Document Retrieval System

By

Saurabh Goyal

tf-idf weight

It's full form is term frequency inverse document frequency that reflects the importance of word in a corpus.

It has two components in its weight namely term frequency and inverse document frequency.

tf-weight

The number of times a term occurs in a document is called term frequency.

- A document with 10 occurrence of a term is more relevant than a document with 1 occurrence of a term
- But not 10 times more relevant.
- Hence relevance doesn't increase proportionally with frequency.

Thus, $W_{t,d} = 1+\log_{10}(f_{td})$ if $f_{td}>0$ otherwise 0

Idf-weight

To give importance to the meaningful terms in a document those are less frequent inverse document frequency is used.

- Frequent terms are less informative than rare terms, since frequent terms like (high, increase, line) are not a sure indicator of relevance.
- Thus df weights are used to capture the importance of less frequent weights.

Thus, $idf_t = \log (N/df_t)$ where N is the total number of documents.

There is only one idf value for each term in the collection.

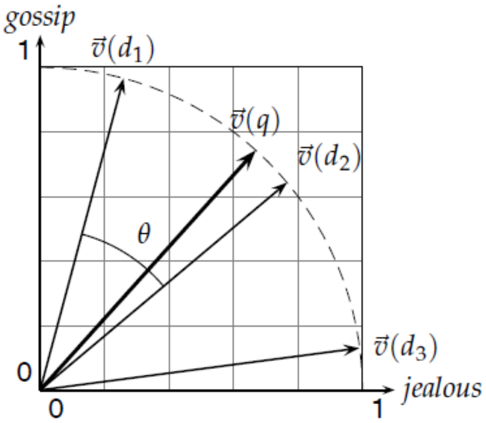
Thus over all tf-idf weight is

$$\mathbf{W_{t,d} = (1+\log tf_{t,d}) * (\log_{10}(N/df_t))}$$

The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Vector Space Model

It is an algebraic model for representing text documents as vectors. It is fundamental to a host of information retrieval operations ranging from scoring documents on a query, document classification and document clustering.



It can be seen above that d1 vector contains term gossip more whereas d3 will have jealous more. Hence we plot query term vector also on the graph and compare the cosines of all the vectors of document with the query. One with the lowest angle will have the highest cosine score and hence the highest similarity.