COMP3009J Information Retrieval

Worksheet 6: Vector Space Model with TF-IDF Weights

For this Worksheet, repeat the tasks from Worksheet 5, but instead of using binary weights (1 and 0), use TF-IDF.

Advanced

a) Instead of the small document collection included in Worksheet 5, load the documents from the npl-doc-text.txt file we used last week, and allow a user to enter a query and see the first 10 results.

b) Compare the results of using the Vector Space Model on the npl-doc-text.txt collection using binary weights and using TF-IDF. Can you identify queries that give very different results for the same query (**Hint**: think of a query that combines rare words with frequent ones)? Does this help or hinder the retrieval in your opinion? Why?