Week 1:
Introduction

- This assignment is due on **21st October, 2015 (15:00)**.
- You should work in groups of two.
- 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 (week1.pdf)
  - put your names and matriculation numbers on *each* sheet in the pdf file

Assignment 1: Group Formation

To get started you need to form groups of size two.

a) Get familiar with the submission system
`https://www.dcl.hpi.uni-potsdam.de/submit`. **0 P**

b) When you submit your first assignment be sure to include both names of the two members of your group in the author list. **1 P**

Assignment 2: Information Retrieval Introduction

a) In which aspects is Web search similar and different to information retrieval? **3 P**

b) What is an information need? And how does it differ from a database SQL statement? **3 P**

c) How is the relevance notion defined in information retrieval? And what makes a document relevant or not relevant to a particular query? **3 P**

Assignment 3: U.S Patent Retrieval

a) What are the main differences between Web search and Patent search? **4 P**

b) Why would someone search for patents? Considering different kinds of patent searchers, i.e. inventors, examiners, lawyers, scientists, etc. which are the main reasons for searching patent inventions? **4 P**

Assignment 4: (Programming) Getting Ready

During this course each group will implement their own search engine in Java. At the end of the course we will evaluate all search engines with respect to quality of search results as well as speed. We will build our search engines on the U.S. patent dataset. The patent data are in XML format. The programming assignments will guide your development and implementation process. Don't submit any source code for the assignments; just the output of your program as described in the task.

- Install your favorite Java programming environment, e.g. Eclipse, Netbeans.
- Download the Java source files from the course's folder. (`WS 2015-16/Excercises/Code`)
- For now, you will use a small test dataset containing only 20 "utility" patents to build and test your search engine. This file is called testData.xml and is included in the course's folder (`WS 2015-16/Excercises/Data`). It is ready for you to use it during the development of your search engine.
- In general, there are different types of patents, but we are only interested in the patents with application type "utility". That is why the testData file contains only this kind of patents. The application type is denoted in the *appl-type* attribute inside the *application-reference* element.
- Later on in the course we will use the whole U.S. patent dataset. Thus, you should download the data at `http://www.google.com/googlebooks/uspto-patents-grants-text.html` and filter the xml files, keeping only the patents with application type "utility". (you will need these filtered data later)
- You should only download the patent files of the form "ipg", that is the zip files from 2005 until 2015. Each file contains all the patent grants announced in the U.S. in a particular week of the year. For example, ipg140107.zip refers to the week of 2014 that includes 7th January 2014. In each zip file, you can find an XML file containing all the weekly patent grants.
- Get familiar with Java's SAX Parser `http://sax.sourceforge.net/` to parse XML streams (hint: Although we are currently only considering the patents in the test dataset, the overall dataset is too large for loading it in the main memory. Make sure to read and process the XML files sequentially.).
- Implement a Java method using the SAX parser to output the patent invention titles with their document numbers. The invention titles are included in the *invention-title* element and the document numbers in the *doc-number* element inside the *publication-reference* element (hint: there is a second *doc-number* element included in the *application-reference* element, which is not the one we will use).

a) Print the invention titles along with the document numbers contained in the development XML file. **2 P**