TDT4117 Information Retrieval - Autumn 2019
Assignment 1
Deadline for delivery is 18.09.2019

September 4, 2019

Note: Use log base 2 for all your computations.

Task 1: Basic Definitions

Explain the main differences between:

1. Information Retrieval vs Data Retrieval
2. Structured Data vs Unstructured Data

Task 2: Term Weighting

Explain:

1. Term Frequency (tf)
2. Document Frequency (*df*)
3. Inverse Document Frequency (*idf*)
4. Why *idf* is important for term weighting

Task 3: IR Models

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.

```
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 and draw a figure to illustrate.
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 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).
5. Rank the documents for query $q5$ using cosine similarity.

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?
2. Rank the documents using the BM25 model. Set the parameters to $k = 1.2$ and $b = 0.75$. (Here we assume relevance information is not provided.)
Hint: To avoid getting negative numbers, you need to use $idf = \log\left[\frac{N}{df_t}\right]$ in the BM25 model.

Important notes

Please carefully read the following notes and consider them for the assignment delivery. Submissions that do not fulfill these requirements will not be assessed and should be submitted again.

1. The assignment must be delivered in **pdf format**. Other formats such as .docx and .txt are not allowed.
2. The assignment must be **typed**. Handwritten assignments are not accepted.
3. Final scores are **required**, but not sufficient. You need to explicitly write the details of your computations (with no redundancy).
4. You may work in groups of maximum 2 students.