Week 11:
Web Search

- The theoretical assignments are due on **13th January, 2015 (13:30)**
- The programming assignment is due on **20th January, 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 zip files
  - one file per group per week (week11.zip)
  - put your names on *each* sheet in the pdf file

Assignment 1: Web Retrieval vs. Classical IR

a) What are the main differences between Web Retrieval and Classical IR? (2)

b) What are the types of queries you find in the Web? Give an example for each query type. (2)

c) What are reasonable business models for Web Search? Describe one scenarios where different business models would make sense. (3)

Assignment 2: Hidden Web

a) Name three characteristics of the deep/hidden web. (2)

b) Describe in detail how a search engine could index documents from a deep web source, such as a patent search engine. (6)

Assignment 3: (Programming) Web Search

This week we will scale up our implementation to run on the patent dataset spanning from 2011 to 2015.

- Generate your index using all xml files contained in the zip file (PatentData.zip) in the course's directory. The xml files consist only of utility US patents grants.
- Make sure that your index and seek list are working properly, and your search engine is still efficient when using only 1GB Ram memory. This is the memory limit we will be using from now on. If any memory problems occur, you could reconsider your compression techniques. Keep in mind that your search engine will be evaluated not only for its accuracy but also for its speed.

Print the ranking (along with your snippets) of the patents that match the following queries and the corresponding NDCG values. Use the value of `20` for *topK*.

a) ”responsive applications for android tablet” (1)

b) ″cloud computing security issues″ (1)

c) ″cloud NOT smart″ (1)

d) ″″3-D miniatures″″ (phrase query) (1)

e) ″healthcare AND services″ (1)