TDT4117 - Assignment 1

Sander Lindberg

Task 1

1

Explain the main differences between Data retrieval and Information retrieval

IR deals with unstructured data and small errors are allowed. With Information retrieval we want to retrieve *information* rather than *data*.

Data retrieval deals with structured data, allows no errors. Either we get a precise result, or no result at all.

I like to think about data retrieval as retrieving data from a database and information retrieval as getting information about a subject.

2

Explain the main differences between Structured data and Unstructured data

Structured data is a pre-defined data model and straightforward and easy to analyse. Such data can for example be data retrieved with an SQL-query.

Unstructured data on the other hand does not have this pre-defined model, and are therefore harder to analyse. It can be for example text-heavy (such as a document), audio or video.

Task 2

Explain:

1. Term frequency (*tf*)
    - Term frequency is a factor used in calculating the ranking of a document. It represents the frequency of a given term (word) in a document.
2. Document frequency (*df*)
    - Df is the number of documents that contains a specific term.

3. Inverse document frequency (*idf*)
    - The inverse document frequency is a factor that diminishes the weight of terms that occur often, such as the term ”the”.
4. Why *idf* is important for term weighting
    - *idf* is important because it introduces less weight to the less important terms, such as ”the” and more weight to important terms such as ”onomatopoeia”. This is done to get a more correct ranking of documents

Task 3

Given the following document collection containing words from the set O = Big, Cat,Small, Dog , answer the questions in subtasks 3.1 and 3.2

$$
\begin{aligned}
doc1 &= \{Big \quad Cat \quad Small \quad Dog\} \\
doc2 &= \{Dog\} \\
doc3 &= \{Cat \quad Dog\} \\
doc4 &= \{Big \quad Cat \quad Big \quad Small \quad Cat \quad Dog\} \\
doc5 &= \{Big \quad Small\} \\
doc6 &= \{Small \quad Cat \quad Dog \quad Big\} \\
doc7 &= \{Big \quad Big \quad Big\} \\
doc8 &= \{Dog \quad Cat \quad Cat\} \\
doc9 &= \{Cat \quad Small\} \\
doc10 &= \{Small \quad Small \quad Big \quad Dog\}
\end{aligned}
$$

Subtask 3.1

Given the following queries:

$$q1 = "Cat \quad AND \quad Dog"$$

$$q2 = "Cat \quad AND \quad Small"$$

$$q3 = Dog \quad OR \quad Big"$$

$$q4 = Dog \quad NOT \quad Small"$$

$$q5 = "Cat"$$

1

Which of the documents will be returned as the result for the above queries using the Boolean model? Explain your answers and draw a figure to illustrate.

q/doc	1	2	3	4	5	6	7	8	9	10
1	Yes	No	Yes	Yes	No	Yes	No	Yes	No	No
2	Yes	No	No	Yes	No	Yes	No	No	Yes	No
3	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
4	No	Yes	Yes	No	No	No	No	Yes	No	No
5	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No

For query 1, I have checked if the documents contains both cat and dog, the ones that do get's retruned. The same for query 2. For query 3 I have returned all the documents that contains either dog or big. For query 4 the ones that has dog and not small and for query 5 all the documents that contains cat.

2

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

Each dimension represents a term. Since there are four terms in the collection vocabulary the dimension is 4.

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.)

term/doc	1	2	3	4	5	6	7	8	9	10
Big	1	0	0	2	1	1	2.58	0	0	1
Cat	1	0	1	2	0	1	0	2	1	0
Small	1	0	0	1	1	1	0	0	1	2
Dog	1	1	1	1	0	1	0	1	0	1

Term frequency

Term	Ni	$\log_2 \frac{N}{n_i}$
Big	6	0.74
Cat	6	0.74
Small	6	0.74
Dog	7	0.51

Inverse Document frequency

Term/doc	1	2	3	4	5	6	7	8	9	10
Big	0.74	0	0	0	0.74	0.74	1.9	0	0	0.74
Cat	0.74	0	0.74	1.48	0	0.74	0	1.48	0.74	0
Small	0.74	0	0	0	0.74	0.74	0	0	0.74	1.48
Dog	0.51	0.51	0.51	0.51	0	0.51	0	0.51	0	0.51

$(1 + \log_2 tf_{i,j}) * \log_2 \frac{N}{n_i}$

4

Study the documents 2, 3, 5 and 7 and compare them to document 9. Calculate the similarity between document 9 and these four documents according to Euclidean distance. (Use tf-idf weights for your computations).

