Representaciones Vectoriales de Textos: Bag of Words y Word Embeddings

Fabián Villena

Julio 2025

El análisis vectorial de textos permite representar el contenido semántico de un documento en un espacio vectorial, facilitando tareas de clasificación, agrupamiento y búsqueda semántica. En este laboratorio explorará tanto el modelo clásico de Bolsa de Palabras (*Bag of Words*), que representa los textos como vectores de frecuencia, como las técnicas modernas de **word embeddings**, donde palabras con significado semántico similar poseen representaciones vectoriales próximas.

Todo el desarrollo debe realizarlo en un *Jupyter Notebook* en Python. Los ejercicios usan como base resúmenes de publicaciones científicas:

- **Conjunto reducido:** `https://users.dcc.uchile.cl/~fvillena/files/abstracts_small.zip`
- **Conjunto grande:** `https://users.dcc.uchile.cl/~fvillena/files/abstracts_large.zip`

Parte 1: Bolsa de Palabras y Similaridad Coseno

Utilizando el conjunto reducido de resúmenes, realice lo siguiente:

1. Implemente una función para calcular la frecuencia de cada palabra en un documento. Utilice esta función para representar cada resumen como un vector de frecuencia (espacio $\mathbb{N}_0^{|V|}$, donde $|V|$ es el tamaño del vocabulario).
2. Implemente una función que calcule la similaridad coseno entre dos vectores y aplíquela a cada par de documentos.
3. Construya un gráfico de mapa de calor visualizando las similaridades entre todos los resúmenes.

Parte 2: Word Embeddings y Agrupamiento Semántico

Utilizando el conjunto grande de resúmenes, y embeddings pre-entrenados en español:

1. Descargue e los word embeddings pre-entrenados[1]. Cargue el modelo en Python y ejemplifique cómo obtener el vector para una palabra cualquiera.
2. Desarrolle una función que reciba un texto y devuelva un vector representativo utilizando los word embeddings.
3. Aplique un algoritmo de clustering (por ejemplo, KMeans) sobre las representaciones vectoriales de los resúmenes y visualice, mediante reducción de dimensionalidad (por ejemplo, t-SNE o PCA), la distribución de los documentos.

[1] `https://zenodo.org/records/6647060/files/mix.vec?download=1`