Tesis

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

Calonge, Federico Matias
calongefederico@gmail.com

11 de enero de 2022

Resumen

En la Tesis de Ingeniería en Informática 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 Computer 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 measuring 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-**.

Ver si agregar KNN y Kmeans

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.

Este Proyecto de Tesis consiste en tres grandes secciones:

1. Análisis e investigación inicial. Esta sección abarca principalmente la parte teórica del trabajo, haciendo hincapié en el análisis e investigación de:

   - El estado del arte (actual y pasado) de los sistemas de lectura y análisis de Curriculum Vitae.
   - Proyectos y librerias existentes.
   - Técnicas usadas para el procesamiento del lenguaje natural (NLP).
   - Técnicas para medir similitudes entre textos: .... y .... .
   - Algoritmos de Vectorización: .... y .... .

   Agregar KNN y Kmeans

2. Investigación y desarrollo del algoritmo core del sistema. Esta sección hace referencia a la aplicación práctica dentro del marco teórico desarrollado en la primera sección, la cual abarca:

   - Obtención de sets de datos: curriculums vitaes y descripciones laborales.
   - Preprocesamiento de los textos.
   - Implementación del algoritmo CORE: utilizando las técnicas para medir similitudes entre textos (... y ....) y algoritmos de vectorización (... y ...).
   - Pruebas entre distintas técnicas y librerías disponibles.
   - Análisis y primeras visualizaciones de los resultados.

   Agregar KNN y Kmeans

3. Investigación y desarrollo de la parte funcional del sistema. Esta última sección abarca lo que es la interfaz de usuario y su integración a un sistema web capaz de manejar las actividades entre los Candidatos y Reclutadores dentro del sistema. Los items principales son:

   - Definición de usuarios y roles.
   - Integración de frameworks y bases de datos.

- Modelado, filtrado y visualización de los datos.

VER si poner de primero hacer análisis en Jupyter notebooks -sección 2- y despues la parte web en django integrando estos jupyters -sección 3-

2 Reclutamiento laboral en IT.

En este capítulo se va a realizar un acercamiento a la historia del Reclutamiento laboral, especialmente en el sector de IT, sus etapas y problemáticas. Luego se realizará un análisis en el Estado de Arte actual de los Sistemas de lectura y análisis de CV actuales.

2.1 Introducción.

FALTA

2.2 Problemáticas.

FALTA

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

FALTA

3 Aprendizaje supervisado y no supervisado en Machine Learning.

Para comenzar el marco teórico de esta tesis, es necesario explicar qué es Machine Learning y cómo se pueden clasificar a los distintos algoritmos según su tipo de aprendizaje. En este capítulo se va a explicar el objetivo del aprendizaje supervisado y no supervisado en Machine Leanring haciendo énfasis en los algoritmos K-Nearest Neighbor (KNN) y K-means.

3.1 Introducción.

FALTA Los métodos que utilizaremos en este Proyecto de Tesis serán...

3.2 K-Nearest Neighbor (KNN).

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

3.3 K-Means.

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

3.3.1 Elbow Method.

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

4 Natural Language Processing.

FALTA

4.1 Introducción.

FALTA

4.2 Preprocesamiento de textos.

FALTA

4.3 Similitud entre textos.

FALTA

4.4 Técnicas para medir Similitud entre textos.

FALTA

4.4.1 Cosine Similarity.

FALTA

4.4.2 Word Mover's Distance (WMD).

FALTA

4.5 Algoritmos de vectorización.

FALTA

4.5.1 TF-IDF.

FALTA

4.5.2 Word Embeddings.

FALTA

4.5.2.1 ¿Cómo entrenar Word Embeddings?

FALTA

5 Desarrollo.

FALTA El desarrollo de este Sistema se trabajó en dos grandes partes: 1-Obtención del modelo predictivo. En esta primera parte se obtuvieron y preprocesaron datasets de Curriculum Vitae de distintos Candidatos y Descripciones de puestos de trabajo de IT publicados por distintas empresas, para luego ser comparados y obtener similitudes entre los textos utilizando las técnicas para medir distancias y obtener dichas similitudes (WMD y Cosine Similarity) y los algoritmos de aprendizaje (KNN y K-Means) anteriormente mencionados. De esta manera, el resultado final es la obtención de un modelo predictivo capaz de predecir nuevas muestras (en nuestro caso, predecir nuevos Candidatos para los distintos puestos laborales de IT disponibles).

2-Integración al Sistema Web: Esta fue una etapa posterior a la primera parte, donde se integró el modelo junto con la lógica de los algoritmos utilizados, a un Sistema Web. De esta manera, nuestro sistema cuenta con una interfaz gráfica permitiendo interactuar entre Candidatos y Reclutadores.

