TF-IDF Vectorize

TF-IDF or Term Frequency - Inverse Document Frequency. To understand what is TF-IDF, we need to know what TF and IDF are:

+ TF: As the name suggested, it is the term frequency. To make it easier to understand, we will show an example:

Suppose we have a set of document texts and the problem is how to rank those texts based on the query such as “the music”. One simple way to find is to eliminate any documents that do not have the words or basically we call terms “the”, “music” to solve, but the remaining texts are still many. To further describe them, those terms need to be counted in each document, the number of terms appearing in document is term frequency.

For anyone who knows Bag of Word, TF is similar to it in counting how many the terms are appearing. The difference between them is TF use frequency which is number of terms appearing divided by total number of words in the document. TF easily shows how many percentage those terms consist of in the document

+ IDF: It is contrast to the TF. If TF is to count the terms, then IDF is to determine how rare, important the terms are.

Using the example above, suppose the problem is how determine how rare, important those words in the given documents. To determine the value of the terms, the documents that have those terms need to be counted to know if those terms are used in common documents.

IDF is the frequency so the algorithms for the IDF is by number of documents divided by the number of document that have the terms appeared. The logarithm is added to dampen the importance of those values to be used for situation such as TF-IDF to make the two values TF and IDF have the same important value.

=> So what is TF-IDF ?: TF-IDF is by multiplying those two values to adjust the value of the term based on following properties: how common the terms are in each document and how rare, importance the terms are used in given documents.

The algorithm: TF: 

 Number of times the word (t) appear in one document (doc)

: Total number of words in the document

IDF: 

: total number of documents

: the number of documents that contain the word (t)

TF-IDF: 