COMP3009J Information Retrieval

Exercises: Phrase queries and Vector Space Model

Phrase Queries

Q1. Below is part of a positional index relating to the term “friend”. In creating this index, stopword removal and stemming have not been used. Postings lists begin at 1 for the first term in each document. Which document(s) could contain the phrase “a friend in need is a friend indeed”? Explain your answer. (From 2019 exam paper)

```
<friend: 22980;
d1: 1, 6, 10, 16;
d2: 28, 97, 138, 143, 150, 151;
d3: 25, 30, 31, 65, 188, 209;
d4: 4, 34, 119, 121, 159, 177;
d5: 55;
...>
```

Q2. Below is part of a positional index relating to the term “fear”. In creating this index, stopword removal and stemming have not been used. Which document(s) could contain the phrase “the only thing we have to fear is fear itself”? Explain your answer. (From 2018 exam paper)

```
<fear: 215230;
1: 2, 4, 130;
2: 20;
3: 134, 199;
4: 7, 100, 102, 156, 279;
5: 8, 88, 888, 890, 891;
...>
```

Q3. Below is part of a positional index relating to the term “another”. In creating this index, stopword removal and stemming have not been used. Which document(s) could contain the phrase “another day another dollar”? Explain your answer. (From 2018 resit paper)

```
<another: 41825;
1: 3, 5, 10;
2: 20;
3: 121, 162;
4: 4, 101, 103;
5: 7, 77, 777, 779, 780;
...>
```

VECTOR SPACE MODEL

Q4. 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 a simple binary weighting scheme (i.e. $w_{i,j} = 1$ if document *j* contains term *i*, and $w_{i,j} = 0$ if document *j* does not contain term *i*). 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.

Q5. 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 a simple binary 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.

Q6. 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 a simple binary weighting scheme, using the stopword list provided (do not perform stemming). Show all workings. (Adapted from 2019 exam paper)