5.1 Obtención del modelo predictivo.

5.1.1 Introducción.

Como inicio definamos qué es un modelo. En nuestro caso, un modelo representa .... Este modelo se construyó en base a ... Que sea predictivo quiere decir que... El desarrollo de este Sistema se realizó en Python...

5.1.2 Esquema.

Como podemos ver la figura 5.1 representa el procedimiento utilizado para la obtención de nuestro modelo predictivo.

5.1.3 Obtención de sets de datos.

En primer lugar debemos definir qué es un set o conjunto 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 set de datos; y cada fila representa a un miembro determinado del mismo.

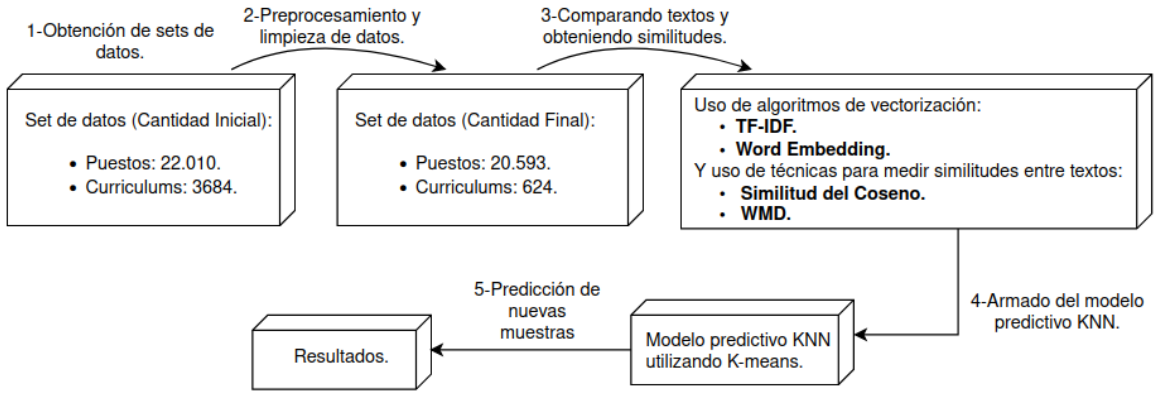

Figura 5.1: Flow del Core del Sistema

Para este Proyecto utilizamos dos grandes sets de datos que se obtuvieron mediante la recolección de distintos archivos alojados en la Web, los cuales estan descriptos a continuación.

5.1.3.1 Curriculum Vitae.

Los set de datos de Curriculum Vitae de Candidatos se obtuvieron de las siguientes fuentes:

1. 228 Curriculums en formato docx y posteriormente convertidos a pdf, obtenidos del sitio Kaggle (https://www.kaggle.com/palaksood97/resume-dataset). Estos pdfs son Candidatos de la India con experiencia en el rubro de IT.
2. 2484 Curriculums en formato CSV, obtenidos del sitio Kagle (https://www.kaggle.com/snehaanbha dataset). Este CSV cuenta con CVs obtenidos del sitio web de postulación de trabajos 'livecareer.com'.
3. 962 Curriculumns en formato CSV, obtenidos del sitio Kaggle (https://www.kaggle.com/gauravdut dataset). Este CSV cuenta con CVs repartidos en distintas categorías de IT.
4. 10 Curriculums en formato PDF, los cuales los cuales fueron obtenidos como ejemplos mediante una recolección propia de distintos sitios web.

5.1.3.2 Descripciones Puestos Laborales.

Los set de datos de descripciones de puestos laborales se obtuvieron de las siguientes fuentes:

1. 22.000 descripciones en formato CSV; obtenido del sitio Kaggle (https://www.kaggle.com/Prompt technology-jobs-on-dicecom). El CSV cuenta con descripciones de puestos obtenidos del sitio web de USA de postulación de trabajos del rubro de IT 'Dice.com'.
2. 10 descripciones en formato CSV; obtenidas como ejemplos mediante una recolección propia del sitio Indeed (https://www.indeed.com/q-USA-jobs.html) para puestos de trabajo de IT.

5.1.4 Preprocesamiento y limpieza de datos.

Previamente a utilizar las técnicas para medir distancias y obtener similitudes entre textos (WMD y Cosine Similarity) y los algoritmos de aprendizaje (KNN y K-Means) necesitamos que los datos que comparemos e introduzcamos en los algoritmos estén lo más limpios posible; ya que de lo contrario las mismos podrían clasificar o predecir de forma errónea. Este análisis previo sobre los datos debe ser minucioso ya que puede haber valores incoherentes o absurdos.

