COMP3009J INFORMATION RETRIEVAL

Exercises for Vector Space Model: Solution for Q1

VECTOR SPACE MODEL: TF-IDF

Q1. Below is a small document collection, containing three documents. Answer the questions that follow. (Adapted from 2018 resit exam paper)

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.

Step 1: preprocessing

- We are asked to remove stopwords but not perform stemming. At this stage, we would also convert any uppercase letters to lowercase. After this process, the documents' terms become:
    - Document 1: bank, had, many, may, rolls, coins
    - Document 2: coins, rolled, river, bank
    - Document 3: plane, rolled, banked

Step 2: construct TF-IDF vectors

- To find the number of dimensions in the vectors, we must count how many distinct (unique) terms there are in the document collection. They are:
    - bank, had, many, rolls, coins, rolled, river, plane, banked
- Now the IDF scores can be calculated (we will organise the terms in the vector in alphabetical order here, just to make it easier to display – in reality the order is not important as long as it is the same for all vectors):

Term	bank	banked	coins	had	many	plane	river	rolled	rolls
IDF	0.5850	1.5850	0.5850	1.5850	1.5850	1.5850	1.5850	0.5850	1.5850

- Next we can calculate TF scores. Note that for document D1, the term "many" appears twice. This means that the most common frequency in that document is 2, which explains why the other terms have a TF of 0.5.

Term	bank	banked	coins	had	many	plane	river	rolled	rolls
D1	0.5	0.0	0.5	0.5	1.0	0.0	0.0	0.0	0.5
D2	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0
D3	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0

- Finally we finish creating our vectors by multiplying the TF scores for each document by the IDF scores.

Term	bank	banked	coins	had	many	plane	river	rolled	rolls
D1	0.2925	0.000	0.2925	0.7925	1.5850	0.0000	0.0000	0.0000	0.7925
D2	0.5850	0.0000	0.5850	0.0000	0.0000	0.0000	1.5850	0.5850	0.0000
D3	0.0000	1.5850	0.0000	0.0000	0.0000	1.5850	0.0000	0.5850	0.0000

Step 3: construct the query vector

We are now ready to receive a query. Here, the query is “many coins rolled”. For the IDF, we take the IDF score that was calculated for the document collection. We do **not** recalculate a new IDF score. The TF is calculated in the same was as for documents. In this case because each query term occurs exactly 1 time, the TF for every term is 1. Thus the query vector is:

Term	bank	banked	coins	had	many	plane	river	rolled	rolls
Q	0.000	0.000	0.5850	0.000	1.5850	0.0000	0.0000	0.5850	0.000

Step 4: compute cosine similarity and rank the documents

We compute the cosine similarity between the query and each of the documents. First we need the lengths of all the vectors:

- **Q:** 1.7879
- **D1:** 1.9848
- **D2:** 1.8811
- **D3:** 2.3165

Finally the cosine similarities can be calculated and the documents can be ordered from the highest similarity to the lowest:

- **D1:** 0.7562
- **D2:** 0.2035
- **D3:** 0.0826