COMP3009J Information Retrieval

Worksheet 6

Question 1 (from 2018 exam paper)

Below is a small document collection, containing three documents. Answer the questions that follow.

Stopwords: he, his, in, was, is

Document 1: He washed his coat in New York.

Document 2: My dog's coat was washed yesterday.

Document 3: My new coat is very, very warm.

(i) Describe the preprocessing steps you would use when creating an index for these documents.

(ii) Calculate a vector to represent each document, using the TF-IDF weighting scheme. You should use the stopword list provided, but do not perform stemming.

(iii) Calculate the cosine similarity for each vector using the query "his new coat", and show the final ranked list of documents for this query.

(iv) Describe what changes you would make so that users can search for two-word phrases (e.g. "new coat").

Question 2 (from 2018 resit exam paper)

Below is a small document collection, containing three documents. Answer the questions that follow.

Stopwords: and, of, over, the, then.

Document 1: The bank had many, many rolls of coins.

Document 2: The coins rolled over the river bank.

Document 3: The plane rolled and then banked.

(i) Calculate a vector to represent each document, using the TF-IDF weighting scheme. You should use the stopword list provided, but do not perform stemming.

(ii) Calculate the cosine similarity for each vector using the query “many coins rolled”, and show the final ranked list of documents for this query.

(i) If stemming had been used, what effect would it have had on the index?

Question 3

A user wishes to search the NPL dataset (this dataset is available on Moodle as `npl-doc-text.txt`).

The user enters the query "systems data coding information transfer".

Before any feedback occurs, what are the initial values for the following probabilities for each term?

i) $P(k_i|R)$
ii) $P(k_i|\bar{R})$

Following the first iteration, the user marks the following documents as relevant (i.e. the set *V* contains these documents):

1136, 2042, 2175, 3595, 4056, 4412, 5826, 6138, 7762, 7985

Using the formulae on Slide 31 of Lecture 4 (entitled "Improving Probabilities"), **calculate the new probabilities for each term.**

For this question, you may choose to write a program to calculate these probabilities, or combine some programming and some manual work to calculate them.