

TRABAJO FIN DE GRADO

INGENIER´IA INFORMÁTICA

**Modelos de Generación de Texto en Bases de Datos Orientadas a Grafos**

**Autor**

Alberto López Povedano

**Directores**

Juan Francisco Huete Guadix

**Ciencias de la computación e Inteligencia Artificial**



Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, noviembre de 2022

**Modelos de Generación de Texto en Bases de Datos Orientadas a Grafos**

Alberto López Povedano

**Palabras clave:**  lenguaje natural, generación de texto, grafos, modelo del lenguaje, n-gramas

**Resumen:**

Desarrollo de un sistema que tiene como objetivo sugerir recomendaciones a un usuario escribiendo texto en castellano. El sistema implementa modelación del lenguaje con n-gramas construido sobre una base de datos orientada a grafos.

**Text Generation Models in Graph-Oriented Databases**

Alberto López Povedano

**Keywords:** natural language, text generation, graphs, language model, n-grams

**Abstract:**

Development of a system that aims to suggest recommendations to a user typing text in Spanish. The system implements language modeling with n-grams built on a graph-oriented database.

Yo, **Alberto López Povedano**, alumno de la titulación Ingeniería Informática de

la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación** **de la Universidad de Granada**, con DNI 77448870G, autorizo la ubicación de la siguiente

copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser

consultado por las personas que lo deseen.

Fdo: Alberto López Povedano

Granada a 08 de Noviembre de 2022

D. **Juan Francisco Huete Guadix**, Catedrático en el departamento de Ciencias de la Computación e Inteligencia Artificial en la Universidad de Granada.

**Informan:**

Que el presente trabajo, titulado **Modelos de generación de texto basados en bases de datos orientados a grafos**, ha sido realizado bajo su supervisión por **Alberto López Povedano**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 8 de noviembre de 2022.

**Los directores:**

**Juan Francisco Huete Guadix**

**Agradecimientos**

A Juan Francisco Huete Guadix por la supervisión en la realización del trabajo de final de grado y su inestimable ayuda aportando recursos didácticos, sugerencias y nuevos enfoques.

A Ernesto Martínez del Prieto por dirigirme para sentar las bases del sistema.

Índice

[Figuras 13](#_Toc119083621)

[Tablas 14](#_Toc119083622)

[Términos 15](#_Toc119083623)

[Abreviaciones 16](#_Toc119083624)

[1. Introducción 17](#_Toc119083625)

[1.1. Propósito 17](#_Toc119083626)

[1.2. Justificación 17](#_Toc119083627)

[1.3. Objetivos 18](#_Toc119083628)

[1.3.1. Objetivos Primarios 18](#_Toc119083629)

[1.3.2. Objetivos Secundarios 18](#_Toc119083630)

[1.3.3. Requisitos no funcionales 19](#_Toc119083631)

[1.4. Metodología 19](#_Toc119083632)

[1.4.1. Terminología 20](#_Toc119083633)

[1.4.2. Licencia 20](#_Toc119083634)

[1.5. Estado del arte 20](#_Toc119083635)

[1.5.1. GPT-3 20](#_Toc119083636)

[2. Fundamentos Teóricos 21](#_Toc119083637)

[2.1. Modelación del lenguaje 21](#_Toc119083639)

[2.1.1. Modelos de N-gramas 21](#_Toc119083640)

[2.2. Grafos 23](#_Toc119083641)

[3. GPG – Grafo Preentrenado Generativo 25](#_Toc119083642)

[3.1. Conjunto de datos 25](#_Toc119083644)

[3.1.1. Fuente de los datos 25](#_Toc119083645)

[3.1.2. Limpieza de los datos 26](#_Toc119083646)

[3.2. Arquitectura del sistema 27](#_Toc119083647)

[3.2.1. Flujo de control 28](#_Toc119083648)

[3.2.2. Flujo de datos 30](#_Toc119083649)

[3.3. Implementación del sistema 30](#_Toc119083650)

[3.3.1. Grafos creados 31](#_Toc119083651)

[3.3.2. Funcionalidades 31](#_Toc119083652)

[3.4. Evaluación 31](#_Toc119083653)

[3.4.1. Perplexity 31](#_Toc119083654)

[4. Demo técnica 34](#_Toc119083655)

[5. Conclusiones 36](#_Toc119083656)

[5.1. Conclusiones sobre objetivos 36](#_Toc119083659)

[5.2. Trabajo Futuro 36](#_Toc119083660)

[5.3. Valoración personal 36](#_Toc119083661)

[6. Tecnologías 37](#_Toc119083662)

[6.1. Python 37](#_Toc119083664)

[6.2. Docker 38](#_Toc119083665)

[6.3. ArangoDB 38](#_Toc119083666)

[6.4. Git 38](#_Toc119083667)

[7. Bibliografía 40](#_Toc119083668)

# Figuras

[Figura 1.1: Tablero Kanban a día 14/08/2022 20](#_Toc119069342)

[Figura 2.1: Formación de n-gramas usando como token las palabras 22](#_Toc119069343)

[Figura 2.2: Ejemplo de grafo de unigramas a nivel de palabras 23](#_Toc119069344)

[Figura 3.1: Flujo de control global de GPG 28](#_Toc119069345)

[Figura 3.2: Flujo de control del menú de GPG 29](#_Toc119069346)

[Figura 3.3: Flujo de control del entrenamiento del modelo 30](#_Toc119069347)

# Tablas

[Tabla 3.1: Número de palabras y caracteres del conjunto de entrenamiento 26](#_Toc119069348)

[Tabla 3.2: Número de palabras y caracteres del conjunto de prueba 26](#_Toc119069349)

[Tabla 3.3: Resultados perplexity sobre corpus de GPG 32](#_Toc119069350)

[Tabla 3.4: Resultados perplexity sobre corpus del Wall Street Journal 32](#_Toc119069351)

# Términos

**Chatbot** Sistema software cuyo propósito es simular conversaciones con

usuario humano.

**Corpus** Colección de texto o audio oficial perteneciente a un lenguaje

organizado en conjuntos de datos.

**Graphical User Interface** Interfaz de usuario gráfica.

**Object Oriented Programming** Modelo de programación que organiza el

diseño del software en torno a la idea de objeto, una entidad que

contiene datos y código en forma de métodos.

**Object-Relational Mapping** Método para convertir los datos almacenados en

una base de datos a un modelo definido en un lenguaje orientado a

objetos.

**Perplexity** Métrica utilizada comúnmente para realizar evaluaciones de

modelos del lenguaje basados en n-gramas.

**Plugin** Extensión de un programa software que añade nueva funcionalidad sin

alterar la proporcionada por el programa original.

**Smoothing** En el terreno estadístico hace referencia al hecho de crear una

función que otorgue los considerados como patrones importantes en los

datos a la vez que deja lo considerado como ruido fuera de ellos.

**Sprint** Ciclo de tiempo acotado por un equipo en el desarrollo del software

siguiendo los principios de las metodologías ágiles.

