Week 5:
Retrieval Models: Vector Space, Probabilistic and Language Model

- This assignment is due on **18th November, 2015 (13:30)**
- You can discuss the problems with other groups of this course or browse the Internet to get help. However, copy and paste is cheating.
- There are 13 weekly exercises in total. In each one of them, all assignments sum up to 20 points. You need to achieve at least 80% of all assignments during the course in order to participate in the final exam. Hence, you need to achieve at least 208 points in total (*13*20*0.8=208*).
- Submission at
  https://www.dcl.hpi.uni-potsdam.de/submit
  - only pdf files
  - one file per group per week (week5.pdf)
  - put your names on *each* sheet in the pdf file

Assignment 1: Vector Space Model

a) Can the tf-idf weight of a term in a document exceed 1? **1 P**

b) What is the purpose of normalizing a documents vector representation for document length? **2 P**

c) If each term represents a dimension in a t-dimensional space, the vector space model is making an assumption that the terms are *orthogonal*. Explain this assumption and discuss whether you think it is reasonable. Why do we normalize the vector representation of documents in the vector space model? Is it always a good idea? **2 P**

Assignment 2: Probabilistic Model

a) What is 'binary' in the binary independence model (BIM)? **1 P**

b) What is 'independent' in the binary independence model (BIM) and is this a reasonable assumption? Explain. **2 P**

c) What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance information is available? **2 P**

d) What is parameter b good for in the BM25 model? **2 P**

Assignment 3: Comparing Models

Given the following document collection:

doc ID	document text
1	click click test click
2	click click
3	foo bar
4	click here foo bar test

Build a **query likelihood model** using maximum likelihood estimates, a **BM25 model** and a **tf-idf model**. Use Jelinek-Mercer smoothing with $\lambda = 0.2$ for the query likelihood model. For BM25 assume that there is no relevance information and that $k1$=1.2, $k2$=100 and b=0.75. Compute the ranking of the four documents for the queries

a) click **2 P**

b) test **2 P**

Assignment 4: (Programming) Different Models for Ranked Retrieval

This week we will implement one of the three above-mentioned retrieval models to rank our search results.

- You should implement a vector space, or probabilistic or language model. The goal is to get good results for our patent corpus. Therefore you need to change the implementation of your `ArrayList<String> search(String query, int topK, int prf)` function for the queries that do not contain ″AND″, ″OR″, ″*″, or ″NOT″.
- Set the variable ′topK′ to be ″10″ and limit the results to the first K patents. Ignore the last parameter of the `search` method for now.

a) Compute the ranking of the patents that match the following queries. Print their invention titles and document numbers.

- ″processing″ **1 P**
- ″computers″ **1 P**
- ″′mobile devices′″ **1 P**
- ″data″ **1 P**

b) Include the type of the model that you chose to implement in your pdf file.

c) Print your dictionary in your pdf file.