Tesis

Automatización de lectura de Currículum Vitae para selección de personal en el Sector IT

Calonge, Federico Matias
calongefederico@gmail.com

9 de enero de 2022

Resumen

En la Tesis de Ingeniería que se presenta, se diseña un *sistema de lectura automática de Curriculum Vitae*. La finalidad del mismo es ayudar al reclutador laboral a elegir a los mejores candidatos para los puestos laborales que tenga disponible mediante una medición de similitud entre textos: Curriculum Vitae de los candidatos por un lado, y descripciones de puestos laborales por el otro. El sistema esta desarrollado utilizando el lenguaje de programación Python, permitiendo verificar la teoría desarrollada.

Abstract

This Engineering Thesis introduces an *automatic Curriculum Vitae reading system*. The purpose of it is to help the job recruiter to choose the best candidates for the available job positions by means of a measurement of similarity between texts: Curriculum Vitae of the candidates on the one hand, and job descriptions on the other. The system is developed using the Python programming language allowing to verify the developed theory.

Índice

1 Introducción.

El proceso de **selección de personal** se ha vuelto crucial para el manejo de recursos humanos en el mundo laboral moderno. Con la transformación digital de las empresas y del mercado laboral en general, identificar los perfiles más acordes a las necesidades de la empresa se convirtió en uno de los retos más ambiciosos de Recursos Humanos, en especial cuando hablamos del **Sector IT**.

En estos últimos años se implementaron una gran cantidad de **herramientas de Software que permiten automatizar y gestionar información de los candidatos de una manera mucho más intuitiva e inteligente**. Gracias a la ayuda de este tipo de sistemas, el reclutador consigue a los candidatos más cualificados para cada puesto.

El tema de este Proyecto de Tesis será desarrollar un **Sistema de lectura automática de Curriculum Vitae** que ayude al reclutador laboral a elegir a los mejores candidatos para los puestos laborales que tenga disponible mediante una **medición de similitud** entre textos: los Curriculum Vitae de los candidatos por un lado, y las descripciones de puestos laborales por el otro.

La medición de similitudes de documentos es uno de los problemas más cruciales del **Procesamiento del Lenguaje Natural (NLP)**. Encontrar similitudes entre documentos se utiliza en varios dominios, tales como recomendación de películas, libros o artículos similares, identificación de documentos plagiados o documentos legales, etc.

Para que las máquinas puedan descubrir esta similitud entre documentos, se necesita definir una forma de medir matemáticamente la similitud, la cual debe ser comparable para que la máquina pueda identificar qué documentos son más similares (o menos). Previamente a esto necesitamos representar el texto de los documentos en una forma cuantificable (que suele ser en forma vectorial), de modo que podamos realizar cálculos de similitud sobre él.

Por lo tanto, los pasos necesarios para que las máquinas puedan medir la similitud entre documentos son:

1. Convertir un documento en un objeto matemático (vector).
2. Definir y emplear una medida de similitud.

Para el primer paso se utilizarán los algoritmos de vectorización **TF-IDF** y **Word Embeddings**; y para el segundo paso se emplearán las técnicas **Cosine Similarity** y **Word Mover's Distance -WMD-**.

1.1 Objetivos del Proyecto.

1.1.1 Objetivo general.

El objetivo de este Proyecto de Tesis es lograr la implementación de un Sistema de lectura automática de Curriculum Vitae accesible vía Web, que servirá para la elección de los mejores candidatos para cada puesto laboral; basándose en la **similitud** entre el Currículum Vitae del candidato y el puesto laboral.

1.1.2 Objetivos específicos.

Los objetivos específicos de este Proyecto de Tesis son:

- Describir el estado del arte actual de los Sistemas de lectura y análisis de Curriculum Vitae en Recruiting.
- Implementar un Sistema de lectura automática de Curriculum Vitae basado en la comparación entre textos, para finalmente obtener una visualización de los mejores candidatos para un puesto laboral determinado.
- Aprender los conceptos y técnicas principales utilizadas dentro del procesamiento de lenguaje natural (NLP) aplicando técnicas de preprocesamiento y limpieza de textos.
- Implementar diferentes técnicas para medir similitudes entre los textos (Cosine Similarity y Word Mover's Distance -WMD-) y diferentes algoritmos de vectorización (TF-IDF y Word Embeddings), analizando su funcionamiento tanto teórica como matemáticamente, ventajas y desventajas.
- Evaluar los resultados de cada técnica y algoritmo e implementar la mejor solución en el Sistema.
- Evaluar los Frameworks disponibles para tener una UI accesible vía web e integrar el mismo al Sistema.
- Almacenar datos de Candidatos, Reclutadores y puestos laborales en una base de datos; contando con un sistema de login y registración para los mismos.

