Tarea de Text Mining

Objetivo

El objetivo de esta tarea es aplicar las técnicas de *TF-IDF* (Term Frequency-Inverse Document Frequency) y *cosine similarity* para analizar los documentos asociados a los procesos de compra pública. Este análisis permitirá clasificar los documentos según su relevancia y, potencialmente, identificar procesos de alto riesgo basados en el contenido textual, así como detectar anomalías en los procesos de compra.

Descripción del Dataset

El dataset contiene información de procesos de compra pública bajo la modalidad de Subasta Inversa Electrónica (SIE) del sistema SOCE de Ecuador, proporcionado por el proyecto Kapak. El dataset incluye tres tablas principales:

- **Tabla de Procesos de Compra (**process_info.csv**)**:
  - Columnas: sl_contract_id (identificador numérico del proceso), sd_entidad (entidad que realiza la compra), sd_objeto_de_proceso (categoría de la compra), sie_ic_promedio (indicador ponderado de riesgo de corrupción).
- **Tabla de Preguntas y Aclaraciones (**preguntas_y_aclaraciones.csv**)**:
  - Columnas: sl_contract_id, pregunta_id, pregunta_aclaracion (texto de la pregunta), respuesta_aclaracion (texto de la respuesta).

  Cada proceso puede tener una o más preguntas y respuestas.

- **Tabla de Archivos Asociados (**sample_files.csv**):**
  - Columnas: sl_contract_id, file_name, relative_file_path (ruta del archivo txt).

  Cada fila representa un archivo asociado a un proceso de compra, que contiene texto extraído para facilitar el procesamiento.

  Como cada proceso puede tener uno o varios archivos, estos se encuentran en la carpeta cosine_sample en subdirectorios con el nombre de su sl_contract_id.

Instrucciones para el Análisis

Vectorización de Textos con TF-IDF

Los textos deben convertirse en vectores numéricos utilizando la técnica de *TF-IDF*. Para ello, se debe realizar el preprocesamiento o normalización de los textos primero. Librerías como NLTK o SpaCy facilitan este proceso.

Luego se deben crear vectores de cada archivo de la tabla de archivos. Y también se deben crear vectores que combinan la información de cada proceso. En específico, se requiere concatenar el

contenido de los archivos de cada proceso y convertirlos en un vector. De tal manera que quede un vector de cada proceso. De igual manera, se requiere concatenar el contenido de todas las preguntas y respuestas (de la tabla preguntas_y_aclaraciones) asociadas a un mismo proceso y convertirlo en vector. De tal manera que también quede un vector de preguntas y respuestas por proceso

Nota: Considerar que la similaridad del coseno funciona correctamente solo entre vectores de la misma dimensionalidad. Es decir, con un vocabulario común.

1. Relevancia de Documentos

Cada proceso de compra pública (identificado por su sl_contract_id), está asociado a uno o más documentos, pero no todos son igualmente relevantes. Para determinar cuáles destacan, se intentará clasificar los documentos utilizando la similitud del coseno.

A. Dado un proceso de compra (identificado por su sl_contract_id). Mostrar un ranking de sus documentos con respecto a una pregunta o query realizado. El ranking se hace en base a la similaridad del coseno entre el query y los vectores individuales de los documentos de dicho proceso de compra.
B. Dado un proceso de compra (identificado por su sl_contract_id). Mostrar un ranking de sus documentos con respecto a las preguntas y respuestas de dicho proceso. El ranking se hace en base a la similaridad del coseno entre el vector de preguntas y respuestas de ese proceso y los vectores de cada archivo de ese proceso de compra.

Nota: El vector del query debe tener la misma dimensionalidad que los otros vectores de documentos. Es decir, el mismo vocabulario

2. Análisis de Procesos de Alto Riesgo

La tabla process_info incluye un indicador compuesto de riesgo de corrupción (sie_ic_promedio), que mide el riesgo del proceso entre 0 y 1, donde los procesos con valores mayores a 0.5 se consideran de alto riesgo. En esta sección, se analizará si los vectores de procesos de alto riesgo presentan alta similitud con otros procesos de alto riesgo.

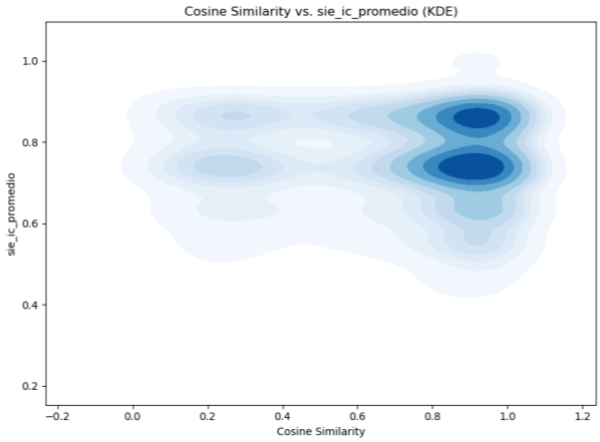
Seleccionar un proceso de compra de alto riesgo (preferiblemente con un sie_ic_promedio cercano a 1) y comparar su vector de proceso con los vectores de los otros procesos. (Estos son los vectores hechos de la concatenación de todos los documentos de un proceso).

A. Imprimir en orden la similaridad de coseno de este proceso de alto riesgo con el resto de los procesos.
B. Hacer un kde plot para mostrar si los procesos parecidos al seleccionado también son procesos con alto riesgo (similaridad vs sie_ic_promedio)

C. Si consideramos que una similaridad mayor a 0.5 “clasifica” a un proceso como de alto riesgo. Comparar la cantidad real de procesos de alto riesgo (sie_ic_promedio > 0.5) con la cantidad de procesos clasificados como de alto riesgo (similaridad > 0.5).

Nota: Para elaborar el kde plot se puede usar la librearía seaborn

```
sns.kdeplot(data=results_single_process, x='cosine_similarity', y='sie_ic_promedio')
```



3. Análisis de Anomalías (Opcional, +1)

Calcular la similitud entre todos los vectores de procesos (Estos son los vectores hechos de la concatenación de todos los documentos de un proceso). Es decir, se comparan todos contra todos, y el resultado es una matriz de similitud (n x n) con la que se puede calcular la similitud promedio entre cada vector de un proceso con los demás (n x 1). Los procesos con baja similitud promedio pueden considerarse atípicos, lo que podría indicar compras poco comunes o potencialmente sospechosas. Encontrar procesos atípicos o hacer un gráfico útil para este análisis.