Fernanda Pérez Ruiz A01742102

DIA | MES | AÑO

1-Investigue la estrategia de vectorización TF-IDF. ¿Cómo se calcula? ¿En qué situaciones es más efectivo usar TF-IDF para tareas de clasificación de texto? ¿Con qué bibliotecas se pueden implementar?

Se calcula multiplicando la frecuencia del término (TF) por la ~~[illegible]~~ inversa de la frecuencia en los documentos (IDF) La fórmula es:

$$TF-IDF(t,d,D) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

N es el número total de documentos y DF(t) es el numero de documentos que contiene el término t.

TF-IDF es más efectivo en tareas donde se quiere destacar palabras importantes en la clasificación de textos, como la extracción de palabras clave o categorización de documentos.

Las bibliotecas que se pueden implementar con scikit-learn (TfidVectorizer), NLTK y Gensim

Norma

2- ¿Qué problemas de los N-gram resuelve el "Laplace smoothing"? ¿Cómo trabaja? ¿y qué pasa con un modelo de NLP cuando se emplea esta técnica?

Evita que un N-gram tenga una probabilidad de 0 cuando no aparece en los datos de entrenamiento, lo cual puede romper el modelo.

Suma un valor (generalmente 1) a todos los conteos de N-grams, incluyendo los que no han aparecido. Esto garantiza una probabilidad no nula para cada N-gram.

Hace que el modelo sea más robusto frente a datos no vistos, pero puede sobreestimar la probabilidad de N-grams raros.

3- ¿Qué pasa cuando una palabra en el test set no se encuentra en el vocabulario del modelo de los N-gram? ¿Cómo se puede modelar la probabilidad de palabras 'out-of-vocabulary'? (OOV?)

Si una palabra en el test set no está en el vocabulario, el modelo no puede calcular su probabilidad, lo que afecta su desempeño.

Para modelarlo se puede introducir un token especial (<UNK>) para palabras desconocidas o aplicar un suavizado (como Laplace) para asignar probabilidades a palabras fuera del vocabulario.