30/01/2024

TF-IDF

El termino de frecuencia/frecuencia inversa del documetno, cuantifica la importancia de una palabra dentro de un texto y de una determinda coleccion de corpus. Usos

* Clasificar en un motor de busqueda
* Hacer un resumen de texto
* Construer chatbots inteligentes

Frecuencia de una palabra en un document, frecuencia de una palabra en el corpus.

La tecina asigna puntuaciones a las palabras:

* Si una palabra aparece de forma frecuente en un documento, podemos determinar que tiene una puntuación alta. (TF)
* Si la palabra que aparece mucho en un documento no lo tiene en el resto de documentos, tiene una puntuación baja (IDF)

Pasos a seguir:

1. Se calculan los valores de frecuencia de un termino
2. Se calcula los valores de frecuencia inversa en el doucmetno de corpus
3. Se combina 1 y 2
4. TF(t, d): Frecuencia del termino entre el numero de términos en el documento
5. IDF(t, D): rareza del termino en el conjunto de documentos D

Log(numero total de documentos/numero de documentos que contiene t + 1)

1. Multiplicar paso 1 y 2

Ejemplo

1. Partimos de un corpus
2. Preprocesar el texto
3. Crear una matriz tf
4. Normalizar valores (ocurrencias entre totales)
5. Crear valores IDF)
6. Calcular TF-IDF asociada

Implementación

La librería scikit-learn proporciona 3 clases para este modelo

* TfidfVectorizer: obtiene la matriz tf-idf