Assignment 2: Transformation

- This assignment is due on **20th 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-assignment2.pdf)
  - put your names and matriculation numbers on *each* page in the pdf file

Task 1: Tokenization

Transform the news articles (headline, subtitle, captions, body) to make it ready for indexing. To start, you need to tokenize your text, i.e. identify individual words. Decide what to do with punctuation (remove or keep?) and whether to lowercase all words or not. You dont need to store the tokenized text; you will use it next week to build your index on the fly...

a) Print example sentences from your news articles: the original version and the tokenized version. **4 P**
   - One sentence with an abbreviation
   - One sentence with a person name

b) Briefly state why or why not you removed punctuation and why or why not you lowercased everthing. **6 P**

Task 2: Stemming

Use a library to stem your tokens. Again, no need to store the stems for now; we will use the stems next week to build our index.

a) Which stemming library/algorithm did you use? **2 P**

b) Find an example where the stemming algorithm makes a mistake or the result is at least problematic. **4 P**
   - Discuss why the mistake happens and how it could be prevented.

c) Print an example sentence from your corpus: The original version and the tokenized+stemmed version. **4 P**