D5: Databases and Information Systems
Advanced Topics in Information in IR, WS 2014/15
Dr. Klaus Berberich
Assignment 1, due: **13 Nov 2014**

max planck institut informatik

Please submit your solution as a PDF to `atir2014@mpi-inf.mpg.de` *by the indicated due date!*

Language Models and Smoothing

Problem 1. The first lecture included a quick recap of language models and smoothing methods. The likelihood of generating the query q from the language model θ_d for document d is

$$P[\, q \mid \theta_d \,] = \prod_{v \in q} P[\, v \mid \theta_d \,]^{\, tf(v,q)}$$

where the probability $P[\, v \mid \theta_d \,]$ can be estimated without smoothing as

$$P[\, v \mid \theta_d \,] = \frac{tf(v,d)}{\sum_w tf(w,d)} \; ,$$

using Jelinek-Mercer smoothing with the document collection D as

$$P[\, v \mid \theta_d \,] = \lambda \cdot \frac{tf(v,d)}{\sum_w tf(w,d)} + (1-\lambda) \cdot \frac{tf(v,D)}{\sum_w tf(w,D)} \; ,$$

or using Dirichlet smoothing as

$$P[\, v \mid \theta_d \,] = \frac{tf(v,d) + \mu \cdot \frac{tf(v,D)}{\sum_w tf(w,D)}}{(\sum_w tf(w,d)) + \mu} \; .$$

(a) Smoothing introduces a relative weighting of query terms similar to the inverse document frequency in tf.idf term weighting. Devise a **minimal example** (at most three documents) to demonstrate this effect.

(b) Show that making the parameter λ in Jelinek-Mercer smoothing depend on the document length, as in the posting model by Weerkamp et al. [11], makes it equivalent to using Dirichlet smoothing.

Effectiveness Measures

Problem 2. The first lecture revisited the notions of precision and recall. We also saw that tasks in Information Retrieval can often be categorized as *precision-oriented* (typical example: web search) and *recall-oriented* (typical example: prior art search in patent retrieval).

(a) Think about other examples of precision-oriented and recall-oriented tasks in Information Retrieval. Describe at least one example per category.

(b) Revisit the definitions of mean average precision (MAP) and mean reciprocal rank (MRR) as two other common effectiveness measures. Do they measure precision (as in the system's ability to return only relevant results), recall (as in the system's ability to return all relevant results), or both?

(c) When was the last time Google (or your favorite web search engine) let you down? That is, it did not satisfy your information need or you had to invest serious effort (e.g., reformulate the query multiple times, combine insights from multiple queries, etc.). Describe the situation (only if it's not too personal of course). How could the web search engine automatically address this in a better way?

Opinion Retrieval

Problem 3.
Read 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

Problem 4.
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

Problem 5. 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?

Data Analysis (Programming Assignment)

Problem 6. Below are two data analysis tasks related to what we discussed in the lecture. Please pick **one of them** and work on it using your favorite programming language (e.g., Java, Python, R). In the end, we're not interested in your code but only in the results of your analysis. These should be included in the solution that you submit using a suitable form of presentation (e.g., list, table, plot).

(a) **Twitter**: We provide a sample of 100K tweets by U.S. users on the course website (`twitter-sample.json.gz`). Details about the format can be found at: `https://dev.twitter.com/overview/api/tweets`. Answer at least the following questions; we're happy to see answers to your own questions.

  (i) What does the distribution of tweet lengths look like?

  (ii) Distinguish between hashtags (starting with `#`), users (starting with `@`), URLs (starting with `http:`), and words (the rest). For each of them, what does the frequency distribution look like? Does Zipf's law hold?

(b) **Amazon Product Reviews**: We provide a sample of 463K video-game reviews on the course website (`amazon-video-games.txt.gz`). Apply the methods used by He et al. [4] (Kullback-Leibler divergence and Bose Einstein statistics) to extract words that are indicative of positive reviews. Positive reviews have a score of 4.0 or 5.0. Perform the same analysis for helpfulness. Helpful reviews have been marked as such by more than 50% of users. For both measures and aspects, please submit the list of top-25 most indicative terms. Compare the indicative terms for positive and helpful reviews. Do you see any correlation?