1. True or false: The boolean model does not rank documents in the answer, while the vectorial model produces a ranking.
2. Suppose you are given the frequency of every term in a given document. What other

information do you need to compute its representation in tf-idf weights?

1. Consider the following collection of five documents:

Doc1: we wish efficiency in the implementation for a particular application

Doc2: the classification methods are an application of Li's ideas

Doc3: the classification has not followed any implementation pattern

Doc4: we have to take care of the implementation time and implementation efficiency

Doc5: the efficiency is in terms of implementation methods and application methods

Assuming that every word with 6 or more letters is a term, and that terms are ordered in order of appearance,

a . Give the representation of each document in the boolean model.

b. Give the representation in the vector model using tf-idf weights of documents Doc1 and

Doc5. of documents Doc1 and Doc5. Compute the similarity coefficient, using the cosine

measure, among these two documents.

1. Suppose that terms A, B, C, and D appear, respectively, in 10,000, 8,000,5,000, and 3,000 documents of a collection of 100,000.

Compute the similarity of the documents “A B B A C C” and “D A DB B C C” using tf-idf weighting and the cosine measure.

1. We have an indexed collection of one million documents that includes the following terms:



1. Compute the similarity between the following documents D1 and D2 using tf-idf weights and the cosine measure:

D1 = "p2p programs help users sharing files, applications, other programs, etc. in computer networks"

D2 = "p2p networks contain programs, applications, and also files"

1. Assume we are using the cosine measure and tf-idf weights to compute document

similarity. Give a document containing two different terms exactly that achieves maximum similarity with the following document

"p2p networks contain programs, applications, and also files"

Compute this similarity and justify that it is indeed maximum among documents with two terms.