TDT4117 Information Retrieval - Autumn 2021
Assignment 1 - Solution

September 12, 2021

Task 1 : Basic Definitions

Explain the main differences between:

1. Information Retrieval vs Data Retrieval:
   Answer: Information retrieval is about retrieval of semi- or unstructured data, which can be semantically ambiguous. Data retrieval involves the selection of a fixed set of data based on a well-defined query (e.g SQL, OQL).

2. Structured Data vs Unstructured Data:
   Answer: Unstructured data does not have clear, easy-for-a-computer structure. It is the opposite, structured data tends to refer to information in "tables".

Task 2: Term Weighting

Distinguish the concepts of:

1. Term Frequency (tf):
   Answer: $tf_{t,d}$ is the schema of assigning weight to each term t in the document d. The simplest approach is the number of occurrences of term t in document d.

2. Document Frequency (df):
   Answer: df_t is the number of documents in the collection that contain a term t.

3. Inverse Document Frequency (idf):
   Answer: idf is defined as :
   $$\text{idf}_t = \log \frac{N}{df_t}.$$
   It means that the idf of a rare term is high, whereas the idf of a frequent term is likely to be low.

Task 3: IR Models

Assuming the following document collection, which contains only the words from the set $O = \{Big, Cat, Dog, Small\}$.

```
doc1  = {Big Cat Small Dog}
doc2  = {Dog}
doc3  = {Cat Dog}
doc4  = {Big Cat Big Small Cat Dog }
doc5  = {Big Small}
doc6  = {Small Cat Dog Big}
doc7  = {Big Big Big}
doc8  = {Dog Cat Cat }
doc9  = {Cat Small}
doc10 = {Small Small Big Dog}
```

SubTask 3.1: Boolean Model and Vector Space Model

Given the following queries:

```
q1 = "Cat AND Dog"
q2 = "Cat AND Small"
q3 = "Dog OR Big"
q4 = "Dog NOT 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.

   Answer:

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
q_1	R	-	R	R	-	R	-	R	-	-
q_2	R	-	-	R	-	R	-	-	R	-
q_3	R	R	R	R	R	R	R	R	-	R
q_4	-	R	R	-	-	-	-	R	-	-
q_5	R	-	R	R	-	R	-	R	R	-

2. What is the dimension of the vector space representing this document collection when you use the vector model and how is it obtained?
   Answer: Each dimension represents a term. Here we have four terms in the collection vocabulary and consequently 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.)
   Answer: TF-IDF weighting is the **product** of tf and idf values, where each is calculated as follows.

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{idf}_t = \log \frac{N}{df_t}.$$

	idf_i	$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	$tf_{i,4}$	$tf_{i,5}$	$tf_{i,6}$	$tf_{i,7}$	$tf_{i,8}$	$tf_{i,9}$	$tf_{i,10}$
"cat"	0.737	1.0	0	1.0	2.0	0	1.0	0	2.0	1.0	0
"dog"	0.515	1.0	1.0	1.0	1.0	0	1.0	0	1.0	0	1.0
"big"	0.737	1.0	0	0	2.0	1.0	1.0	3.0	0	0	1.0
"small"	0.737	1.0	0	0	1.0	1.0	1.0	0	0	1.0	2.0

Table 1: TF and IDF values for each term of vocabulary.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
"cat"	0.737	0	0.737	1.474	0	0.737	0	1.474	0.737	0
"dog"	0.515	0.515	0.515	0.515	0	0.515	0	0.515	0	0.515
"big"	0.737	0	0	1.474	0.737	0.737	1.905	0	0	0.737
"small"	0.737	0	0	0.737	0.737	0.737	0	0	0.737	1.474

Table 2: TF-IDF weights.

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 and idf for your computation).
   Answer: if $q = (q_1, q_2, ..., q_n)$ and $d = (d_1, d_2, ..., d_n)$ are two points in Euclidean n-space, the euclidean distance is:

$$dis(q, d) = \sqrt{(q_1 - d_1)^2 + (q_2 - d_2)^2 + ... + (q_n - d_n)^2}.$$

   To calculate the euclidean distance between two documents, each document can be represented by boolean values, term frequency or tf-idf weights. For instance, document d_4 can be represented as $d_4 = (2, 1, 2, 1)$, where the values refer to frequency of terms *big, small, cat* and *dog* in document d_9, respectively.

	d_2	d_3	d_5	d_7
dis_{d_9,d_i}	1.163	0899	1.04	1.56

Table 3: Euclidean distance (tf.idf is used as weights).

5. Rank the documents by their relevance to the query $q5$ (use cosine similarity to calculate the similarity scores).
Answer: The similarity between two documents d_1 and d_2 is the cosine similarity of their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|},$$

where the numerator represents the dot product ($\vec{x} \cdot \vec{y} = \sum_{i=1}^{M} x_i y_i$) and the denominator is the product of their Euclidean lengths.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim_{q_5,d_i}	0.535	0	0.819	0.649	0	0.535	0	0.94	0.707	0

Table 4: Cosine similarity (using tf-idf).

Ranking(tf-idf) : $d_8, d_3, d_9, d_4, d_1, d_6, d_2, d_5, d_7, d_{10}, d_{10}$.

SubTask 3.2: Probabilistic Models

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?
Answer: BM25 can be computed without any relevance information provided by the user. In contrary, the classic probabilistic model assumes that there is a subset of all documents that the user prefers as the answer set to the query q. This assumption is troublesome, because relevance might be affected by variables *outside* the system.

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.
Answer:

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

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.

An alternative version of BM25 model is:

$$RSV_d = \sum_{t \in q} \log \frac{N - df_t + \frac{1}{2}}{df_t + \frac{1}{2}} \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{ave})) + tf_{td}}.$$

This variant behaves slightly strangely: if a term occurs in over half the documents in the collection then this model gives a negative term weight, which is undesirable. But, assuming the removal of stop words, this normally doesn't happen, and the value for each summand can be given a floor of 0.

In this task, we used the first formula, with $k = 1.2$ and $b = 0.75$.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim_{q_1,d_i}	1.16	0.711	1.46	1.16	0	1.16	0	1.54	0.862	0.459
sim_{q_2,d_i}	0.658	0	0	0.533	0.862	0.658	0.0	0.0	0.862	0.936