

COMP3009J INFORMATION RETRIEVAL

Exercises for Week 5 (Phrase queries and Vector Space Model)

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.

Q2. Below is a small document collection, containing three documents. Answer the questions that follow. (Adapted from 2018 exam paper)

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) 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 “his new coat”, and show the final ranked list of documents for this query.

Q3. Below is a small document collection, containing three documents.

Stopwords: a, as, had, of, was

Document 1: Mary had a little lamb, little lamb, little lamb.

Document 2: Mary had a little lamb.

Document 3: Whose fleece was white as snow.

Calculate the cosine similarity for each document using the query “white lamb of mary” and show the final ranked list for this query. You should use the TF-IDF weighting scheme, using the stopwords list provided (do not perform stemming). Show all workings. (Adapted from 2019 exam paper)