1.2 Alcance del Proyecto.

El alcance de esta Tesis de Grado de Ingeniería incluye el desarrollo de conceptos de análisis de datos, procesamiento de lenguaje natural, técnicas de preprocesamiento y limpieza de los datos, algoritmos de vectorización y técnicas para medir la similitud entre textos, integración con frameworks, visualización de datos, y gestión de Base de Datos, de acuerdo a lo enunciado en los objetivos específicos.

1.3 Organización.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

2 Reclutamiento laboral en IT.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

2.1 Introducción.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

2.2 Problemáticas.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

2.3 Sistemas de lectura y análisis de CV : Estado del arte.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

3 Aprendizaje supervisado y no supervisado en Machine Learning.

3.1 Introducción.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Los métodos que utilizaremos en este Proyecto de Tesis serán...

3.2 K-Nearest Neighbor (KNN).

K-Nearest Neighbor (o K Vecinos más Próximos en español), blablabal....

3.3 K-Means.

K-Means (o K-Medias en español), blablaba...

3.3.1 Elbow Method.

Elbow Method (o Método del codo en español), blablaba...

4 Natural Language Processing.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.1 Introducción.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.2 Preprocesamiento de textos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.3 Similitud entre textos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.4 Técnicas para medir Similitud entre textos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.4.1 Cosine Similarity.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.4.2 Word Mover's Distance (WMD).

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.5 Algoritmos de vectorización.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.5.1 TF-IDF.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.5.2 Word Embeddings.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

4.5.2.1 ¿Cómo entrenar Word Embeddings?

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5 Desarrollo.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.1 Introducción.

El desarrollo de este Sistema se realizó en Python...

5.2 Esquema del sistema.

Como podemos ver en la figura 5.1, blablaba.

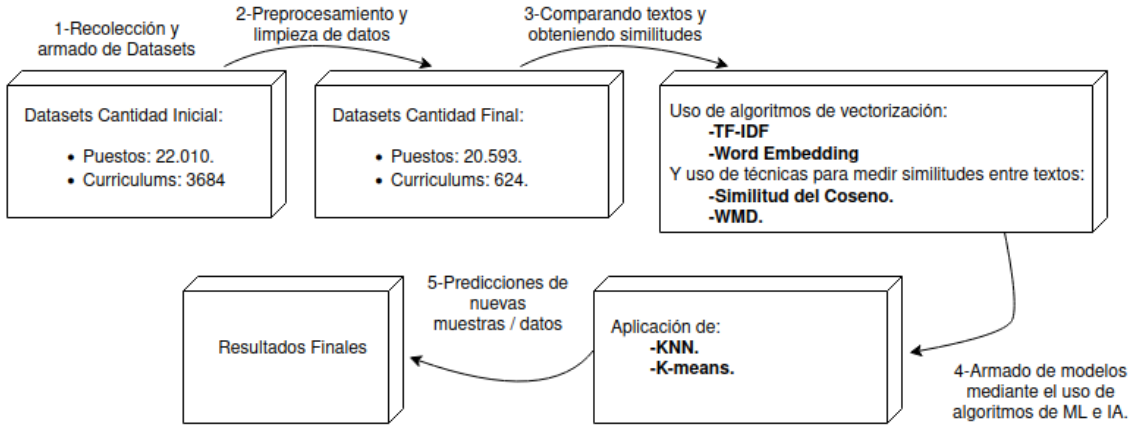

Figura 5.1: Flow del Core del Sistema

5.3 Set de datos.

Un Set o conjunto de datos es una tabla de una base de datos o, matemáticamente, una matriz estadística de datos. Cada columna de la tabla representa una variable del Data Set; y cada fila representa a un miembro determinado del conjunto de datos.

Previamente a utilizar algoritmos de ..... necesitamos que los datos estén limpios, por esta razón es importantísimo manipular y analizar estos datos en una primera instancia para luego si, meter estos datos limpios en los algoritmos y que estos funcionen correctamente; ya que de lo contrario las máquinas podrían clasificar o predecir de forma errónea. Este análisis previo sobre los datos debe ser minucioso ya que puede haber valores incoherentes o absurdos.

Para este Proyecto utilizamos dos grandes sets de datos, descriptos a continuación.

5.3.1 Curriculum Vitae.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.3.2 Descripciones Puestos Laborales.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.4 Medición de similitud entre textos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.4.1 FALTA: Pasos de las pruebas que hice: "Mejorando... / Ïmplementación de... / "Problemas al...".

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.5 Resultados Obtenidos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.6 Conclusiones.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7 Agregando funcionalidades.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.1 Base de datos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.2 Framework Web.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.3 Roles y Usuarios.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.4 Manejo de los datos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.4.1 Modelado.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.4.2 Filtrado.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.7.4.3 Visualización.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.8 Caso de Uso.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

5.9 Limitaciones del sistema.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

6 Próximos pasos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

7 Glosario.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

8 Anexo.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

9 Bibliografía.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.

10 Agradecimientos.

Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo. Texto ejemplo.