**Token** Unidad mínima que puede representar hechos, emociones, etc… En

este documento es utilizado para designar la unidad atómica que

consideramos en el análisis del texto.

**Tokenización** Acción de convertir datos complejos en una o una colección de

unidades individuales conocidos como tokens.

# Abreviaciones

**GUI** Graphical User Interface.

**OOP** Object Oriented Programming.

**ORM** Object-Relational Mapping.

**NLP** Natural Language Processing.

# Introducción

Los primeros sistemas conversaciones datan de 1966 y 1971, con los *chatbots* ELIZA y PARRY respectivamente. Ambos sistemas tuvieron cierta relevancia y plantearon nuevas cuestiones éticas en cuanto a la privacidad de datos de los usuarios, sin embargo, fue con el desarrollo de la arquitectura del sistema GUS (1977) que se establecería el paradigma predominante en el desarrollo de sistemas conversaciones durante los 30 próximos años.

(Jurafsky & Martin, Bibliographical and Historical Notes, 2022)

Las aplicaciones del procesamiento del lenguaje natural por una computadora son increíblemente extensas, por mencionar alguna de ellas:

* Traducción automática de textos
* Clasificación de información
* Detección de estados emocionales
* Generación de texto
* Etcétera

Por sus grandes aplicaciones y debido al gran interés que despertó el terreno de la inteligencia artificial, el procesamiento del lenguaje natural se ha convertido en un campo de la inteligencia artificial con décadas de estudio a sus espaldas. Varias son las técnicas que se han aplicado en su investigación, hablamos de sistemas basados en reglas, aprendizaje automático mediante métodos estadísticos y redes neuronales.

En particular la modelización del lenguaje mediante la implementación de modelos basados en n-gramas, modelos probabilísticos, es la técnica utilizada en esta ocasión.

## Propósito

El propósito de este trabajo es el de crear un software que nos sugiera recomendaciones, en forma de palabras u oraciones, en función de las palabras introducidas por el usuario. Este proyecto surge de la idea de crear un sistema que facilite la escritura de textos en un idioma distinto al del usuario. Idea propuesta por el antiguo trabajador de la empresa 8Belts Ernesto Martínez del Prieto. Al día de publicarse, y durante gran parte del desarrollo no ha existido ningún enlace entre este trabajo y dicha empresa.

## Justificación

Las razones que me han llevado a seleccionar este tema para el desarrollo de mi tesis de final de grado son:

* El interés y deseo de introducirse en el estudio de las técnicas de aprendizaje automático.
* Mejorar mi habilidad con el lenguaje de programación Python y la tecnología de contenedores Docker.
* Conocer y asentar las bases en el campo del procesamiento del lenguaje natural.

## Objetivos

Teniendo como objetivo general de este proyecto aprender, reforzar, y poner en práctica los conocimientos más fundamentales de los modelos del lenguaje por métodos estadísticos, mediante la propuesta y desarrollo de un sistema básico de generación de texto, se plantean los siguientes objetivos:

### Objetivos Primarios

1. Obtener una **fuente de datos** para el entramiento del modelo.
2. Implementar una forma de leer y almacenar los datos en una **base de datos** adecuada.
3. Representar los datos en una **estructura de datos** que nos permita representar las conexiones entre las diferentes palabras.
4. Construir **dos modelos del lenguaje**. El primero enfocando los n-gramas a nivel de caracteres y el segundo cambiando el enfoque a nivel de palabras.
   1. El modelo a nivel de caracteres **estará compuesto por 3 submodelos**, a nivel de unigramas, bigramas, trigramas y cuatrigramas.
   2. El modelo a nivel de palabras **estará compuesto por 3 submodelos**, a nivel de unigramas, bigramas y trigramas.
5. Implementar **funcionalidades que nos permitan interactuar y realizar consultas** a nuestro modelo.
   1. Obtener **lista de tokens, ordenada por frecuencia**, que aparezcan a continuación de un token dado.
   2. **Sugerir un token** a continuación de una entrada dada.
   3. Poder **autocompletar palabras**.
   4. Poder **formar frases sencillas**.
   5. Habilitar la interacción mediante **una interfaz**.

### Objetivos Secundarios

1. Configurar un **contenedor** que permita un despliegue rápido y sencillo de la base de datos.
2. Realizar una **evaluación** que nos aporte información de como de bien funciona nuestro modelo.
3. Escribir un **fichero *readme.md*** que aporte instrucciones para el despliegue y uso del proyecto.
4. Montar una **demo** mediante una GUI que ofrezca al usuario sugerencias a medida que este va escribiendo.

### Requisitos no funcionales

Durante el proceso de desarrollo al igual que se consiguen los objetivos mencionados se han de tener los siguientes principios en mente:

1. Como en todo proyecto software se ha de **mantener una estructura de ficheros** y directorios que sea **clara y organizada**. Esta es detallada en la Sección 3.2.1\*\*
2. Los datos utilizados se obtendrán de **fuentes que aporten una licencia libre** para su uso.
3. El código ha de ser reutilizable y sencillo de comprender. La longitud de **comentarios en el texto ha de ser breve** y **resaltar aquello que no se puede extraer con facilidad** al leer el código.
4. El sistema ha de ser poder **desplegado con relativa facilidad en diferentes equipos haciendo pocas instalaciones de dependencias**, para facilitar tanto el desarrollo como para posibilitar el uso o desarrollo por diferentes usuarios.

## Metodología

La metodología utilizada en el desarrollo del proyecto no ha respetado estrictamente las normas y formas de proceder de una metodología ya conocida, pero si se ha inspirado y aproximado al terreno de las conocidas como metodologías ágiles.

En este grupo encontramos principios como la organización del trabajo en torno a *sprints,* identificar el trabajo en pequeñas tareas lo más atómicas posibles otorgándoles una estimación de tiempo o el uso de herramientas como los tableros Kanban.

De las prácticas mencionadas anteriormente se ha llevado a cabo principalmente la definición de pequeñas tareas a realizar siendo representadas en tarjetas de un tablero Kanban proporcionado por la plataforma Trello. A continuación, en la Figura 1.1

se aporta un ejemplo de cómo se vería el tablero Kanban utilizado:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 1.1: Tablero Kanban a día 14/08/2022

### Terminología

Con el fin de evitar malentendidos y búsquedas confusas de términos usados en este documento se utilizarán términos en inglés, especialmente en los casos que la traducción de un término al castellano no produzca búsquedas satisfactorias o sea incluso confundido con acepciones relativas a otros campos del conocimiento. La terminología se utiliza, así, en castellano e inglés indistintamente. A la hora de utilizar términos en inglés, estos serán escritos utilizando cursiva.

En las primeras secciones no numeradas de este documento se han incluido un listado de Términos y otro de Abreviaciones, que exponen de forma breve los conceptos referenciados.

