

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, 8 de julio de 2022

**Modelos de Generación de Texto**

Alberto López Povedano

**Palabras clave:** orientado a grafos, lenguaje natural, generación de texto, grafos, inteligencia artificial

**Resumen:**

Implementación de un sistema recomendador de texto en función las palabras escritas por el usuario.

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 Julio 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 6 de julio de

2020.

**Los directores:**

**Juan Francisco Huete Guadix**

**Agradecimientos**

A Juan Francisco Huete Guadix por la supervisión en la realización del trabajo de final de grado.

A Ernesto Martínez del Prieto por su inestimable ayuda ofreciéndome recursos didácticos, asesoramiento y recomendaciones en la realización del trabajo de final de grado.

Índice

[1. Introducción 12](#_Toc114743625)

[1.1. Propósito 12](#_Toc114743626)

[1.2. Justificación 12](#_Toc114743627)

[1.3. Objetivos 12](#_Toc114743628)

[1.4. Metodología 12](#_Toc114743629)

[1.5. Estado del arte 12](#_Toc114743630)

[1.5.1. GPT-3 12](#_Toc114743631)

[2. Fundamentos 12](#_Toc114743632)

[2.1. Grafos 12](#_Toc114743634)

[2.2. Distribución discreta 12](#_Toc114743635)

[2.3. Teoría de modelos de generación de texto 12](#_Toc114743636)

[3. GPG – Grafo Preentrenado Generativo 12](#_Toc114743637)

[4. Conclusiones 12](#_Toc114743638)

[5. Tecnologías 12](#_Toc114743639)

[5.1. Python 12](#_Toc114743643)

[5.2. Docker 13](#_Toc114743644)

[5.3. ArangoDB 13](#_Toc114743645)

[5.4. Git 14](#_Toc114743646)

[6. Bibliografía 15](#_Toc114743647)

# Introducción

## Propósito

## Justificación

## Objetivos

## Metodología

## Estado del arte

## GPT-3

# Fundamentos



## Grafos

## Distribución discreta

## Teoría de modelos de generación de texto

# GPG – Grafo Preentrenado Generativo

# Conclusiones

# 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 modelos de grafos de los datos almacenados en la misma, quitando dicha carga de programación en la parte del lenguaje. Relacionado con el uso de la misma 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 programación orientada a objetos (POO) 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 *Object-relational mapping (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

## Docker

Docker es un software de código abierto, en cuyo desarrollo encontramos que han participado grandes titanes de la industria, hablamos de Micrososft, 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 con un sistema operativo Windows 10.

## 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 lo 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 enlace: <https://github.com/alberto-lopov/bachelor-thesis>

# Bibliografía

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