Foundations of Information Retrieval

Midterm test from 19.12.2013

Allowed resources: A non-programmable calculator
Duration: 45 minutes
Good luck!

I. Web Information Retrieval

1. Given is the following document collection containing two documents:

[5(1a)+5(1b)=**10 P.**]

Document 1:
World War Z is a 2013 British-American film starring Brad Pitt.
Document 2:
Brad Pitt, an American actor and film producer, wildly varies his film choices.

1a. Create an inverted index for this document collection. **Tokenization rules**: word wise, case-folding, ignore punctuation. **Stop list***: is, a, an, and, to, his.* Include TF and DF values at a suitable position in the index.

1b. Specify search results can be obtained from this index for the following queries?

Q_1= *Brad Pitt*
Q_2= *American actor*

Compute the relevance scores for each query and search result using the following function:

$$w_{Q,d} = \sum_{q \in Q,d} \left(1 + \log\left(TF_{q,d}\right)\right) * IDF_q$$

Explain the results!

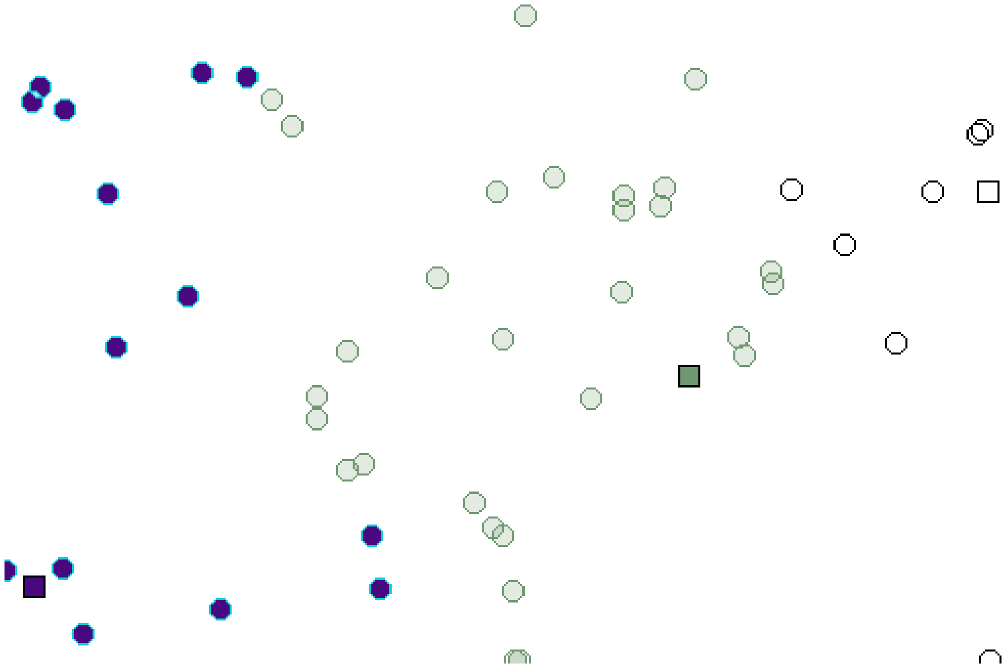
2. A collection of documents contains 10 documents that are relevant for a query *q*. For this query, the search engines S_1 and S_2 return the following relevant (R) and non-relevant (N) documents:

S_1:	NNNRR NNRRR
S_2:	RRNNN NNRRN

Draw a precision-recall diagram for the both search results and compare the quality of the search results based on the **interpolated precision at 20% recall**. [**10 P.**]

II. Text Classification and Clustering

1. The figure below describes a state of the k-means algorithm with *k*=3. In the figure, the squares represent the centroids and the circles represent the data points. The color encoding corresponds to the current cluster assignment. Describe the next step of the algorithm and sketch the changes that will be performed by the algorithm in this step. [**5 P.**]



2. Text Classification. Given is a Naive Bayes-Classifier with two classes C_1 and C_2 with: [**5 P.**]

$P(C_1)$	0.4
$P(C_2)$	0.5
P("*world*"|C_1)	0.7
P("*world*"|C_2)	0.6
$P(t_x\|C_1)$, $t_x \neq$" *world*"	0.1
$P(t_x\|C_2)$, $t_x \neq$" *world*"	0.1

Classify the following document using this classifier (computation required!):

D_1:
World War Z was chosen to open the 35th Moscow International Film Festival.

III. Query Optimization and Tolerant Retrieval

1. A document collection contains 15,000 documents. Given is the following query: [**5 P.**]

NOT Helsinki AND (Hannover OR Braunschweig) AND Moskow AND Saint Petersburg

Specify the most efficient order of execution for this query that can be determined from the following table:

Term	**DF**
Helsinki	13000
Hannover	1500
Braunschweig	2500
Moskow	3000
Saint Petersburg	2800

Describe a possible term distribution for which the order you proposed would not be optimal.

2. Name two data structures that enable processing of wildcard queries. For each of these structures, specify the queries for the wildcard query Con*dat* (Consolidation). Compare the both structures with respect to the precision of search results. [**5 P.**]

3. Given are two terms: “World” and “Wort”. Compute the Levenshtein distance and bi-gram based similarity between these terms. [**5 P.**]