Tarea 2 Eryk Elizando Ganzález - 101284899

1. Estrategia de vectorización 1 FoIDF - Définicion : Es una (estrategia) téchica de vectorización que comb ina la frecuencia de un término en un documento CTF) y su rareza en toda la colección (IDF). - Cálculo: Se calala como el producto de TF (Term Frequency) Y IDF (Inverse Downent Frequency)

* TF-IDF (+,d) = TF (+,d) X IDF (+,D) * 1 F (t,d) = Número de veces que aparece el término t en documento d Primero total de términos en el documento d *IDF(+,0) = log (Número total de documentos Womero de dommentos que contienen t) - ¿ Coundo es mais efectivo? · Tareas de clasificación de texto y busqueda de texto coando: c Enfatizar términos que son importantes en un documento o Evitar que las palabras comunes tengan un peso en la clasificação o Procesar textos pequeños - medianos con vocabularios amplios - Bibliotecas para implementar TF-IDF · Scikit-learn (Tfidf Vectoricer) · NLTK (Parcialmente) · spacy (Parcialmente) 2. Laplace Smoothing - i Qué problema de los N-grams resuelve Laplace smoothing? · Problema de probabilidad O en los modelos N-gramal

· Problema de los N-grams resuelve Laplace smoothing?
· Problema de probabilidad O en los modelos N-gram a i
ciñadir un valor constante a los conteos, asignando probabilidades a secuencias ne vistas. Esto mejora la
generalización del modelo, pero puede dar aportunida
probabilidades más altas a secuencias poco comunes.

Maeria

3. Modelado de palabras fuera del Jocabulario (001) · Coando una palabra del test set no está en el Vocabulario (Oov), el modelo no puede asignarle probabilidad. Para mone jarlo, se usa un token oov especial, técnicas de smoothing o vocabularios mas cemplios para evitar probabilidades nulas * 1 + (t,d) = Nomero de veces que aparece el término + en document Nomero total de términos en el documen MIDFG, D) = log (Numero total de documentos Nomera de documentos que cartienen +) Tareas de clasificación de texto y bisqueda de texto cuando: e Enforticar terminos que son importantes en un documento o Evitor que lois palabras comunes tengan un pesa en la cles se o Procesar textos pequeños - medianos con vocabularios amp - Bibliotecas pare implementar 15-10F Scikit - learn CTfidf Vectoricer) NUTIC (Parcialments) spacy (Possialmente) Laplace Smoothing EQUE problems de Pos N-grams resuelve Laplace smoothing · Problem de probabilidad Q en los modelos U-grame curadir un valor constante a los contros, asignando pro babilidadis a secuencias no vistas. Esta mejara a eneralización del modelo, pero prede dar ignaturados probabilidades más altas a socienciais paro comunes.