Description of implementation

Vector space model-cosine similarity

The documents and queries are vectors in t-dimensional vector space where t represents the number of terms.

Cosine similarity(Q,D) =

Q is the query

D is the document

Q.D = Sum(qt\*dt)

||Q|| is sqrt(sum of squares of term qt)

||D|| is sqrt(sum of squares weights of all terms in the document)

qt is tf-idf weight of term t in query Q(which is equal to term frequency of the term in query. Idf =1 in this case)

dt is the tf-idf weight of term t in document D

tf-idf=tf\*idf

tf=term frequency

idf=log(N)/document frequency of term

where N is number of documents in the corpus

For query term:

tf is term frequency of query term

idf is 1 in weight calculation for term in query

For document:

tf is the term frequency of term in query in the document

idf is inverse document frequency which determines the importance across all documents. Idf is calculated by taking log of N/document frequency of the term

I have arranged the documents in the decreading order of cosine similarity measures indicating the top documents are more relevant to the given query.