Assignment 1: Crawling

- This assignment is due on **7th May, 2019 (11:00 am, CET)**
- You can discuss the problems with other groups of this course or browse the Internet to get help. However, copy and paste is cheating. Code plagiarism is a severe offense!
- There are 8 assignments in total. In each one of them, all tasks sum up to 20 points. You need to achieve at least 70% of the points in 7 assignments and at least 50% in the remaining one in order to participate in the final evaluation.
- Submission via studip
  - only pdf files
  - one file per group per assignment (groupName-assignment1.pdf)
  - put your names and matriculation numbers on *each* page in the pdf file

Task 1: Setup

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 and memory consumption. You will build your own search engine for newspaper articles. 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 description.

We provide a small test dataset containing only 10 newspaper articles to build and test your search engine. This file is called testData.csv and is included in the course's folder. Later on in the course, you will use your own crawled article dataset. Thus, you should finish your crawler's code as soon as possible so that you can focus on the other parts of your search engine.

a) Download the testData.csv and print meta-data of the first article. **10 P**

Task 2: Crawling

- Download the Java source files from the course's folder and have a look at the example scraper (`/Excercises/Assignment_1`).
- Store the data in a simple csv format with
  `article_id, article_url, article_authors, article_text, article_headline, publication_timestamp, article_categories`
- Implement a Java method using the provided template that crawls the newspaper articles for a given date. The method should return a csv file with the beforementioned structure.

a) Print the csv file for ten newspaper articles from May 3rd, 2019. **10 P**

Hints for Crawling:

- Find article URLs, visit each URL only once
- Some articles are distributed across multiple pages
- Regular recrawling: which pages have changed?
- User-agent: "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36"
- Referrer URL: http://facebook.com/l.php?u=[insert WSJ URL here]
- Javascript, dynamic content