### Licencia

Gracias a los conocimientos adquiridos en la asignatura: Dirección y gestión de proyectos. Se ha decidido lanzar el proyecto en un repositorio de GitHub público bajo la licencia GNU General Public License v3.0. Por lo tanto se entiende que el código es *Free Software* según la definición otorgada por (Free Software Foundation, 2022).

Es importante resumir que esto implica la libre utilización del código, sin importar que sea un uso comercial, teniendo como condición el mantenimiento de la licencia utilizada en proyectos derivados a partir de este (*copyleft*).

## Estado del arte

El procesamiento del lenguaje natural

\*\* Mencionar transformadores

### GPT-3

\*\* Hablar sobre el modelo GPT-3 desarrollado por la empresa OpenAI

# Fundamentos Teóricos



## Modelación del lenguaje

La modelación del lenguaje es un término que engloba todos aquellos métodos que intentan solucionar el problema de modelar el lenguaje natural. La modelación del lenguaje tiene como objetivo permitir el aprendizaje de la máquina sobre el lenguaje para resolver problemas determinados.

Existen diferentes aproximaciones para construir un modelo del lenguaje, en concreto los modelos que funcionan mediante la asignación de altas probabilidades a frases bien estructuradas y formadas han de tener en cuenta la asunción de Márkov.

De primeras sin tener en cuenta la asunción de Márkov, para calcular la probabilidad de aparición de una frase se desglosaría utilizando la regla de la cadena en:

Como se puede observar este cálculo es bastante laborioso, pero según (Gudivada, Rao Vijay, & Raghavan, 2015) la asunción de Márkov nos permite, aplicado en el contexto de modelos del lenguaje, presumir que la probabilidad de la siguiente palabra en una secuencia depende únicamente de la palabra previa o secuencia de palabras previas, siendo un número pequeño de palabras. El tamaño de indica el -orden del proceso de Márkov. Simplificando la ecuación anterior:

Entendiendo que hace referencia a la probabilidad de ver tras la secuencia de palabras

Dentro de la modelación del lenguaje se han desarrollado varios modelos diferentes con sus propias ventajas y desventajas. Encontramos modelos exponenciales, modelos de n-gramas y modelos del lenguaje neuronales. Como se puede asumir tras la explicación del proceso de Márkov, los modelos de n-gramas son en los que se va a centrar el proyecto.

### Modelos de N-gramas

Los modelos de n-gramas puede ser definidos como una cadena de Márkov de orden n-1. En ellos se trabaja con una distribución de probabilidad sobre las secuencias de palabras o caracteres. Estos modelos se caracterizan por trabajar a base de n-gramas. Los n-gramas son conjuntos de datos formados por palabras, caracteres o sílabas, esto depende en la *tokenización* elegida a la hora de analizar el *corpus* de texto*.*

Dependiendo de la *tokenización* elegida se pueden clasificar diferentes tipos de modelos. Lo más conocidos son los que dividen y analizan el *corpus* a nivel de palabras haciendo uso de espacios y signos de puntuación como separadores. Diferenciamos entre unigramas, bigramas, trigramas, … en función del número de *tokens* que formen el n-grama. Un ejemplo de lo explicado se puede visualizar en la Figura 2.1:

Texto

Descripción generada automáticamente

Figura 2.1: Formación de n-gramas usando como token las palabras

Para el desarrollo de este proyecto se ha construido dos modelos del lenguaje a nivel de palabras y a nivel de caracteres. Cada uno de estos cuenta con otros 3 submodelos que analizan los tokens a nivel de unigramas, bigramas y trigramas.

#### Modelos de N-gramas a nivel de caracteres

Los modelos basados en n-caracteres presentan pocas aplicaciones como se expone en (Russell & Norvig, N-gram character models, 2009), principalmente destacan en la identificación del lenguaje de un texto escrito, aunque también son usados para detectar fallos ortográficos. En este caso se está utilizando un modelo de caracteres para poder ofrecer sugerencias que completen palabras a medio escribir.

Es importante tener en cuenta que el conjunto de unigramas, bigramas y trigramas posibles a nivel de caracteres en el lenguaje español es mucho menor a la explosión combinatoria resultante al analizar n-gramas a nivel de palabras. Esto se debe a que los vocabularios que analizamos en ambos casos son en tamaño muy diferentes. Para visualizar esto pensemos en que la cota superior de posibles n-gramas por nivel puede ser calculada como , siendo el vocabulario del lenguaje a analizar y el nivel de n-grama. Estos números quedan reducidos por la imposibilidad de algunas combinaciones de letras o palabras en el lenguaje.

#### Modelos de N-gramas a nivel de palabras

Los modelos de n-gramas de palabras afrontan el problema conocido como *data sparsity* al presentar vocabularios muchos más extensos. Este problema aparece dada la explosión combinatoria que se produce al querer tener un mayor tamaño en los n-gramas construidos, pues al presentar mayores niveles más contexto podremos tener en cuenta y por lo tanto nuestras respuestas tendrán mayor parecido con la realidad.

El problema de intentar hacer los n-gramas de gran tamaño es que cuanto mayor es el tamaño más posibilidad hay de que aparezcan n-gramas que no hayan sido observados en el conjunto de datos utilizado para el entrenamiento, también se considera que a partir de los trigramas, es decir, haciendo uso de tetragramas nuestro modelo memoriza en vez de generar texto. Para enfrentar este problema se hace uso comúnmente de técnicas de *smoothing*, y no se suele analizar n-gramas con un orden mayor a

(Russell & Norvig, N-gram word models, 2009)

También se pueden encontrar los modelos de lenguaje neuronales como una solución al problema mencionado anteriormente. Pues con el modelo expuesto en (Vincent, 2003) se introdujo la idea de que las redes neuronales son capaces de aprender representaciones similares de palabras dentro de un contexto, otorgando así respuestas a n-gramas que no se hayan observado en el *corpus* de entrenamiento usado.

## Grafos

Los n-gramas y las conexiones entre ellos pueden ser almacenados y visualizados como un grafo. Un grafo es una estructura constituida por nodos y arcos entre esos nodos. El ámbito de estudio de los grafos es conocido como teoría de Grafos, la cual es una rama de la matemáticas y ciencias de la computación.

En este caso las palabras serían representadas por los nodos y los arcos establecerían relaciones de sucesión entre ellas, es decir, imaginemos el nodo que simboliza a la palabra “por” dicho nodo podría presentar arcos que lo relacionaría con las palabras “lo” y “ejemplo” tras haber analizado las frases: “… por lo tanto esta comida está fría” y “por ejemplo, aquí tienes mi estuche”, pues como se puede ver la palabra “lo” y “ejemplo” suceden a la palabra “por”. La *Figura 2.2* ilustra el ejemplo descrito, además nótese como los arcos reflejan conexiones unidireccionales, si estudiáramos una sentencia donde “lo” precediera a “por” entonces el arco entre ellas pasaría a reflejar una conexión bidireccional.

