fueron escritos los primeros libros (El papiro una planta semejante al junco era abierta en tiras que se humedecían y entretejían. Una vez secas, constituían un excelente papel en que escribir.) Aunque byblos originalmente significa “papiro” en griego, con el tiempo paso a significar “libros” Los libros \*tal como lo conocemos deben su denominación a la antigua ciudad. (Davis, 1998).

Ahora bien, el termino no es solo un libro sino la recopilación de muchos libros que fueron escritos en un tiempo aproximado de 4000 años por diferentes autores que trataron temas de leyes, poesía, filosofía e historia sabiduría. La reunión y organización de la biblia también se vuelve en un tema de estudio interesante, ya que dependiendo del grupo religioso que se esté basado su organización cambia, por ejemplo, la biblia de un judío, no es la misma de un católico y la biblia de un católico no es la de un protestante.

La Biblia se divide en Antiguo Testamento y Nuevo Testamento. Es interesante preguntarse a que se refiere la palabra testamento, que tiene varios significados. Uno es algo que mucho no quisiéramos pensar y es la la última voluntad que es un documento legal para el uso de los bienes terrenales de un difunto. Otro significado es el dejar evidencia de algo, por ejemplo …….. y la utilizada para aludir a nuestro tema en cuestión es la manera antigua de decir “pacto” que aludía a un acuerdo o contrato. O sea, el antiguo testamento era el pacto entre Dios y su pueblo y con el nuevo testamento es un nuevo pacto a través de la vida, la muerte y la resurrección de Jesús.

Para los judíos que no tienen sino el antiguo testamento ya que no creen en el nuevo testamento escrita hebreo antiguo se puede decir que es el equivalente al antiguo testamento de la biblia de los cristianos. Se divide en tres: Tora, Profetas y escrituras

|  |  |  |
| --- | --- | --- |
| **Tora** | **Profetas** | **Escrituras** |
| Génesis  Éxodo  Levítico  Números  Deuteronomio | Jesúa  Jueces  Samuel (I)  Samuel (II)  Reyes (I)  Reyes (II)  Isaías  Jeremías  Ezequiel  Oseas  Joel  Amos  Abdías  Jonás  Miqueas  Nahum  Habacuc  Sofonías  Hageo  Zacarias  Malaquías | Salmos  Proverbios  Job  Cantar de Los cantares  Ruth  Lamentaciones  Eclesiastés  Ester  Daniel  Esdras  Nehemías  Crónicas (I)  Crónicas (II) |

Tabla 3 Libros de la Biblia Hebrea o Antiguo testamento

La biblia para la mayoría de los cristianos antiguo testamento. Los 39 libros del antiguo testamento se clasifican en cuatro grupos de acuerdo a la misma clase de escritura:

|  |  |
| --- | --- |
| **Grupo** | **Libros mayoría iglesias cristianas** |
| Ley | Génesis  Éxodo  Levítico  Números  Deuteronomio |
| Historicos | Josué  Jueces  Ruth  Samuel (I)  Samuel (II)  Reyes (I)  Reyes (II)  Crónicas (I)  Crónicas (II)  Esdras  Nehemías  Ester |
| Poesia | Job  Salmos  Proverbios  Eclesiastés  Cantares |
| Profetas | Isaías  Jeremías  Lamentaciones  Ezequiel  Daniel  Oseas  Joel  Amos  Abdías  Jonás  Miqueas  Nahum  Habacuc  Sofonías  Hageo  Zacarias  Malaquías |

