D5: Databases and Information Systems
Advanced Topics in Information in IR, WS 2014/15
Dr. Klaus Berberich
Assignment 6, due: **6 Feb 2015**

max planck institut informatik

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

Interesting Phrase Mining

Problem 1.
Bedathur et al. [2] describe different index representations and query-processing methods to identify interesting phrases in an ad-hoc document set D' (e.g., a query result) relative to the entire document collection D. Their measure of interestingness, though, is rather simplistic.

a) Pointwise mutual information (PMI) is an another measure of interestingness from the literature. Adapted to our setting, for a phrase $p = \langle p_1, \ldots, p_{|p|} \rangle$, it can be defined as

$$PMI(p, D') = \frac{df(p, D')/|D'|}{\prod_i (df(p_i, D)/|D|)} \; .$$

b) Assume that you want to modify the definition of interestingness to take into account the relevance of documents in D'. Let $w(q, d) \in [0, 1]$ denote a weight that indicates how relevant d is to the query q. The interestingness of a phrase can then be defined as

$$I(p, D') = \frac{\sum_{d \in D'} w(q, d) \cdot \mathbb{1}(p \in d)}{df(p, D)}$$

with $\mathbb{1}(p \in d)$ indicating whether document d contains the phrase p.

Please discuss for both modified notions of interestingness how you would adapt the different index representations and query-processing methods to support them. In particular, can you still terminate query processing early when using the frequency-ordered phrases approach?

Document Dating (Programming Assignment)

Problem 2.
On the course website you can download all articles published by The New York Times in May 1987, 1997, and 2007 (`198705.tgz`, `199705.tgz`, `200705.tgz`). In this assignment, we want to evaluate how well a document dating method based on language models works.

a) Parse the documents (same instructions as for Assignment 04).

b) Randomly split the documents from each year into 50% training and 50% testing data.

c) Estimate a unigram language model on the training portion from each year.

d) Date documents in the testing data, i.e., assign them to the year whose language model has the lowest Kullback-Leibler divergence from the document language model.

Please report a 3×3 confusion matrix as well as precision, recall, and F1 for each year.

Non-Traditional Effectiveness Measures & Rank Correlation

Problem 3.

a) Compute expected reciprocal rank for the following result with graded assessments $(0 \ldots 2)$

Rank	1	2	3	4	5
Grade	0	1	2	1	0

b) Compute rank-biased precision ($p = 0.5$) for the following result with binary assessments

Rank	1	2	3	4	5
Grade	1	0	0	1	1

c) Compute Kendall's τ between the following two permutations of the integers $1 \ldots 5$

$$\pi_1 = \langle\, 1, 3, 2, 4, 5 \,\rangle \quad \pi_2 = \langle\, 5, 3, 2, 4, 1 \,\rangle$$

Minimal Test Collections

Problem 4.
Consider the following results returned by two systems S_1 and S_2

$$S_1 = \langle\, d_3, d_1, d_5, d_9, d_7, d_2 \mid d_6, d_4, d_{10}, d_3 \,\rangle$$

$$S_2 = \langle\, d_4, d_2, d_9, d_5, d_{10}, d_6 \mid d_5, d_3, d_7, d_3 \,\rangle$$

We are interested in the relative order of the two systems when using P@6 as an effectiveness measure. Apply the minimal test collection method discussed in the lecture.

a) Which documents need to be judged at all?

b) Iteratively judge the documents in the order of their identifier using the following assessments

d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
1	0	1	1	1	0	1	0	1	0

At each step, determine the upper bound and the lower bound for $\Delta P@6(S_1, S_2)$ and terminate as early as possible.

Assessor Agreement

Problem 5.

a) Assume two assessors have judged 100 documents on a 4-grade scale (0 : *irrelevant*, 1 : *somewhat relevant*, 2 : *relevant*, 3 : *highly relevant*)

	0	**1**	**2**	**3**
0	10	0	3	2
1	5	20	10	5
2	0	2	10	
3	0	8	5	20

Compute Cohen's kappa for the two assessors.

b) Cohen's kappa is often criticized for its dependence on marginal probabilities. To illustrate this: Can you come up with two simple examples (two categories) where assessors agree on the category for an equal number of subjects, but Cohen's kappa values differ substantially?

c) Assume each of 10 documents has been judged by seven assessors on a 4-grade scale (0 : *irrelevant*, 1 : *somewhat relevant*, 2 : *relevant*, 3 : *highly relevant*)

	0	**1**	**2**	**3**
d_1	0	0	3	4
d_2	1	5	0	1
d_3	1	1	1	4
d_4	4	2	0	1
d_5	7	0	0	0
d_6	2	3	2	0
d_7	6	1	0	0
d_8	1	5	1	0
d_9	0	0	7	0
d_{10}	0	3	4	0

Compute Fleiss' kappa between the assessors.

First-Mention Detection

Problem 6. Assume you have a large-scale longitudinal document collection such as a newspaper archive available. Documents in the collection come with a publication date, i.e., you have reliable information about when the document was published. In this assignment, we are interested in efficiently determining when a phrase (e.g., a meme like "lipstick on a pig") was first mentioned in our document collection. We consider two scenarios. For each of them, please describe your approach in detail, i.e., which index structures and algorithms would you use.

a) You're given a dictionary P consisting of all phrases of interest. Phrases in this dictionary can have arbitrary length. For each of them, determine when it was first mentioned, i.e., the earliest publication date of any document that contains the phrase.

b) For any phrase p occurring in our document collection and consisting of up to σ words, determine when it was first mentioned in our document collection.