Assignment 1 - TDT4117

Hermann Owren Elton, Olaf Rosendahl

September 20, 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)

The Boolean model has a very sparse Term-Document matrix meaning that there is a lot of zeroes. There is also no ranking of the documents so how relevant or how accurate a document is to a query is not taken into consideration. In the boolean model a document also need to be an exact match to the query to be counted as a relevant/correct document.

The Vector Space model assumes that the terms in a query are independent so it assumes that the order of the words isn't relevant. The model also doesn't allow partial matching meaning no substrings of a term will result in a match. Similarly the order of terms in a document has no relevance to the rank of the document.

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

1 The document collection is volatile

When new documents are added in a volatile document collection, the document term-frequency matrix must be updated.

2 Queries are likely to predominate

When receiving new queries, one have to do similarity computations. A single of these computations can be very heavy as they might require the dot product in large dimensions, demanding enormous amounts of arithmetic operations.

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

Vector normalization is preformed so that the model doesn't discriminate between large and small documents. This is done by representing each document as weighted term vectors where the dimension of the matrix corresponds to the size of the vocabulary of the given system. So if the vocabulary has a size of 3 the dimension of the weighted term vector is 3. We then normalize the vector to a unit

length by converting it to a unit vector through vector normalization. This is done by dividing each component of the vector with the magnitude of the vector itself. This standardizes the size and shape of the vectors corresponding to each document, by the size of the vocabulary in the system.

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.

   The boolean model considers if the document contains all the term of the query this, when applying the boolean model to this problem given the query and the document selection, the system would return document, q1: 1,2,4,7,8,10 q2: 1,3,4,7,10 q3: 1,2,3,4,5,6,7,8,10 q4: 3,9 q5: 1,3,4,5,7,9,10

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

   For this document collection the dimension of the vector space would be 3 because the vocabulary size for the system is 3 because all documents only contain the words Cloudy, Sunny and Rainy.

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.)

Table 1: TF

	tf1	tf2	tf3	tf4	tf5	tf6	tf7	tf8	tf9	tf10
cloudy	1	1	1	4	-	-	1	1	-	1
sunny	1	-	2	2	1	-	1	-	1	2
rainy	2	1	-	2	-	2	1	2	2	1

Table 2: IDF

cloudy	0.5146
sunny	0.5146
rainy	0.3219

Table 3: Weight

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
cloudy	0.5146	0.5146	0.5146	1.5438	-	-	0.5146	0.5146	-	0.5146
sunny	0.5146	-	1.0292	1.0292	0.5146	-	0.5146	-	0.5146	1.0292
rainy	0.6438	0.3219	-	0.6438	-	0.6438	0.3219	0.6438	0.6438	0.3219

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).

   Euclidean distance is given by = the magnitude of doc1 - magnitude of doc2

$$(5-1)d = \sqrt{(0-1)^2 + (1-1)^2 + (0-2)^2} = 2.236$$
$$(5-2)d = \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2} = 1.732$$
$$(5-4)d = \sqrt{(0-4)^2 + (1-2)^2 + (0-2)^2} = 4.5826$$
$$(5-10)d = \sqrt{((0-1)^2 + (1-2)^2 + (0-1)^2} = 1.732$$

   We can see that a the documents are pretty different, but document 2 and 10 are pretty similar when it comes to magnitude compared with document 5 since they produce the same euclidean distance.

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

Table 4: q5 Relevance

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
q5	1	0	1.9996	1.9996	1	0	1	0	1	1.9996

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?

   The model introduced by Robertson-Jones was mainly a model to be used as a framework for future models. The framework has a lot of holes which makes it inefficient when ranking documents. The algorithm does not have index with weights and it assumes that terms are independent. BM25 is a collection of a lot of scoring functions which ranks documents, with different equations and parameters.

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.

   - Precalculations

$$\text{RSV}_d = \sum_{t \in q} log \left[\frac{N}{df_t} \right] * \frac{(k_1+1)tf_{td}}{k_1((1-b)+b*(\frac{L_d}{L_{ave}}))+tf_{td}}$$

Using these values from the slides:

$\text{k}_1 = 1$
$b = 0.75$

With the following static variables:

$\text{N} = 10$
$\text{L}_{ave} = 33/N = 3.3$

Which gives us this formula:

$$\text{RSV}_d = \sum_{t \in q} log \left[\frac{10}{df_t} \right] * \frac{2*tf_{td}}{(0.25+0.75*(\frac{L_d}{3.3}))+tf_{td}}$$

We'll use the computed values for IDF from subtask 1.3

- `q1 = Sunny Rainy`:

$\text{RSV}_1 = 0.5146 * \frac{2*1}{(0.25+0.75*(\frac{4}{3.3}))+1} + 0.3219 * \frac{2*2}{(0.25+0.75*(\frac{4}{3.3}))+2} = 0,8871$
$RSV_2 = 0,3786$
$RSV_3 = 0,7036$
$RSV_4 = 0,8258$
$RSV_5 = 0,6977$

$RSV_6 = 0,4768$
$RSV_7 = 0,8688$
$RSV_8 = 0,4401$
$RSV_9 = 0,9746$
$RSV_10 = 0,9526$

- `q1 = Cloudy`:

$\mathrm{RSV}_1 = 0.5146 * \frac{2*1}{(0.25+0.75*(\frac{4}{3.3}))+1} = 0,4785$
$RSV_2 = 0,6053$
$RSV_3 = 0,5345$
$RSV_4 = 0,6803$
$RSV_5 = 0$
$RSV_6 = 0$
$RSV_7 = 0,5345$
$RSV_8 = 0,5345$
$RSV_9 = 0$
$RSV_10 = 0,4785$