Week 6:
Querying

- This assignment is due on **25th 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 one zip file
  - one zip file per group per week (week6.zip)
  - put your names on *each* sheet in your pdf file

Assignment 1: Query Transformation using User Feedback

a) When does relevance feedback not work? **3 P**

b) What is the idea underlying Rocchios algorithm for pseudo relevance feedback? Do not use any formulas. **3 P**

c) What are the main advantages and disadvantages of pseudo relevance feedback? Use your own words. **3 P**

Assignment 2: Spelling Correction

In case of a misspelled query term, string comparison can be used to obtain the similarity between words and suggest to the user a possible corrent term for the misspelled one. One can calculate the distance between the incorrectly spelled word and some potential correct ones and then propose to the user the best candidates.

a) Compute the following measures for the terms ”data” and ”unit”: **2 P**

- Mutual information, Expected mutual information , Chi-square and Dice’s coefficient
- Consider that $n_a = 1{,}000$, $n_b = 3{,}000$, $n_{a,b} = 2{,}000$ and $N = 15{,}000$.

Assignment 3: Query Expansion using Synonyms

Another method for improving a search engine’s results is to exploit the synonyms of the query terms. One can expand the query with the query terms’ synonyms and attempt to retrieve more relevant documents to the original query. This technique can be helpful in order to retrieve documents with similar topic to the original query, but different vocabulary terms.

a) Which are the pros and cons of using synonyms for query expansion? **3 P**

b) In general, lexicons can be used to find synonyms of a given word. Are you aware of any publicly available lexicons? Are they domain dependent or general lexicons? **3 P**

Assignment 4: (Programming) Querying

This week we will implement pseudo-relevance feedback and as a bonus task, a spelling correction mechanism.

- The straight-forward way to implement a pseudo-relevance algorithm is to use the most frequent terms from the top-x documents in your result set, expand the original query with these terms and execute the new query. The goal is to retrieve more relevant documents to the user's information need with our expanded query. You are welcome to decide the technique you prefer to implement.
- There is a 'int' variable called prf which can be set to the number of top patents that should be included in the pseudo-relevance feedback. A value of '0' means no relevance feedback. Check the value of prf and if it is higher than zero, then compute your ranking using pseudo-relevance feedback including the top-prf documents. For answering the below queries, set the value of prf to be 2.

a) Explain your pseudo-relevance feedback method briefly in your pdf file.

b) Execute the following queries with and without using pseudo-relevance feedback. Print the titles and the document numbers of the patents returned for the following queries. Use two different txt files for each case and name them as follows: prfAnswers.txt, answers.txt .

- "digital" **1 P**
- "rootkits" **1 P**
- "network access" **1 P**

Assignment 5: (Programming) Spelling Correction : Bonus Task

As a bonus task for this week, you can implement a spelling correction mechanism.

- This week you can choose a spelling correction mechanism that you prefer and incorporate it as a new feature in your patent engine. For this pursuse, you could use the frequencies of the terms in the patent data or their frequencies in the AOL query logs `http://www.cim.mcgill.ca/~dudek/206/Logs/AOL-user-ct-collection/`. In case you choose the latter, you should be aware that there might not be many patent related terms, since this is not a patent search engine.
- In general, if a given query term does not appear in your dictionary, you could assume that it is misspelled and try to find potential correct query terms for it. You are welcome to implement any method you prefer to identify the candidate terms. You are also welcome to assume that the first letter in the misspelled query term is correct and thus only compare it to the terms in the dictionary that begin with the same letter.
- In addition, the presentation of the search results is essential, especially when you correct misspelled query terms. You should let the user know about the new correctly spelled query that you executed instead of the original one.

a) Explain in a few words in your pdf file how your spelling correction method works.

b) Execute the below queries without enabling your pseudo-relevance feedback mechanism. As in the above assignment, write the titles and the document numbers of the patents returned for the below queries in a txt file, which is now called bonus.txt. Write the correctly spelled queries in a second txt file called correctedQueries.txt.

- "commom" **1 P**
- "kontrol" **1 P**
- "incluce" **1 P**
- "streem" **1 P**