D5: Databases and Information Systems
Advanced Topics in IR, SS 2016
Dr. Vinay Setty and Dr. Jannik Strötgen
Assignment 6, due: **14 Jul 2016**

max planck institut informatik

Please submit your solution as a PDF and source code together in a single tar/zip file, name your pdf file in the following format: "Lastname-Immatrikulationnumber-AssigmentX.pdf", also include Immatrikulationnumber in the title of your document and email it to `atir16@mpi-inf.mpg.de` *before the due date mentioned above!*

Opinion Retrieval (20 points)

Problem 1.
Refer to the papers by He et al. [4] and Huang and Croft [5]. Think about the following two questions regarding their approaches. Your answers should provide sufficient detail and be justified (i.e., just saying "[5] is more efficient" is not good enough).

(a) *Which one is more efficient?* To answer this question, think about (a) which data structures you would use for the implementation (b) which statistics can be precomputed (once or periodically) or have to be computed at query-processing time.

(b) *Which one is more effective?* To answer this question, compare their experimental setups and see whether the reported results are comparable at all. Are they reproducible? Do you trust the results that are reported?

Feed Distillation (20 points)

Problem 2.
Consider the blogger model and posting model developed by Weerkamp et al. [11].

(a) At first glance, the difference between the two is all but obvious. Take a second look and see whether you can illustrate by means of an example how they differ. That is, you should come up with a small example (consisting of two or more blogs and posts therein) and a high-level topic (i.e., query) and show that the returned result differs. You may ignore smoothing for this problem.

Hint: Have a look at Table 7 in Weerkamp et al. [11].

(b) One aspect neither of the two models captures is whether a blog posts about the topic of interest over an extended period of time. To see this, note that the order of posts in a blog does not matter, so that a blog that posts regularly on-topic is considered as good as a blog that published (equally relevant) on-topic posts long ago. Extend one of the two models so that this aspect is taken into account. You may assume that every post p comes with a publication timestamp t at day granularity.

Top-Story Identification (20 points)

Problem 3. Consider the top-story identification approach described in the slides (a simplified version of the approach by Lee and Lee [11]). Sometimes, it can be useful to identify news stories that are not only *important*, as shown by intensive coverage in the blogosphere, but also *controversial*, as indicated by different blog posts expressing very different opinions about the story. Assuming that you have a lexical resource (e.g., SentiWordNet available at `http://sentiwordnet.isti.cnr.it`) from which you can find out positive, objective, and negative words. Can you come with an approach that identifies the most controversial news articles published around a given day d?

Representative Terms (Programming Assignment)(40 points)

Problem 4. By now you should be familiar with the TREC sample corpus we provided.
Apply the methods used by He et al. [4] (Kullback-Leibler divergence and Bose Einstein statistics) to extract words that are indicative of the top articles from the TREC corpus for a given query. For the queries below first retrieve top-1000 documents, then compute the KL divergence and Bose Einstein statistics for each term in these top-10 documents using the top-1000 documents as the background model. For both measures, please submit the list of top-20 most indicative terms. Compare the indicative terms from both approaches. Do you see any differences?

1. Hubble Telescope Achievements
2. African Civilian Deaths
3. Implant Dentistry
4. Radio Waves and Brain Cancer
5. Alzheimer's Drug Treatment