The Euclidean distance is given by:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + (p_5 - q_5)^2}$$

Using the weights I get:

$$d(9,2) = \sqrt{(0-0)^2 + (0.74-0)^2 + (0.74-0)^2 + (0-0.51)^2} = \sqrt{0 + 0.5476 + 0.5476 + 0.2601} = 1.1641$$
$$d(9,3) = \sqrt{(0-0)^2 + (0.74-0.74)^2 + (0.74-0)^2 + (0-0.51)^2} = \sqrt{0 + 0 + 0.5476 + 0.2601} = 0.8987$$
$$d(9,5) = \sqrt{(0-0.74)^2 + (0.74-0)^2 + (0.74-0.74)^2 + (0-0)^2} = \sqrt{0.5476 + 0.5476 + 0 + 0} = 1.046$$
$$d(9,7) = \sqrt{(0-1.9)^2 + (0.74-0)^2 + (0.74-0)^2 + (0-0)^2} = \sqrt{0.5476 + 0.5476 + 0 + 0} = 2.169$$

5

For the vocabulary v = {Big Small Dog Cat} i get:

$$d_1 = \{1,1,1,1\}$$
$$q = \{0,0,0,1\}$$

I'll just show the computation for d_1, then I'll summarize the rest in a table:

$$sim(d_1,q) = \frac{(1*0) + (1*0) + (1*0^{)} + (1*1)}{\sqrt{1^2+1^2+1^1+1^1} * \sqrt{0^2+0^2+0^2+1^2}} = \frac{1}{2}$$

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Dco10
$Sim(doc,q)$	0.5	0	0.707	0.63	0	0.5	0	0.707	0.707	0

When ranked, we get: [Doc3, Doc8, Doc9] [Doc4] [Doc1, Doc6] [Doc2, Doc5, Doc7, Doc10]

Subtask 3.2

Given the following queries: q1 = ”Cat Dog”, q2 = ”Small”

1

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

The main difference is that the Robertson-Jones model ws originally a framework for future models and that it does not have any weighted index terms like BM25. Also, it has no accurate estimate for the first probabilities.

2

I’ll use the formula

$$RSV_d = \sum_{t \in q} \log_2 [\frac{N}{dt_f}] \cdot \frac{(k1+1) \cdot tf_{td}}{k1((1-b)+b(\frac{L_d}{L_{ave}}))+tf_{td}}$$

for my calculations. I will show calculation for only one document and sumarize the rest in a table:

Generally in every calculation, I have:

$$L_{ave} = 3.1$$
$$N = 10$$
$$dt_{cat} = 6$$
$$dt_{dog} = 7$$
$$dt_{small} = 6$$

for document 1 and query 1 I have:

$$L_d = 4$$
$$tf_{cat} = 1$$
$$tf_{dog} = 1$$
$$tf_{small} = 1$$

which gives me:

$$\begin{aligned}
RSV_d &= \sum_{t \in q} \log_2 [\frac{N}{dt_f}] \cdot \frac{(k1+1) \cdot tf_{td}}{k1((1-b)+b(\frac{L_d}{L_{ave}}))+tf_{td}} \\
&= \log_2 [\frac{N}{dt_{cat}}] \cdot \frac{(k1+1) \cdot tf_{cat}}{k1((1-b)+b(\frac{L_d}{L_{ave}}))+tf_{cat}} + \log_2 [\frac{N}{dt_{dog}}] \cdot \frac{(k1+1) \cdot tf_{dog}}{k1((1-b)+b(\frac{L_d}{L_{ave}}))+tf_{dog}} \\
&= \log_2 [\frac{10}{6}] \cdot \frac{(1.2+1) \cdot 1}{1.2((1-0.75)+0.75(\frac{4}{3.1}))+1} + \log_2 [\frac{10}{7}] \cdot \frac{(1.2+1) \cdot 1}{1.2((1-0.75)+0.75(\frac{4}{3.1}))+1} \\
&= (\log_2 \frac{10}{6} + \log_2 \frac{10}{7}) \cdot \frac{682}{763} \\
&\approx 1.12
\end{aligned}$$

Below is a table summarizing:

q/doc	1	2	3	4	5	6	7	8	9	10
1	1.1186	0.7118	1.4640	1.1744	0.0	1.1186	0.0	1.5440	0.8621	0.4599
2	0.6587	0.0	0.0	0.5329	0.8621	0.6587	0.0	0.0	0.8621	0.9368

The ranking is now:

For Query 1: Doc8 Doc3 Doc4 [Doc1 Doc6] Doc9 Doc2 Doc10 [Doc5 Doc7]

For Query 2: Doc10 [Doc5 Doc9] [Doc1 Doc6] Doc4 [Doc2 Doc3 Doc7 Doc8]