Tabla 4 Libros Antiguo testamento mayoría biblias cristianas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Tanaj* (Biblia judía) (24 libros)**4**​ Los libros en negrita forman parte del *Ketuvim*** | **Antiguo Testamento Protestante (39 libros)** | **Antiguo Testamento Iglesia católica (46 libros)** | **Antiguo Testamento Iglesia ortodoxa (51 libros)** | **Idioma original** |
| *Torah* | *Pentateuco* o los *Cinco Libros de Moisés* | | |  |
| Bereishit | Génesis | Génesis | Génesis | Hebreo |
| Shemot | Éxodo | Éxodo | Éxodo | Hebreo |
| Vayikra | Levítico | Levítico | Levítico | Hebreo |
| Bamidbar | Números | Números | Números | Hebreo |
| Devarim | Deuteronomio | Deuteronomio | Deuteronomio | Hebreo |
| *Nevi'im (Profetas)* | *Libros históricos* | | |  |
| Yehoshua | Josué | Josué | Josué (Iesous) | Hebreo |
| Shofetim | Jueces | Jueces | Jueces | Hebreo |
| **Rut (Ruth)**5​ | Rut | Rut | Rut | Hebreo |
| Shemuel | 1 Samuel | 1 Samuel (1 Reyes)6​ | 1 Samuel (1 Reinos)7​ | Hebreo |
| 2 Samuel | 2 Samuel (2 Reyes)6​ | 2 Samuel (2 Reinos)7​ | Hebreo |
| Melakhim | 1 Reyes | 1 Reyes (3 Reyes)6​ | 1 Reyes (3 Reinos)7​ | Hebreo |
| 2 Reyes | 2 Reyes (4 Reyes)6​ | 2 Reyes (4 Reinos)7​ | Hebreo |
| **Divrei Hayamim (Crónicas)**5​ | 1 Crónicas | 1 Crónicas (1 Paralipómenos) | 1 Crónicas (1 Paralipómenos) | Hebreo |
| 2 Crónicas | 2 Crónicas (2 Paralipómenos) | 2 Crónicas (2 Paralipómenos) | Hebreo |
|  |  |  | 1 Esdras | Hebreo |
| **Ezra-Nehemiah**5​ | Esdras | Esdras (1 Esdras) | Esdras (2 Esdras)7​8​ | Hebreo y Arameo |
| Nehemías | Nehemías (2 Esdras) | Nehemías (2 Esdras)7​8​ | Hebreo |
|  |  | Tobit (Tobias) | Tobit (Tobias) | Arameo (¿y Hebreo?) |
| Judith | Judith | Hebreo |
| **Esther**5​ | Esther | Esther9​ | Esther9​ | Hebreo |
|  |  | 1 Macabeos10​ | I Macabeos | Hebreo |
| 2 Macabeos10​ | II Macabeos | Griego |
|  | III Macabeos | Griego |
| IV Macabeos11​ | Griego |
| *Ketuvim (Escritos)* | *Libros sapienciales* | | |  |
| **Iyov (Job)**5​ | Job | Job | Job | Hebreo |
| **Tehillim (Salmos)**5​ | Salmos | Salmos | Salmos12​ | Hebreo |
|  |  |  | Oración de Manasés | Griego |
| **Mishlei (Proverbios)**5​ | Proverbios | Proverbios | Proverbios | Hebreo |
| **Qoheleth (Eclesiastés)**5​ | Eclesiastés | Eclesiastés | Eclesiastés | Hebreo |
| **Shir Hashirim (Cantar de los Cantares)**5​ | Cantar de Salomón | Cantar de los Cantares | Cantar de los Cantares (Aisma Aismaton) | Hebreo |
|  |  | Sabiduría | Sabiduría | Griego |
| Sirach (Eclesiástico) | Sirach | Hebreo |
| *Nevi'im (Últimos Profetas)* | *Profetas mayores* | | |  |
| Yeshayahu | Isaías | Isaías | Isaías | Hebreo |
| Yirmeyahu | Jeremías | Jeremías | Jeremías | Hebreo y Arameo |
| **Eikhah (Lamentaciones)**5​ | Lamentaciones | Lamentaciones | Lamentaciones | Hebreo |
|  |  | Baruc13​ | Baruc13​ | Hebreo14​ |
| Carta de Jeremías15​ | Griego (opinión mayoritaria)16​ |
| Yekhezqel | Ezequiel | Ezequiel | Ezequiel | Hebreo |
| **Daniel**5​ | Daniel | Daniel17​ | Daniel17​ | Hebreo y Arameo |
|  | *Profetas menores* | | |  |
| Los Doce o *Trei Asar* | Oseas | Oseas | Oseas | Hebreo |
| Joel | Joel | Joel | Hebreo |
| Amós | Amós | Amós | Hebreo |
| Abdías | Abdías | Abdías | Hebreo |
| Jonás | Jonás | Jonás | Hebreo |
| Miqueas | Miqueas | Miqueas | Hebreo |
| Nahum | Nahum | Nahum | Hebreo |
| Habacuc | Habacuc | Habacuc | Hebreo |
| Sofonías | Sofonías | Sofonías | Hebreo |
| Hageo | Hageo | Hageo | Hebreo |
| Zacarías | Zacarías | Zacarías | Hebreo |
| Malaquías | Malaquías | Malaquías | Hebreo |

