Actividad 3: Representación de texto moderna Word Embeddings

Teoría

1. ¿Qué es un Word embeddings?
2. ¿Qué podemos representar en un espacio vectorial?
3. ¿Qué elementos necesitamos para entrenar / utilizar un modelo de embeddings? 4
4. Enumerar y describir qué mediciones se usan para comparar dos embeddings
5. Explicar cómo se obtienen word embeddings utilizando la técnica Word2Vec (2 enfoques de entrenamiento).
6. Enumere y explique las diferencias de este enfoque y problemas que resuelve con respecto a los anteriores (tradicionales).
7. Realizar una tabla comparativa general entre enfoques tradicionales (Word vectorization) y modernos de representación de texto (Word embeddings).
8. Realizar una tabla comparativa con todos los modelos de la familia Word2Vec para representar texto (Word2Vec, Doc2Vec, GloVe y FastText).

Práctica

En esta actividad práctica aplicar los conceptos vistos en la teoría. Para ello realizar una copia, ejecutar y analizar los colabs 06 - Modelo Word2Vec preentrenado.ipynb y 10 - Modelo GloVe preentrenado.ipynb utilizando la biblioteca Gensim para responder las siguientes preguntas a modo de reflexión:

1. ¿Por qué se utiliza un benchmark de analogías para evaluar los resultados obtenidos del entrenamiento de embeddings?
2. ¿En qué consiste la mejora que introdujo Gensim en 2016 para evaluar la similaridad de palabras de los modelos entrenados?
3. ¿Cómo se puede hacer finetuning de un modelo de Gensim con nuestras palabras adicionales?
4. Enumere y explique las diferentes funciones exploradas en el colab de GloVe para trabajar con similaridad entre palabras.
5. Investigar las técnicas PCA (Principal Components análisis) y TSNE e indicar para que se utilizaron en el colab de GloVe
6. ¿Qué usos útiles puede encontrar aplicar similaridad entre documentos de texto ?.