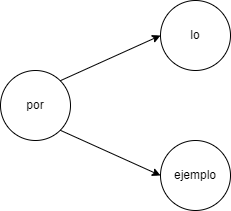


Figura 2.2: Ejemplo de grafo de unigramas a nivel de palabras

La representación de las conexiones entre n-gramas puede ser realizada por varias estructuras de datos diferentes, matrices de vectores, vectores de instancias de objetos, etcétera. La razón por la cual se ha escogido el grafo como estructura para representar dichas conexiones es dado a la facilidad que presenta para recorrerlo, sobre todo teniendo en cuenta que la base de datos elegida para el proyecto es una base de datos orientada a grafos.

# GPG – Grafo Preentrenado Generativo

En este capítulo se expone la arquitectura, flujos de datos, implementación del sistema a desarrollar. También se explican en detalle las funcionalidades que ofrece cada modelo implementado en el sistema.

Tras tener claro el funcionamiento y detalle del sistema el nombre que se le ha decido dar es GPG, abreviación de grafo preentrenado generativo.



## Conjunto de datos

### Fuente de los datos

Actualmente existen muy diversas fuentes de datos que pueden ser utilizadas para entrenar este tipo de modelos. En este caso por facilidad de lectura de formato se ha decidido elegir libros electrónicos escritos en la extensión de archivo TXT, es decir, los libros se encuentran en texto plano, además, presentan codificación de caracteres ANSI.

Los libros, todos en castellano, se han obtenido de la plataforma web del Proyecto Gutenberg (The Project Gutenberg Literary Archive Foundation, s.f.), disponen de aproximadamente 60000 libros en diferentes idiomas de los cuales la mayoría pertenecen a dominio público, por lo tanto, las licencias de uso no suponen un problema.

Se ha divido el conjunto de datos, en un conjunto de entrenamiento que supone aproximadamente un 80% del *corpus* total y un conjunto de prueba o test que ocupa el 20% restante.

A continuación, se enumeran los libros utilizados en el **conjunto de entramiento**:

1. Al primer vuelo (Pereda, Al primer vuelo, 1986)
2. Algo de todo (Valera, 1883)
3. Amar es vencer (Caro, 2008)
4. Amistad funesta: Novela (Martí, 1885)
5. Amparo (Memorias de un loco) (Fernández y González, 1858)
6. Ariel (Rodó, 1900)
7. Arroz y tartana (Blasco Ibánez, 1894)
8. Cádiz (Pérez Galdós, 1874)
9. Cartas de mi molino (Daudet & Cabañas, 1869)
10. Don Quijote de la Mancha (de Cervantes Saavedra, 1605)
11. Novelas Ejemplares (Cervantes Saavedra, 1613)
12. Viajes por España (de Alarcón y Ariza, Viajes por españa, 1883)
13. El niño de la bola: Novela (de Alarcón y Ariza, El niño de la bola, 1880)
14. La Regenta (Leopoldo Alas, La Regenta, 1884)

Lista de libros utilizados para el **conjunto de prueba:**

1. La vida de Lazarillo de Tormes y de sus fortunas y adversidades (Anónimo, 1554)
2. Su único hijo (Leopoldo Alas, Su único hijo, 1890)
3. Crónicas de Marianela (Palma Román, 1917)
4. La quinta de Palmyra (Gómez de la Serna, 1944)
5. La Montálvez (Pereda, La Montálvez, 1888)

Para saber la importancia del corpus utilizado a continuación se ofrecen una serie de tablas sobre el conjunto de entrenamiento y de test que ofrecen el número de palabras y caracteres por libro usado en cada uno.

|  |  |  |
| --- | --- | --- |
| Libro | N.º caracteres (sin espacios) | N.º palabras |
| 1 | 463807 | 102184 |
| 2 | 266908 | 56741 |
| 3 | 254249 | 54965 |
| 4 | 279164 | 61413 |
| 5 | 131144 | 28590 |
| 6 | 130052 | 25708 |
| 7 | 445047 | 91430 |
| 8 | 366500 | 78286 |
| 9 | 169742 | 34966 |
| 10 | 1718347 | 381104 |
| 11 | 838985 | 189412 |
| 12 | 337952 | 68386 |
| 13 | 359302 | 75073 |
| 14 | 1496535 | 310692 |
| Total | 7257734 | 1558827 |

Tabla 3.1: Número de palabras y caracteres del conjunto de entrenamiento

|  |  |  |
| --- | --- | --- |
| Libro | N.º caracteres (sin espacios) | N.º palabras |
| 1 | 85920 | 20063 |
| 2 | 425275 | 89836 |
| 3 | 251517 | 51827 |
| 4 | 212632 | 45551 |
| 5 | 494428 | 106976 |
| Total | 1469772 | 314253 |

Tabla 3.2: Número de palabras y caracteres del conjunto de prueba

### Limpieza de los datos

Se han tenido que tomar varias consideraciones a la hora de tratar los datos para proporcionar la información más relevante y menos conflictiva al modelo de lenguaje. A continuación, se listan todas las decisiones tomadas:

* En todos los libros mencionados se puede encontrar en las primeras líneas y últimas líneas unos párrafos en inglés conteniendo metadatos del libro en sí y además indicando los términos y condiciones de la licencia de uso del libro. Se decidide eliminar, directamente del archivo, dichas líneas pues contienen vocabulario en inglés, además de presentarse los metadatos en un formato no natural y poco parecido al lenguaje humano.
* Se ha optado por transformar todas las palabras presentes en los libros a minúscula, es decir, ‘Hola’ y ‘hola’ serían consideradas la misma palabra.
* Los signos de exclamación en interrogación son ignorados y no se tienen en cuenta a la hora de formar una palabra e identificar las palabras que le suceden, es decir, en la oración ‘¿Qué tal estás?’ se detectarán correctamente las palabras [qué], [tal], [estás] y sus correspondientes conexiones.
* Se respetan acentuaciones de las palabras, por lo tanto, ‘que’ y ‘qué’ son palabras distintas. Por esta misma razón se puede diferenciar entre ‘qué’ interrogativo o exclamativo y ‘que’ como conjunción o pronombre relativo.
* En cuanto a signos ortográficos menos comunes, tales como ‘<<’, ‘>>’, ‘-‘, ’\_’, etcétera, son eliminados y no interfieren en la sucesión o detección de palabras.
* El final de una frase se indica por un punto ‘.’, por lo cual el texto de un libro es divido en sentencias. De esta forma no se contabilizan relaciones de continuidad entre la última palabra de una frase y la primera de la siguiente.
* En varios de los libros se ha observado el uso de la preposición “a”, “o”, “u” pero acentuándose, es decir: “á”, “ó”, “ú”. Esto se debe a un uso anticuado, se ha corregido en todo texto usado.

Se pueden consultar los detalles de implementación a código que satisfacen estos requisitos en la Sección \*\*

## Arquitectura del sistema

Es importante antes de introducirse a implementar el sistema tener asentados conceptos de diseño previos. Saber identificar el ciclo de ejecución deseado para nuestro sistema y como el usuario podrá interactuar con el mismo, además de visualizar el proceso de extracción, transformación y almacenamiento de los datos.