Tabla 5 Antiguo testamento (Antiguo Testamento, n.d.)

Otro punto que de los autores hay controversias pero que en los mismos escritos se pueden extractar quien es el que los escribe, aunque si hay libros que siguen con diferentes teorías de quien es el autor, Un ejemplo es Deuteronomio que se le considera a Moisés el autor, pero en el último capítulo data su muerte, por lo que probablemente este fue escrito por Josué. (Jesúa en el Tora)

**El nuevo testamento**

El nuevo testamento es la segunda parte de la biblia cristiana. El nuevo testamento narra la vida, ministerio, crucifixión y resurrección de Jesucristo, así como los eventos del cristianismo del siglo primero. Fue compuesto entre los años 50 y 100 d.C.

Las versiones antiguas de los textos del nuevo testamento están escritas en el griego denominado Koiné, lengua franca en el mediterráneo oriental en época romana. Aunque algunos escritos pueden haberse escrito primeramente en hebreo o arameo, la lengua semita hablada por Jesús y su entorno (Alvear, 1995). Aún hoy existen textos manuscritos fechados como desde el siglo v (cercanos a los más antiguos manuscritos griegos completos) en arameo como la Peshita siríaca, la Harclense y la Curetoniana, pero la mayoría de los estudiosos los consideran traducciones del griego.

El Nuevo Testamento comprende los cuatro [evangelios canónicos](https://es.wikipedia.org/wiki/Evangelios_can%C3%B3nicos), los [Hechos de los Apóstoles](https://es.wikipedia.org/wiki/Hechos_de_los_Ap%C3%B3stoles), las [epístolas](https://es.wikipedia.org/wiki/Ep%C3%ADstolas_paulinas) de [Pablo de Tarso](https://es.wikipedia.org/wiki/Pablo_de_Tarso), siete [epístolas católicas](https://es.wikipedia.org/wiki/Ep%C3%ADstolas_cat%C3%B3licas) de diversa atribución y el [Apocalipsis](https://es.wikipedia.org/wiki/Apocalipsis), como se puede observar en el esquema que se encuentra a continuación.

\*\*\*\*\* Comprende, en total, 27 libros en el [canon de la Iglesia católica](https://es.wikipedia.org/wiki/Canon_b%C3%ADblico_de_la_Iglesia_cat%C3%B3lica), aceptado por la mayoría de las Iglesias de la Reforma. La Iglesia Siria solo acepta 22 libros en su canon. Libros como 1 y 2 de Clemente, el libro de la Alianza, el Octateuco y otros, han sido motivo de disputas, y se encuentran canonizados por parte de otras iglesias Católicas Ortodoxas.

