How to compute TF-IDF

Suppose we are looking for documents using the query *Q* and our database is composed of the documents *D1*, *D2,* and *D3*.

- *Q*: The cat.
- *D1*: The cat is on the mat.
- *D2*: My dog and cat are the best.
- *D3*: The locals are playing.

TF(word, document) = "number of occurrences of the word in the document" / "number of words in the document"

TF scores of the words "the" and "cat" (i.e. the query words) with respect to the documents *D1*, *D2,* and *D3*.

TF("the", D1) = 2/6 = 0.33
TF("the", D2) = 1/7 = 0.14
TF("the", D3) = 1/4 = 0.25
TF("cat", D1) = 1/6 = 0.17
TF("cat", D2) = 1/7 = 0.14
TF("cat", D3) = 0/4 = 0

IDF can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

If the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

IDF(word) = log(number of documents / number of documents that contain the word)

Let's compute the IDF scores of the words "the" and "cat".

IDF("the") = log(3/3) = log(1) = 0
IDF("cat") = log(3/2) = 0.18

Multiplying TF and IDF gives the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

TF-IDF(word, document) = TF(word, document) * IDF(word)

Let's compute the TF-IDF scores of the words "the" and "cat".

TF-IDF("the", D1) = 0.33 * 0 = 0
TF-IDF("the, D2) = 0.14 * 0 = 0
TF-IDF("the", D3) = 0.25 * 0 = 0

TF-IDF("cat", D1) = 0.17 * 0.18= 0.0306
TF-IDF("cat, D2) = 0.14 * 0.18= 0.0252
TF-IDF("cat", D3) = 0 * 0 = 0

The next step is to use a ranking function to order the documents according to the TF-IDF scores of their words.

We can use the average TF-IDF word scores over each document to get the ranking of *D1*, *D2,* and *D3* with respect to the query *Q*.

Average TF-IDF of D1 = (0 + 0.0306) / 2 = 0.0153
Average TF-IDF of D2 = (0 + 0.0252) / 2 = 0.0126
Average TF-IDF of D3 = (0 + 0) / 2 = 0

The word "the" does not contribute to the TF-IDF scores of each document. This is because "the" appears in all of the documents and thus it is considered a not-relevant word.