Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Week 5: Querying

Assignment 1: Query Transformation

a) When does relevance feedback not work?
Relevance feedback does not work in those cases when the relevance of the result set has to be determined by manually looking at the documents, e.g. because the provided snippets provide no valuable information.

b) What is the idea underlying Rocchios algorithm for pseudo relevance feedback? Do not use any formulas.
The result set for a query is split into a relevant and a non-relevant part. Then, the system will continuously refine the query using related relevant documents in order to find the largest distance between the two sets (represented by vectors).

c) What are the main advantages and disadvantages of pseudo relevance feedback? Use your own words.
In pseudo relevance feedback, an algorithm determines a certain number of relevant documents for a query. Advantage: This is done automatically; therefore no additional user input is needed. Disadvantage: The relevance of the final result will depend on the quality of the initial relevant documents, which are determined by the algorithm.

Assignment 2: Showing Results

Compare the result pages of Google and Bing search engines.
a) What elements are similar, different, the same?
For the query "information retrieval", Google and Bing list quite different results. Bing shows e.g. topic pages on Amazon, suggest different professional seminars on the topic, as well as pages from different encyclopaedias such as DUDEN. Google returns quite a few university course pages, as well as two Wikipedia links in the top results.

Bing puts possibly related queries in a more visible spot (both in the sidebar as well as after the first three results), while Google only shows them at the bottom of the page.

The actual results are displayed in a similar fashion, with a title, URL and snippet followed by additional information such as ratings as well as direct links to different sections of the page.

b) Find two different result snippets for the same page and say which one you like better and why.
Query: "andreas polze"
Result page: http://www.hpi.uni-potsdam.de/personen/professoren/polze.html

Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Google:
Andreas Polze entwickelt. Vor allem die Verbindung von Middleware und eingebetteten Systemen und deren vorhersagbares Verhalten in Bezug auf ...

Bing:
"Das Hasso-Plattner-Institut ist ein gutes Beispiel für Innovation im institutionellen Bereich, für Innovation in den Lern- und Arbeitsabläufen im Sinne einer ...

We prefer Google's result, as it contains more text from the actual content of the page, while Bing's snippet is taken from the sidebar, which is not relevant in the context of this query.

Assignment 3: (Programming) Querying

Print the ranking of Wikipedia titles with snippets in your development set using pseudo-relevance feedback for the queries:
a) "deutsch"

```
Document: Anschluss (Soziologie), Rank: -14.458057512383945
     für Philosophie.'' Heidelberg 1966, Fink Verlag, München 1967, S.
45–55.* Jürgen Frese: ''Prozesse im Handlungsfeld.'' Klaus Boer
Document: Ang Lee, Rank: -18.054271460346378
     Schüren, Marburg 2009.  erst spät ganz auf Filmregie und -
produktion – auch weil Lee seinen Berufswunsch seiner Familie und
insbesondere
Document: Actinium, Rank: -24.905664039181943
     radio-active“, in: ''Comptes rendus'', '''1899''', ''129'', S.
593–595 ( http://gallica.bnf.fr/ark:/12148/bpt6k3085b/f593.table
Document: Alan Smithee, Rank: -25.41242488889825
     anmutende Schreibweisen ''Alan Smi Thee'' und ''Sumishii Aran''
gehören – so die Internet Movie Database – dazu.
```

b) "artikel"

```
Document: Alan Smithee, Rank: 3.7781925428598115
     America (DGA) für solche Situationen empfohlen, seither ist es
'''Thomas Lee'''. Los Angeles Times latimes.com:
Document: Ang Lee, Rank: 3.326089205269883
     Lee - 66eme Festival de Venise (Mostra) 2.jpg miniatur Ang Lee
2009'''Ang Lee''' ({{zh c=李安 p=Lǐ Ān}}; * 23. Oktober 1954 in
```