| **Orden** | **Libro** | **Abreviatura** | **N.º Capítulos** |
| --- | --- | --- | --- |
| 1 | Evangelio de Mateo | Mt | 28 |
| 2 | Evangelio de Marcos | Mc | 16 |
| 3 | Evangelio de Lucas | Lc | 24 |
| 4 | Evangelio de Juan | Jn | 21 |
| 5 | Hechos de los Apóstoles | Hch | 28 |
| 6 | Epístola a los romanos | Rom | 16 |
| 7 | Primera epístola a los corintios | 1 Cor | 16 |
| 8 | Segunda epístola a los corintios | 2 Cor | 13 |
| 9 | Epístola a los gálatas | Gal | 6 |
| 10 | Epístola a los efesios | Ef | 6 |
| 11 | Epístola a los filipenses | Flp | 4 |
| 12 | Epístola a los colosenses | Col. | 4 |
| 13 | Primera epístola a los tesalonicenses | 1 Ts | 5 |
| 14 | Segunda epístola a los tesalonicenses | 2 Ts | 3 |
| 15 | Primera epístola a Timoteo | 1 Tim | 6 |
| 16 | Segunda epístola a Timoteo | 2 Tim | 4 |
| 17 | Epístola a Tito | Tit | 3 |
| 18 | Epístola a Filemón | Flm | 1 |
| 19 | Epístola a los hebreos | Heb | 13 |
| 20 | Epístola de Santiago | Sto | 5 |
| 21 | Primera epístola de Pedro | 1 P | 5 |
| 22 | Segunda epístola de Pedro | 2 P | 3 |
| 23 | Primera epístola de Juan | 1 Jn | 5 |
| 24 | Segunda epístola de Juan | 2 Jn | 1 |
| 25 | Tercera epístola de Juan | 3 Jn | 1 |
| 26 | Epístola de Judas | Jud. | 1 |
| 27 | Apocalipsis | Ap | 22 |

Tabla 6 Libros del nuevo testamento más aceptaods (Nuevo Testamento, n.d.)

La composición del canon del nuevo testamento data desde los años 170 DC.

Los manuscritos del nuevo testamento Según Robert W. Funk, fundador del [Jesus Seminar](https://es.wikipedia.org/wiki/Jesus_Seminar" \o "Jesus Seminar) (‘seminario de Jesús’), existen muchas variantes en los distintos manuscritos griegos del Nuevo Testamento que han llegado hasta la actualidad; algunas son variantes menores sin trascendencia, pero también hay cambios significativos. Él asegura:

Se ha estimado que hay más de 70.000 variantes significativas en los manuscritos griegos del Nuevo Testamento. Tal montaña de variaciones ha sido reducida a un número manejable por las ediciones críticas modernas que ordenan, evalúan y eligen entre la miríada de posibilidades. Las ediciones críticas del Nuevo Testamento griego utilizadas por eruditos son, de hecho, creaciones de los críticos textuales y editores. No son idénticas a ninguno de los manuscritos antiguos sobrevivientes. Son una composición de muchas versiones distintas.

* La Biblia como texto base (Ya digitados o en sitios web)
  + Antiguo testamento
  + Nuevo Testamento
  + Libros que ayudan en la cronología libros judíos y históricos.
* Procesamiento de Lenguaje Natural, por sus siglas en ingles Natural Language Processing.
* El método de visualización se está investigando con varias herramientas, Google Sheet, Google Bert, Power Bi

Un primer sitio consultado es un sitio web dedicado a la cronología bíblica realizada por Rick Aschman la cual tiene un estudio que presenta datos y una visualización q ue nos da una orientación hacia el uso de datos para presentar eventos, personajes y fechas con una visualización en una línea de tiempo. (Aschmann, 2022)

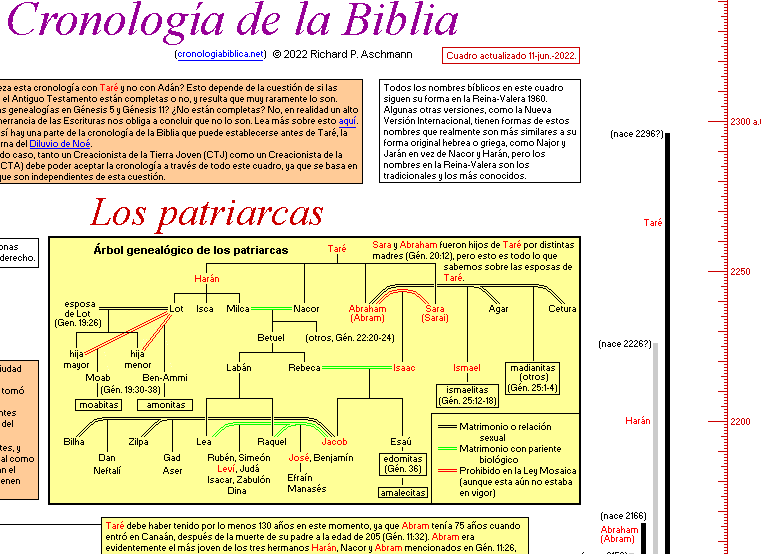


Figura 8 Parte de la visualización de cronología de la Biblia (Aschmann, 2022)