Como se expone en siguientes apartados se pueden identificar pequeños procedimientos que componen el ciclo de ejecución total del programa, procedimientos que generalmente se repiten y que pueden ser encapsulados en trozos de código conocidos como funciones. Teniendo en cuenta las razones anteriores para implementar se decide alejarse del paradigma orientado a objetos y centrarse en un paradigma de programación funcional.

### Flujo de control

El ciclo de vida del sistema es sencillo, este puede ser visto de una forma jerárquica empezando por una vista global como la expuesta en Figura 3.1.

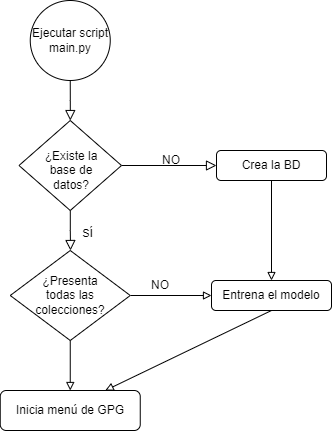


Figura 3.1: Flujo de control global de GPG

Como se puede observar en el mayor nivel de abstracción el flujo de control de GPG es bastante simple. A continuación, se exponen otros diagramas donde se desgranan los procesos complejos que en esta primera vista han sido simplificados.

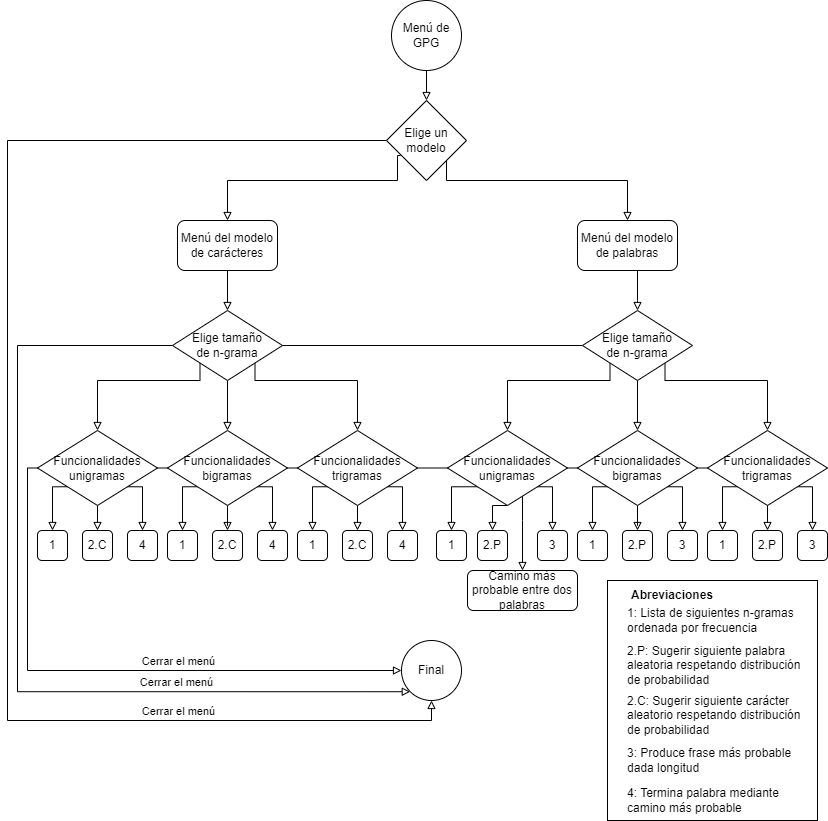


Figura 3.2: Flujo de control del menú de GPG

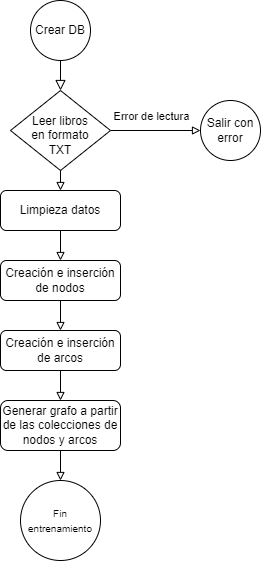


Figura 3.3: Flujo de control del entrenamiento del modelo

### Flujo de datos

Se necesita conocer el camino, transformación y creación de estructuras de datos que van a necesitar los procesos que constituyen n

\*\* hablar sobre estructura de datos intermedias, json, atributos de arcos.

\*\*Especificar el proceso de análisis por el cual el conjunto de datos es analizado y almacenado en la base de datos ArangoDB.

\*\* Explicar como se calcula el valor y atributo de los arcos del grado. Escribir formula.

\*\* Especificar una tabla con los n-gramas más comunes obtenidos tras analizar los libros.

## Implementación del sistema

\*\* organización en directorios.

* Mediante el uso del módulo *re.py* de expresiones regulares presente en Python se ha especificado el siguiente patrón para identificar palabras en una secuencia: **'\b\w+\b'** esta expresión elimina todos los signos ortográficos mencionados en sección, a excepción de ‘\_’ que ha remplazado en los textos por un carácter vacío ‘ ’ **\*\***
* para ello en el mismo código se hace uso de la función *lower()* la cual convierte toda letra mayúscula presente en una cadena a caracteres en minúscula. \*\*

### Grafos creados

\*\* ofrecer numero de nodos y arcos por cada grafo.

### Funcionalidades

\*\* Métodos de predicción realizados, ofrecer lista de pasos de como se procede en cada método.

## Evaluación

Para saber la utilidad o como de bien puede desempeñar nuestro sistema necesitamos encontrar algún tipo de métrica que nos sirva. El problema al que nos enfrentamos en este caso es que el sistema desarrollado no se puede considerar una aplicación de cara al usuario final, sino un modelo del lenguaje simple que podría ser integrado en un servicio más complejo.

Teniendo esto en cuenta podemos diferenciar entre **evaluación extrínseca** y una **evaluación intrínseca.** Las primeras son evaluaciones que toman muchas medidas de una aplicación, donde se valoran el rendimiento (tiempo, memoria), tiempos de respuesta, número de peticiones a servicios, etc… Su intención es que sean realizadas sobre aplicaciones finales. Como se puede ver la evaluación extrínseca no parece una opción en este caso, por lo tanto se realizará una evaluación intrínseca.

La evaluación intrínseca no nos ofrecerá datos del rendimiento de nuestra aplicación, pero nos aportará información sobre nuestro modelo del lenguaje. Una métrica comúnmente utilizada es *perplexity.*

### Perplexity

La *perplexity,* a menudo abreviada como PP, es una medida que valora como de perplejo se encuentra nuestro modelo del lenguaje ante los datos, nunca vistos por el modelo, del conjunto de pruebas.

La *perplexity* de un modelo del lenguaje sobre un conjunto de prueba es la inversa de la probabilidad de aparición, que nuestro modelo asignaría a las sentencias presentes en el test, normalizada por el número de palabras. Para un conjunto de prueba

Como se puede observar, al ser la inversa, a menor *PP* mayor probabilidad de aparición otorga nuestro modelo a las sentencias presentes en el conjunto de prueba.

