Universidad Rey Juan Carlos

TEMA 3 – HOJA DE EJERCICIOS VI

Se plantea comparar distintos métodos de representación de texto para la recuperación de información textual.

Ejercicio 1. Recuperación de la información – Búsqueda con vectores dispersos

La recuperación de información es el proceso de búsqueda, extracción y recomendación de documentos relevantes dentro de un amplio conjunto de datos en función de una consulta específica. Este procedimiento es de vital importancia en aplicaciones como los motores de búsqueda, los sistemas de recomendación y el análisis de texto en general.

Dado que los documentos suelen estar compuestos por datos no estructurados y, en particular, por texto en lenguaje natural, es necesario en primer lugar transformarlos en una representación computable. Con dicho fin, se pueden utilizar las técnicas de vectorización tradicionales basadas en bolsas de palabras vistas en ejercicios anteriores, las cuales convierten los textos en vectores dispersos dentro de un espacio numérico. A partir de estas representaciones vectoriales de documentos, es posible comparar documentos mediante el empleo de métricas y similitudes como la del coseno, facilitando la identificación de los documentos más relevantes en función de una consulta concreta.

Sigue los siguientes pasos:

- Descarga el fichero *train.xlsx* del conjunto de datos *The Spanish Fake News Corpus*. Puedes obtener el fichero a través del link: https://github.com/jpposadas/FakeNewsCorpusSpanish/tree/master

- Vectoriza la columna "Headline" con diferentes preprocesamientos (lematización, filtrado por nombres, adjetivos y verbos, etc), así como vectorizaciones basadas en bolsas de palabras (como matriz TF-IDF).

- Calcula la matriz de similitud coseno de los documentos anteriores. Para el segundo documento (índice 1), obtén sus 4 documentos más similares en función de dicha similitud (5 en total si consideramos al propio documento de consulta). Analiza los resultados. ¿Qué documentos son más similares? ¿Qué grado de similitud tienen los distintos documentos? ¿A qué se debe?

- Repite el proceso anterior para la columna “Text”. Analiza de nuevo los resultados. ¿Qué diferencias ves con los resultados de la columna “Headline”?

Ejercicio 2. Recuperación de la información – Búsqueda con vectores densos estáticos

Además de las anteriores técnicas de vectorización basadas en bolsas de palabras, los documentos también pueden representarse en un espacio numérico mediante el empleo de modelos basados en embeddings. Como vimos en ejercicios anteriores, estas representaciones densas pueden ayudar a identificar mejor las relaciones semánticas presentes entre palabras y documentos.

Apartado 2.1. Embeddings a nivel de palabra

Si bien Word2Vec genera embeddings a nivel de palabra, es habitual emplear este modelo para representar textos mediante diversas estrategias, como el cálculo de la media de los embeddings de las palabras que los constituyen.

Sigue los siguientes pasos:

- Descarga el modelo de Word2Vec en español “SBW-vectors-300-min5.txt” a partir del siguiente enlace. Nota que es necesario loggearse con una cuenta en Kaggle:
  https://www.kaggle.com/datasets/rtatman/pretrained-word-vectors-for-spanish

- Cárgalo mediante el siguiente código de ejemplo. Nota que <TU_PATH> debe reemplazarse por la ubicación donde se encuentra el modelo en tu sistema de archivos:

```
model = KeyedVectors.load_word2vec_format('<TU_PATH>SBW-vectors-
        300-min5.txt', binary=False)
```

- Crea un embedding por cada documento usando la columna “Headline”. Prueba diferentes preprocesamientos y crea el embedding a nivel de documento mediante el cálculo de la media de los embeddings de las palabras que lo componen. Repite el proceso anterior para la columna “Text”.

- Calcula las matrices de similitud coseno para las columnas “Headline” y “Text”. Para el segundo documento (índice 1), obtén sus 4 documentos más similares en función de estas matrices de similitud (5 si contamos al mismo documento).

Analiza los resultados. ¿Qué documentos son más similares? ¿Qué grado de similitud tienen los distintos documentos? ¿A qué se debe? ¿Qué diferencias ves con las vectorizaciones basadas en bolsas de palabras?

Apartado 2.2. Embeddings a nivel de documento

Aunque puede que no se dispongan de los suficientes datos para entrenar de forma adecuada un Doc2Vec, es posible analizar cualitativamente cómo le afectarían distintos preprocesamientos y datasets a la formación de embeddings a nivel de documento para la creación de un sistema recomendador de documentos.

Sigue los siguientes pasos:

- Repite el procedimiento anterior entrenando un Doc2Vec desde el inicio sobre las columnas “Headline” y “Text”. Experimenta diferentes configuraciones de entrenamiento y preprocesamientos de textos.

- Calcula las matrices de similitud coseno y obtén los 4 documentos más similares al primer documento (índice 1).

- Analiza los resultados de Doc2Vec. ¿Cómo varían los resultados en función del número de palabras? ¿Funciona mejor usar solo los titulares o el texto completo?