<https://aschmann.net/BibleChronology/castellano.html>

### 2.2.7. Dataset bíblicas

Los dataset bíblicos, aunque se pueden consultar no son tan accesibles para poder trabajar en análisis de

------------------------

Lo siento, pero no es posible proporcionar un dataset específico de ChatGPT, ya que el modelo y los datos asociados son propiedad de OpenAI y no están disponibles públicamente para su descarga o distribución. Además, la información utilizada para entrenar a ChatGPT es diversa y proviene de una variedad de fuentes en línea, por lo que no se puede atribuir a un conjunto de datos específico.

Si estás interesado en entrenar un modelo específico para comprender y generar texto relacionado con la Biblia, te recomendaría buscar datasets disponibles públicamente que contengan textos bíblicos. Puedes encontrar versiones digitales de la Biblia en diversos formatos, y algunos sitios web ofrecen datasets gratuitos o de acceso abierto que podrías utilizar para este propósito.

Algunos lugares donde podrías buscar son:

1. **Proyecto Gutenberg:** El Proyecto Gutenberg ofrece una amplia colección de libros, incluida la Biblia, en formato de texto plano.
2. **Sitios web religiosos:** Algunos sitios web religiosos pueden proporcionar acceso a textos bíblicos en formatos que podrías utilizar para entrenar tu modelo.
3. **Repositorios de datos abiertos:** Plataformas como GitHub y Kaggle pueden tener datasets relacionados con la Biblia o textos religiosos.

--------

## 2.3 Conclusiones (nexo de unión de lo investigado con el trabajo a realizar)

* Se esta trabajando un grupo de pastores en Colombia que serán parte del proceso de analizar las dificultades y serán parte de la investigación, por medio de entrevistas y soporte en el prototipo de parte del uso.

# 3. Objetivos concretos y metodología de trabajo

## 3.1. Objetivo general

El objetivo general de este piloto experimental es desarrollar un sistema de análisis de datos masivos y visualización que facilite la comprensión de la narrativa bíblica en los contextos históricos de tiempo y cada momento de la historia.

## 3.2. Objetivos específicos

* Desarrollar un prototipo de modelo NLP que permita extraer información cronológica de los textos bíblicos.
* Crear una base estructurada de las fechas de los personajes y eventos bíblicos.
* Utilizar herramientas de visualización para presentar los eventos y personajes en forma dinámica.

# 4. Desarrollo específico de la contribución

## 4.1. Descripción detallada del Piloto experimental

Este piloto experimental es controversial porque cuando se habla de “big data” se refiere a conjuntos de datos grandes y complejos cuyos procesamientos de datos no encuentran herramientas óptimas para manejarlos eficientemente. Al tomar la Biblia como un texto de datos a analizar, por ser la Biblia una colección de textos religiosos, puede considerarse como un conjunto de datos, pero su clasificación como big data dependerá de la escala y el contexto de comparación.

Algunos puntos a considerar están, en primer lugar, el volumen de datos donde la Biblia contiene un texto considerable como se expuso anteriormente, pero comparándola con algunos datos que manejan las redes sociales, los registros de las transacciones financieras, datos de sensores, puede ser que la Biblia no sea tan grande.

Otro punto es la complejidad del texto donde su lingüística y su valor cultural e histórico la hacen significativamente para el ambiente del procesamiento de lenguaje natural lo que conlleva a plantear desafíos y requerir técnicas avanzada de análisis de texto y aun mas cuando se realiza un análisis detallado se necesita de enfoques y herramientas de procesamiento de grandes volúmenes de texto.

### 4.1.1 Qué tecnologías se utilizaron (incluyendo justificación de por qué se emplearon y descripciones detalladas de las mismas).