(Jurafsky & Martin, Evaluating Language Models, 2022)

Se recuerda que puede ser expandido utilizando la regla de la cadena, o simplificado aplicando la asunción de Márkov de orden dependiendo del modelo de -grama que estemos analizando. Esto se explicó en más detalle en la Sección 2.1.

Un problema que encontramos con la *perplexity* es que esta no puede ser medida si el conjunto de entramiento no presenta todo el vocabulario que puede estar presente en el conjunto de pruebas. Para solucionar este problema se ha optado por la aproximación de Laplace a la hora de calcular la probabilidad de aparición.

Mediante este enfoque toda aquella palabra no reconocida en el conjunto de prueba será considerada como el token <UNK>. Cada vez que esto pase toda palabra reconocida aumentará en 1 el número de veces que aparece.

#### Resultados

Todas estas consideraciones hacen que el proceso se observe como algo complejo, afortunadamente la biblioteca NLTK nos aporta herramientas para el cálculo automático de esta medida dados un conjunto de entramiento y un conjunto de prueba.

El código de evaluación del modelo se encuentra en *scripts/evaluate.py* y se ha realizado mediante la documentación de NLTK en (NLTK :: ntlk.lm package, 2022). Los resultados que la ejecución del *script* ha arrojado frente a nuestro conjunto de entramiento formado por 1.558.827 palabras y el conjunto de prueba formado por 314.253 palabras son los siguientes:

|  |  |  |  |
| --- | --- | --- | --- |
| Perplexity | Unigramas | Bigramas | Trigramas |
| Modelo de palabras | 7327,03 | 2838,71 | 3775,03 |
| Modelo de caracteres | 23,75 | 14,67 | 11,22 |

Tabla 3.3: Resultados perplexity sobre corpus de GPG

Para poner estos resultados en contraste también se ofrecen la *perplexity* calculada sobre un *corpus* de datos, formado por 38.000.000 palabras para el entramiento y 1.500.000 palabras para el test y usando un vocabulario cerrado de 19.979 palabras, obtenido del Wall Street Journal (WSJ) y que nos proporcionan (Jurafsky & Martin, Evaluating Language Models, 2022).

|  |  |  |  |
| --- | --- | --- | --- |
| Perplexity | Unigramas | Bigramas | Trigramas |
| Modelo de palabras | 962 | 170 | 109 |

Tabla 3.4: Resultados perplexity sobre corpus del Wall Street Journal

Viendo estos resultados podemos concluir:

* Los modelos de caracteres al presentar vocabularios mucho más pequeños que los modelos de palabras siempre obtendrán una mejor puntuación.
* A pesar de que no es una comparación justa, pues el modelo de WSJ cuenta con un vocabulario cerrado de 19.979 palabras frente a nuestro modelo que tras entrenamiento cuenta con 65.000 palabras aproximadamente, se puede intuir que nuestro modelo está poco entrenado.
* Los bigramas y trigramas presentan menores valores pues al tener contexto de los últimos *tokens* se hace más sencillo predecir el siguiente. Por lo tanto ver que el resultado del modelo de trigramas en palabras es peor que el de bigramas indica que no tenemos un conjunto de entrenamiento lo suficientemente grande.

Se ha de tener en cuenta que debido a los recursos con los que se cuenta al desarrollar el proyecto, no se ha podido aumentar el corpus de datos significativamente, pues el procesamiento de estos mismos se completaría en varios días.

# Demo técnica

Con el objetivo de mostrar la utilidad de GPG, se ha decidido desarrollar una pequeña demo técnica en la que se satisfarían los objetivos por los que inicialmente se planteó este proyecto, ayudar a un usuario conforme escribe un texto mediante la sugerencia de palabras o autocompletando las que este escribe.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 4.1: Demo técnica de GPG

El desarrollo de esta demo implica una simple GUI, desarrollada con la biblioteca Tkinter (tkinter - Interface de Python para Tcl/Tk, 2022), en la que el usuario es capaz de escribir texto y se le ofrecen **dos tipos de sugerencias** distintas.

* Siguiente palabra: Se analiza el texto escrito por el usuario, y se ofrecen en una caja de opciones diferentes sugerencias al input introducido. Se ofrece como máximo una sugerencia por cada modelo de N-Palabra, es decir 3.
* Autocompletar palabra: Dado un texto introducido por el usuario si la última entrada es una palabra a medio escribir, se ofrecen en una caja de opciones diferentes sugerencias. Se ofrece como máximo una sugerencia por cada modelo de N-Carácter, es decir 3.

Téngase en cuenta que para que modelos de bigramas o trigramas requieren de un mínimo de *tokens* escritos, de ahí que estos no ofrecerán sugerencias si el *input* no es suficientemente largo.

Se exponen los diagramas de secuencia de cada tipo de sugerencia. En ellos se da por hecho que se realiza una secuencia de ejecución correcta, es decir, no se reflejan casos como que la entrada no sea suficientemente larga o que el *token* introducido no se encuentre en la base de datos. Si ocurriera alguna de estas situaciones la sugerencia afectada será nula y no se mostrará en pantalla.

