Universidad Rey Juan Carlos

TEMA 3 – HOJA DE EJERCICIOS I

Se plantean una serie de ejercicios relacionados con representaciones clásicas de textos. Esto incluye representaciones basadas en bolsa de palabras, matrices de coocurrencia y modelos de temas latentes.

Ejercicio 1. Bolsa de Palabras

a. Utilizando una representación de bolsa de palabras (BoW) se quiere que los rasgos del vocabulario no sean palabras, sino n-gramas.

   Los n-gramas son secuencias contiguas de n elementos en un texto, cruciales para captar la información contextual. Pueden mejorar la representación de los textos al no considerar palabras sueltas, sino combinaciones de estas, lo que proporciona un contexto más rico.

   Dadas las siguientes frases: “Estamos ya a finales de febrero.”, “En febrero sigue haciendo frío.”, “Esto es una frase de ejemplo que habla de un mes del año.” Se pide lo siguiente:

   - Generar una representación BoW con bigramas (2-grams).
   - Generar una representación BoW con unigramas, bigramas y trigramas.

   En cada caso se debe de imprimir el vocabulario generado y la matriz de rasgos-documentos.

   Nota: con `CountVectorizer` de `sklearn` se pueden especificar los n-gramas a la hora de crear los vectores de la bolsa de palabras.

b. Dadas varias frases, genera una representación BoW e imprime el vocabulario generado y la matriz de rasgos-documentos con:
   - Función binaria
   - Frecuencia de términos
   - TF-IDF

Ejercicio 2. Matriz de coocurrencias

a. Se quiere crear una matriz de coocurrencias para el siguiente párrafo:

"Mysterious tunnels sketched by Leonardo da Vinci in the late 1400s may have been found at the Castle. Secret tunnels at the Sforza Castle."

Los pasos que hay que seguir para hacer esto podrían ser los siguientes:

- Preprocesar el texto eliminando stopwords, convirtiendo las palabras a minúsculas, tokenizando y eliminando los tokens no alfanuméricos.
- Generar el vocabulario (palabras únicas en el texto, sin repeticiones).
- Decidir el tamaño de ventana del contexto y buscar todos los pares de palabras que coocurren dentro de ese tamaño de ventana, contando las veces que coaparecen.

Para este ejercicio poner un tamaño de ventana para el contexto = 2.

- Crear la matriz de coocurrencias utilizando las frecuencias calculadas antes y visualizarla en un DataFrame de Pandas.

b. A partir de la matriz de coocurrencias del ejercicio anterior, se pide obtener la similitud entre diferentes pares de palabras, utilizando la similitud coseno.

Obtener la similitud coseno para los pares:

- "da" – "vinci"
- "may" – "tunnels"
- "tunnels" – "castle"
- "secret" – "tunnels"

Nota: se puede utilizar la librería sklearn para calcular la similitud coseno (`from sklearn.metrics.pairwise import cosine_similarity`). Para indicar las palabras se le pueden decir los índices de la posición que ocupa la palabra correspondiente en el vocabulario.

c. Utilizando la matriz de coocurrencias y la similitud coseno, obtener las 5 palabras más similares a las siguientes palabras:
   - "leonardo"
   - "da"

   Analizar la salida. Es importante tener en cuenta cuál era el tamaño de la ventana del contexto. Si se cambia el tamaño de ventana, ¿cambian las palabras más similares?

Ejercicio 3. Latent Dirichlet Allocation

a. Dado un fichero en formato JSON con noticias se quiere obtener los diferentes topics utilizando LDA. En este primer ejercicio se proporciona el notebook completo hacerlo, los pasos son los siguientes:
   - Lectura del fichero y carga de datos.
   - Preprocesamiento básico de los textos.
   - Entrenamiento del algoritmo Latent Dirichlet Allocation de sklearn. Se recomienda ver la documentación de la librería para ver los diferentes parámetros del algoritmo.
   - Visualización de resultados.
   - Evaluación de los resultados.

   Se pide ejecutar el código proporcionado y realizar ajustes sobre él para ver cómo cambian los resultados:

   - Cambiar el número de topics, ahora está a 2, pero revisando el fichero con las noticias, en realidad serían más.
   - Cambiar el preprocesamiento de los textos, eliminando stopwords al menos y lo que se considere para ver si mejoran los resultados.

   ¿Cómo impacta en los resultados los diferentes cambios de los apartados anteriores?

b. Sobre la base del ejercicio anterior se quiere añadir una nueva noticia y realizar inferencia para ver los topics en ella.

   El texto de la noticia a inferir es el siguiente:

   “Trump ordena suspender toda la ayuda militar de Estados Unidos a Ucrania tras su bronca a Zelenski.”

   Se tiene que mostrar en qué % se ajusta la noticia a los diferentes topics existentes.