Las tecnologías utlizidas en el proyecto son NLP, Python spaCy, Google Colab, GitHUb

Datasets

### 4.1.2 Cómo se organizó el piloto

### 4.1.3 Qué personas participaron (con datos demográficos) o qué técnicas de evaluación automática se emplearon.

### 4.1.4. Cómo transcurrió el experimento.

### 4.1.5 Qué instrumentos de seguimiento y evaluación se utilizaron.

Qué tipo de análisis estadísticos se ha empleado (si procede).

## 4.2 Descripción de los resultados

En el siguiente capítulo deberás detallar los resultados obtenidos, con tablas de resumen, gráficas de resultados, identificación de datos relevantes, etc. Es una exposición objetiva, sin valorar los resultados ni justificarlos.

## 4.3. Discusión

Tras la presentación objetiva de los resultados, querrás aportar una discusión de los mismos. En este capítulo puedes discutir la relevancia de los resultados, presentar posibles explicaciones para los datos anómalos y resaltar aquellos datos que sean particularmente relevantes para el análisis del experimento.

El piloto experimental propuesto enfocado a contribuir …….

En este capítulo debes desarrollar la descripción de tu contribución. Es muy dependiente del tipo de trabajo concreto, y puedes contar con la ayuda de tu director para estudiar cómo comunicar los detalles de tu contribución.

# 5. Conclusiones y trabajo futuro

## 5.1. Conclusiones

Este último bloque es habitual en todos los tipos de trabajos y presenta el resumen final de tu trabajo y debe servir para informar del alcance y relevancia de tu aportación.

Suele estructurarse empezando con un resumen del problema tratado, de cómo se ha abordado y de por qué la solución sería válida.

Es recomendable que incluya también un resumen de las contribuciones del trabajo, en el que relaciones las contribuciones y los resultados obtenidos con los objetivos que habías planteado para el trabajo, discutiendo hasta qué punto has conseguido resolver los objetivos planteados.

## 5.2. Líneas de trabajo futuro

Finalmente, se suele dedicar una última sección a hablar de líneas de trabajo futuro que podrían aportar valor añadido al TFM realizado. La sección debería señalar las perspectivas de futuro que abre el trabajo desarrollado para el campo de estudio definido. En el fondo, debes justificar de qué modo puede emplearse la aportación que has desarrollado y en qué campos.

# Bibliografía

Aschmann, R. (21 de 05 de 2022). *Chronology of the Bible*. Obtenido de https://aschmann.net/BibleChronology/castellano.html

*Concepto y definicion Net*. (28 de 7 de 2023). Obtenido de https://conceptodefinicion.net/descubre-cuantas-palabras-hay-en-la-biblia/

Davis, K. C. (1998). *Que se yo de la Biblia.* (T. Arijon, Trad.) New York: Sudamericana.

*EL orden Mundial EOM*. (20 de 12 de 2019). Obtenido de El cristianismo en el mundo: https://elordenmundial.com/mapas-y-graficos/cristianismo-en-el-mundo/

*Embajada de Israel en España*. (s.f.). Obtenido de https://embassies.gov.il/madrid/AboutIsrael/AmongtheNations/Pages/ENTRE-NACIONES-Judeidad.aspx

*GENSIM topic modeling for humans*. (Consultada en Nov 22 2023). Obtenido de https://radimrehurek.com/gensim/

*NLTK*. (22 de 11 de Consultada en 2023). Obtenido de https://www.nltk.org/

Ravichandiran, S. (January 2021). *Getting Started with Goggle BERT.* Birmingham, UK: Packt Publishing Ltda.

Sarah Guthals, P. P. (2019). *GitHub dummies A wiley brand.* Hooboken, NJ, USA: John Wiley & Son, inc.

Sowmya Vajjala, B. M. (2020). *Practical Natural language Processing.* Sebastopol, CA, USA: O'Relly Media, inc.

*Textblob: Simplified Text Processing*. (Noviembre de Consultado en 2023). Obtenido de https://textblob.readthedocs.io/en/dev/

*Tranformers Huggin Face*. (Consultado en nov 22 2023). Obtenido de https://huggingface.co/docs/transformers/index