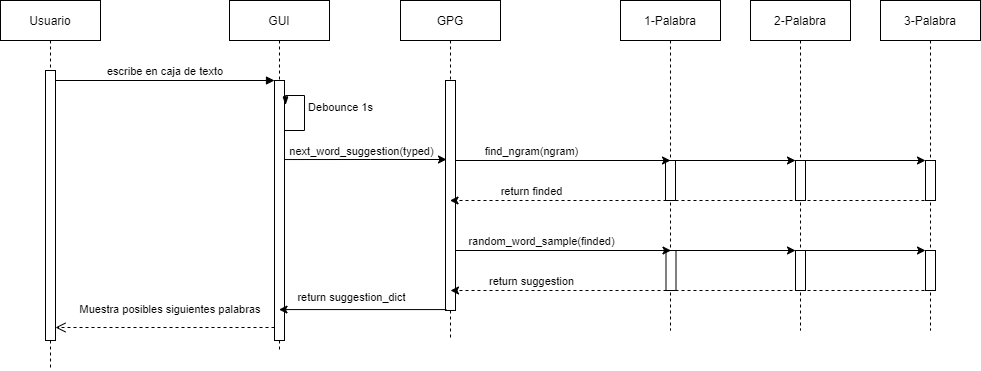


Figura 4.2: Diagrama de secuencia sugerencias de siguientes palabras

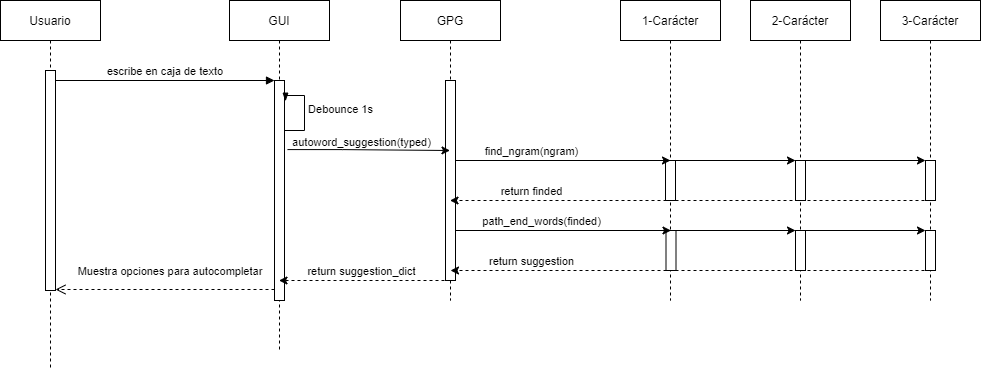


Figura 4.3: Diagrama de secuencia sugerencias para autocompletar la palabra

Tanto en la Figura 4.2 como en la Figura 4.3 se puede apreciar una función llamada debounce, esta se encarga de realizar únicamente la última llamada de las funciones realizadas a GPG tras haber pasado 1 segundo desde que el usuario soltó una tecla. Esto se hace para que no se haga una petición a ArangoDB cada vez que el usuario tecle un nuevo carácter, mejorando así el rendimiento considerablemente.

# Conclusiones



## Conclusiones sobre objetivos

En este punto se revisarán los objetivos mencionados en la Sección 1.3 y como se les ha dado solución en el desarrollo del sistema, empezando por los objetivos primarios y continuando por los secundarios.

Objetivos **primarios:**

Objetivos **secundarios:**

1. Se ha configurado correctamente el despliegue y uso de un **contenedor** mediante el fichero *Dockerfile* presente en [repositorio público del proyecto](https://github.com/alberto-lopov/bachelor-thesis).
2. Se ha realizado una **evaluación** de nuestro modelo como se detalla en la Sección 3.4.
3. Se ha escrito un **fichero *readme.md*** con las instrucciones necesarias para hacer uso del proyecto, este se puede encontrar en el [repositorio público del proyecto](https://github.com/alberto-lopov/bachelor-thesis).
4. Se ha montado una **demo** mediante una GUI realizada con tkinter como se expone en la Sección 4.

## Trabajo Futuro

A continuación, se listan una serie de aspectos que de aplicarse podrían mejorar el producto final con el que se ha concluido este proyecto.

1. Integración del sistema en herramientas de escritura normalmente utilizadas en forma de *plugin*.
2. Desarrollo y despliegue del sistema en una API, sobre todo si se quiere ofrecer un servicio web que utilice el modelo.
3. Solventar el problema de *data sparsity* que se originaría en el modelo de palabras al hacer uso de bigramas o principalmente en los trigramas. Esta solución pasaría parcialmente por conseguir un corpus de datos mucho mayor al actual. Una manera de hacer esto se expone en (Suzuki, Itoh, Nagano, Kurata, & Thomas, 2019) aumentando el conjunto de datos con los ejemplos generados por una red neuronal, o bien con aplicación de técnicas de *smoothing.*

## Valoración personal

El trabajo realizado aún consiguiendo resultado humildes y sencillos no ha sido fácil de desarrollar. El proyecto se ha desarrollado desde cero y ha atravesado numerosas complicaciones, principalmente derivadas por el final de la relación con el encargado de la empresa 8-Belts.

El hecho de haber desarrollado GPG hasta su estado actual me produce satisfacción y orgullo, pues considero que el esfuerzo depositado en él ha sido considerable. Realizar este proyecto me ha aportado nuevos conocimientos en tecnologías y ámbitos desconocidos para mí, como las bases de datos orientadas a grafos o principios del procesamiento del lenguaje natural.

Me alegro de haber tenido como mi tutor a Juan Francisco Huete Guadix por su orientación a la hora de como poder continuar el proyecto, otorgándome así la voluntad para poder terminar el mismo.

# Tecnologías

En este capítulo se introducirán las distintas tecnologías utilizadas en el desarrollo del proyecto.

La elección del lenguaje es una decisión fundamental. En este caso los lenguajes más populares cuentan con las librerías más completas y diversas. Enfocándonos en el ámbito de análisis de datos C++, Python o Java presentan muy buenas opciones. En este caso se ha decidido optar por Python debido a su facilidad de aprendizaje y de uso.

El uso de la base de datos ArangoDB se debe a que nos aporta la posibilidad de producir grafos de los datos almacenados en la misma, quitando dicha carga de programación en la parte del lenguaje. Relacionado con el uso de esta se ha usado Docker para poder aislar dicha tecnología en un contenedor y permitir su despliegue en distinto equipos de trabajo.



## Python

Python es un lenguaje de programación interpretado creado usando C y C++ por Guido van Rossum. Python es un lenguaje multiparadigma, es decir, podemos programar desde diferentes paradigmas de programación, como OOP o programación funcional.

Se caracteriza por ser uno de los lenguajes más sencillos de utilizar. El código escrito en Python es legible y se asemeja al lenguaje humano. También presenta reglas de sintaxis más laxas que otros lenguajes populares como C++.

Python también es uno de los lenguajes más utilizados por programadores en 2022, como nos indica la entrevista anual realizada por Stack Overflow.

(Stack Overflow, 2022)

La versión de Python utilizada es 3.10.2. Para realizar una sencilla, instalación y gestión de paquetes se ha usado el famoso gestor de paquetes Pip en su versión 21.2.4.

Las librerías instaladas utilizadas para el desarrollo del sistema han sido:

* **Python-Arango**: Implementa un *ORM* proporcionándonos una interfaz para interactuar con la base de datos desde el lenguaje. Versión 7.3.1
* **Python-Dotenv**: Habilita el uso de variables de entorno. Usado para poder desarrollar en diferentes equipos de trabajo. En este caso contiene las direcciones absolutas de los libros a leer, ya que estás rutas cambian según el dispositivo en el que trabajemos. Versión 0.20.0
* **SciPy**: Implementa una gran cantidad de algoritmos y modelos matemáticos. En concreto estamos utilizando esta librería por el módulo de estadística de la misma para poder generar una distribución discreta. Versión 1.9.1
* **Nltk:** Es la librería más conocida de Python relativa al procesamiento del lenguaje natura. En concreto se usa esta biblioteca para calcular la métrica *perplexity.*Versión 3.7
* **Tkinter:** Permite el desarrollo de sencillas interfaces gráficas de usuario mediante el uso de la biblioteca gráfica Tcl/Tk. Versión 8.6

## Docker

Docker es un software de código abierto, en cuyo desarrollo encontramos que han participado grandes titanes de la industria, hablamos de Microsoft, Cisco Systems, Google, IBM, o Red Hat.

