TDT4117 Information Retrieval - Autumn 2022
Assignment 1 - Solution

September 1, 2022

Task 1 : IR Models Definitions

A. For each of the classical information retrieval models Boolean, Vector Space, state why it has not been employed within Web search engines?(3 reasons each)

B. Why is the Vector Space Model(VSM) not suited to the following situations:

1. The document collection is volatile
2. queries are likely to predominate

C. With respect to VSM, How and why are the document vectors normalized to unit length?(explain thoroughly)

Task 2: IR Models

Assuming the following document collection, which contains only the words from the set $O = \{Cloudy, Sunny, Rainy\}$.

```
doc1  = {Sunny Cloudy Rainy Rainy}
doc2  = {Rainy Cloudy}
doc3  = {Sunny Sunny Cloudy}
doc4  = {Cloudy Cloudy Sunny Cloudy Rainy Sunny Rainy Cloudy}
doc5  = {Sunny}
doc6  = {Rainy Rainy}
doc7  = {Sunny Cloudy Rainy}
doc8  = {Rainy Rainy Cloudy }
doc9  = {Rainy Rainy Sunny}
doc10 = {Sunny Cloudy Sunny Rainy}
```

SubTask 1: Boolean Model and Vector Space Model

Given the following queries:

```
q1 = "Rainy AND Cloudy"
q2 = "Cloudy AND Sunny"
q3 = "Sunny OR Rainy"
q4 = "Cloudy NOT Rainy"
q5 = "Sunny"
```

1. Which of the documents will be returned as the result for the above queries using the Boolean model? Explain your answers.

2. What is the dimension of the vector space representing this document collection when you use the vector model and how is it obtained?

3. Calculate the weights for the documents and the terms using tf and idf weighting. Put these values into a document-term-matrix. (Tip: use the equations in the book and state which one you used.)

4. Study the documents 1, 2, 4 and 10 and compare them to document 5. Calculate the similarity between document 5 and these four documents according to Euclidean distance (or any other distance measure, if you choose one other than Euclidean distance explain why).

5. Rank the documents by their relevance to the query $q5$ (use cosine similarity to calculate the similarity scores).

SubTask 2: Probabilistic Models

Given the following queries:

```
q1 = \Sunny Rainy"
q2 = \Cloudy"
```

1. What are the main differences between BM25 model and the probabilistic model introduced by Robertson-Jones?

2. Assuming absence of relevance information, rank the documents according to the two queries, using the BM25 model. Set the parameters of the equation as suggested in the literature. Write clearly all the calculations.

Hint: use this formula.

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{df_t} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{ave})) + tf_{td}},$$

where tf_{td} is the frequency of term t in document d, and L_d and L_{ave} are the length of document d and the average document length for the whole collection.