Exercise in Information Retrieval (Master), Summer term 2016
Ioannis Karatassis (LE 324)
Office hours: by appointment
karatassis@is.inf.uni-due.de

Exercise sheet 6  No submission

Exercise notes: By presenting (enough) exercise tasks, it is possible to receive bonus points for the (oral) exam. Those tasks are marked with a (*). You should prepare some slides for the task you want to present. We expect students to participate at discussions and to contribute input – in the form of presentations or questions – to the exercise that allows a good discussion.

http://www.is.inf.uni-due.de/courses/irmai_ss16

Presentation date: 14/7/2016

Task 1: Constructing inverted lists with SPIMI algorithm

(a) Consider the documents below consisting of the given terms:
   
   $d_1$: A C F E J
   $d_2$: A C J H I K
   $d_3$: J K L M N
   $d_4$: A F J L
   $d_5$: C H M N

   The main memory of a computer is limited in this way that it can only keep up to 10 posting at the same time (additionally management informations). Perform the SPIMI algorithm and state the blocks and dictionaries as intermediate steps.

(b) Now, consider a computer that can process 8 threads at the same time. Discuss changes that need to be applied to the SPIMI algorithm exploiting parallelism.
Task 2: Comparison between Web search and classic Information Retrieval

Compared to classic IR, Web search faces particular challenges. Essential differences are listed below:

- Massive amount of data
- Heterogeneous data (HTML, Images, Video, News, PDF, ...)
- Data is unstructured in general (no further structuring or semantic known)
- Quality of data is questionable
- Data is distributed across many Web servers
- Data can change quickly (e.g. Web sites of news papers or news magazines)

Clarify for each item respectively, why it poses a notable challenge and specify components of Web search engines that are affected.