Docker intenta resolver el famoso problema en el de desarrollo de aplicaciones software resumido en la frase “en mi ordenador si funciona”. Esta tecnología nos permite aislar conjuntos de dependencias en los conocidos como “contenedores Docker”. En estos contenedores podemos desarrollar y ejecutar nuestros proyectos, evitando posibles conflictos con otras librerías o programas que pudiéramos haber instalado para previos desarrollos.

Pero sobre todo una de las características clave que ha hecho que sea tan popular es que permite hacer esto de manera automática, es decir, una vez realizada la configuración de dicho contendor el mismo puede ser utilizado en cualquier otro dispositivo que presente el mismo sistema operativo y el proyecto Docker instalado.

En este proyecto se ha utilizado Docker para poder realizar un despliegue automático de la base de datos Arango en diferentes equipos.

## ArangoDB

La base de datos Arango es una base de datos multimodelo lo cual permite su utilización como una base de datos orientada a grafos, una base de datos orientada a documentos y una base de datos orientada a parejas clave-valor.

Aparte Arango cuenta con su propio lenguaje de *queries* (consultas), común a los tres modelos mencionados anteriormente. Dicho lenguaje aspira por parecerse a los lenguajes de alto nivel multipropósito, utilizando palabras reservadas similares como, por ejemplo, *for* para iterar entre los documentos de una colección en la base de datos. Gracias a esto la familiarización con sus consultas resulta más fácil.

Debido a dichas características ArangoDB es una base de datos realmente flexible, que permite una gran eficiencia y escalabilidad horizontal en las aplicaciones donde sea utilizado.

En este caso su uso ha resultado muy conveniente en el desarrollo del modelo, la base de datos nos proporciona las herramientas para el almacenamiento y tratamiento de los datos en un grafo, incorporando herramientas recorrer al mismo.

## Git

Git es un sistema de control de versiones ampliamente conocido por los desarrolladores. Este es de código abierto y libre.

Normalmente un sistema de control de versiones suele ser utilizado junto a un servicio de almacenamiento en la nube de esas mismas versiones. Entre los más conocidos encontramos servicios como Github, Gitlab o BitBucket. Para facilitar el desarrollo entre diferentes equipos se ha creado un repositorio público en la plataforma Github al que se puede acceder buscando “alberto-lopov/bachelor-thesis” o mediante el siguiente hiperenlace: <https://github.com/alberto-lopov/bachelor-thesis>

# Bibliografía

Anónimo. (1554). *La vida de Lazarillo de Tormes y de sus fortunas y adversidades.* Obtenido de https://www.gutenberg.org/ebooks/320

Blasco Ibánez, V. (1894). *Arroz y tartana.* Obtenido de https://www.gutenberg.org/ebooks/16413

Caro, M. P. (2008). *Amar es vencer.* Obtenido de https://www.gutenberg.org/ebooks/24925

Cervantes Saavedra, M. D. (1613). *Novelas Ejemplares.* Obtenido de https://www.gutenberg.org/ebooks/61202

Daudet, A., & Cabañas, F. (1869). *Cartas de mi molino.* Obtenido de https://www.gutenberg.org/ebooks/29706

de Alarcón y Ariza, P. A. (1880). *El niño de la bola.* Obtenido de https://www.gutenberg.org/ebooks/59154

de Alarcón y Ariza, P. A. (1883). *Viajes por españa.* Obtenido de https://www.gutenberg.org/ebooks/26314

de Cervantes Saavedra, M. (1605). *Don Quijote de la Mancha.* Obtenido de https://www.gutenberg.org/ebooks/2000

Fernández y González, M. (1858). *Amparo (Memorias de un loco).* Obtenido de https://www.gutenberg.org/ebooks/27295

Free Software Foundation. (2022). *What is Free Software? - GNU Project.* Obtenido de https://www.gnu.org/philosophy/free-sw.en.html

Gómez de la Serna, R. (1944). *La quinta de Palmyra.* Obtenido de https://www.gutenberg.org/ebooks/63228

Gudivada, V. N., Rao Vijay, D., & Raghavan, V. (2015). Handbook of Statistics. En *Chapter 9 - Big Data Driven Natural Language Processing Research and Applications* (págs. 203-238). Elsevier.

Jurafsky, D., & Martin, J. H. (2022). Bibliographical and Historical Notes. En *Speech and Language Processing* (pág. 552). Obtenido de https://web.stanford.edu/~jurafsky/slp3/ed3book\_jan122022.pdf

Jurafsky, D., & Martin, J. H. (2022). Evaluating Language Models. En *Speech and Language Processing* (págs. 36-39). Obtenido de https://web.stanford.edu/~jurafsky/slp3/ed3book\_jan122022.pdf

Leopoldo Alas, C. (1884). *La Regenta.* Obtenido de https://www.gutenberg.org/ebooks/17073

Leopoldo Alas, C. (1890). *Su único hijo.* Obtenido de https://www.gutenberg.org/ebooks/17341

Martí, J. (1885). *Amistad funesta: Novela.* Obtenido de https://www.gutenberg.org/ebooks/18166

*NLTK :: ntlk.lm package*. (2022). Obtenido de NLTK :: Natural Language Toolkit: https://www.nltk.org/api/nltk.lm.html

Palma Román, M. A. (1917). *Crónicas de Marianela.* (P. Balza, Ed.) Obtenido de https://www.gutenberg.org/ebooks/34565

Pereda, J. M. (1888). *La Montálvez.* Obtenido de https://www.gutenberg.org/ebooks/25812

Pereda, J. M. (1986). *Al primer vuelo.* Obtenido de https://www.gutenberg.org/ebooks/23957

Pérez Galdós, B. (1874). *Cádiz.* Obtenido de https://www.gutenberg.org/ebooks/21906

Rodó, J. E. (1900). *Ariel.* Obtenido de https://www.gutenberg.org/ebooks/22899

Russell, S., & Norvig, P. (2009). N-gram character models. En *Artificial Intelligence: A Modern Approach, 3rd edition* (págs. 861-863). Pearson.

Russell, S., & Norvig, P. (2009). N-gram word models. En *Artificial Intelligence: A Modern Approach, 3rd edition* (págs. 864-865).

Stack Overflow. (2022). *Stack Overflow Developer Survey.* Obtenido de https://survey.stackoverflow.co/2022/

Suzuki, M., Itoh, N., Nagano, T., Kurata, G., & Thomas, S. (2019). Improvements to N-gram Language Model Using Text Generated from Neural Language Model. En *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (págs. 7245-7249).

The Project Gutenberg Literary Archive Foundation. (s.f.). *About | Project Gutenberg*. Obtenido de Project Gutenberg Web Site: https://www.gutenberg.org/

*tkinter - Interface de Python para Tcl/Tk*. (2022). Obtenido de Documentación de Python: https://docs.python.org/es/3/library/tkinter.html

Valera, J. (1883). *Algo de todo.* Obtenido de https://www.gutenberg.org/ebooks/30213

Vincent, J. (2003). Journal of Machine Learning Research. En D. R. Bengio Y, *A neural probabilistic language model* (págs. 1137-1155).