El procedimiento para la Limpieza de los Curriculum Vitae y las descripciones de los puestos laborales fue el siguiente:

1. Convertimos todo a minúscula.
2. Eliminamos datos no relevantes para nuestros análisis (mails y páginas web).
3. Eliminamos signos de puntuación y caracterés especiales (incluyendo números).
4. Eliminamos stop words.
5. Eliminamos common words no relevantes para nuestros análisis.
6. Aplicamos Lematización y Tokenización.
7. Eliminamos repetidos.
8. Obtenemos y usamos bi-gramas.

Luego de aplicar preprocesamiento y limpieza de datos nos quedarán los siguientes tamaños de nuestros datasets:

- 624 CVs de Candidatos (en formato pdf y csv).
- 20593 Descripciones de Puestos de IT (en formato csv).

5.1.5 Cantidad final del set de datos y su uso en las distintas etapas.

El total de CVs de Candidatos (624) y Descripciones de Puestos de IT (20593) que mencionamos previamente, serán utilizados para el entrenamiento y obtención de vectores mediante TF-IDF (para el futuro cálculo de Cosine Similarity) y para el entrenamiento de Word2Vec y obtención de los Word Embeddings (para el futuro cálculo de WMD).

Por otro lado, para calcular Cosine Similarity y WMD, para utilizarlos en K-means y para entrenar a nuestro algoritmo KNN, utilizaremos únicamente una porción de nuestros datasets:

1-Para el cálculo de Cosine Similarity y WMD: 301 CVs de Candidatos. 201 Descripciones de Puestos de IT. Nota: No obstante, al realizar los cálculos de distancias compararemos cada CV con cada Job Description, obteniendo un dataframe total de 3131 filas con sus respectivo valores de WMD y Cosine Sim.

2-Para el uso de K-means y entrenamiento con KNN (eliminamos un CV y una Descripción de Puesto IT que los utilizamos en '3-'): 300 CVs de Candidatos. 200 Descripciones de Puestos de IT. Nota: como se comentó previamente, nos quedarán 3000 filas / puntos para usar en K-means y entrenar KNN; llegando a representar estos 3000 puntos en un plano de 2 dimensiones.

3-Para la clasificación de nuevas muestras prediciendo con KNN: 1 CV de Candidatos. 1 Descripción de Puesto de IT. Nota: como se comentó previamente, nos quedarán 131 filas para clasificar.

¿Por qué utilizamos solo una porción de nuestros datasets?: Esto es debido a los drawbacks de WMD y KNN.

WMD: posee una alta complejidad en el cálculo de la distancia, teniendo un tiempo de ejecución muy elevado. Como ejemplo, al correrlo localmente, el cálculo de WMD para 3131 filas tardó 7 horas; frente a los 3 segundos que tardó el cálculo de Cosine Similarity para la misma cantidad de filas. KNN: KNN es una gran opción para datasets pequeños con pocas variables de entrada; pero tiene problemas cuando la cantidad de entradas es muy grande. Cada variable de entrada puede considerarse una dimensión de un espacio de entrada p-dimensional. En grandes dimensiones, los puntos que pueden ser similares pueden tener distancias muy grandes. Además, cada vez que se va a hacer una predicción con KNN, busca al vecino más cercano en el conjunto de entrenamiento completo. Por esto, se debe utilizar un dataset pequeño para que el clasificador KNN completa su ejecución rápidamente.

En conclusión, al utilizar solo una porción de nuestros datasets para obtener los distintos cálculos de distancias y entrenar KNN, el cálculo de WMD se podrá realizar en un tiempo finito, y nuestro clasificador KNN funcionará rápida y eficientemente al realizar predicciones.

5.2 Comparando textos y obteniendo similitudes.

FALTA

5.3 Armado del modelo predictivo KNN.

FALTA

5.4 Predicción de nuevas muestras y resultados obtenidos.

FALTA

5.5 Integración al Sistema Web.

FALTA

5.5.1 Framework Web.

FALTA

5.5.2 Base de datos.

FALTA

5.5.3 Roles y Usuarios.

FALTA

5.5.4 Manejo de los datos.

FALTA

5.5.4.1 Modelado.

FALTA

5.5.4.2 Filtrado.

FALTA

5.5.4.3 Visualización.

FALTA

5.6 Conclusiones.

FALTA

5.7 Caso de Uso.

FALTA

5.8 Limitaciones del sistema.

FALTA

6 Próximos pasos.

FALTA - acá poner mejoras

7 Glosario.

FALTA

8 Anexo.

FALTA

9 Bibliografía.

FALTA

10 Agradecimientos.

FALTA