Foundations of Information Retrieval

Midterm test from 20.12.2012

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

I. Web Information Retrieval

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

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

D_1:
Eclipse – Biss zum Abendrot ist ein Film, der auf dem Roman „Biss zum Abendbrot" von Stephenie Meyer basiert.
D_2:
Eclipse (von engl.: eclipse = Sonnenfinsternis, Finsternis, Verdunkelung) ist eine integrierte Entwicklungsumgebung.

Create an inverted index for this document collection. **Tokenization rules**: word wise, case-folding, ignore punctuation. **Stop list**: {der, die, das, dem, den, ein, eine}.
Include TF and DF values at a suitable position in the index.

1b. The function for the query-document relevance is:

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

Compute the relevance of the both documents for each of the following queries:

Q_1= *Die Eclipse IDE*
Q_2= *Abendbrot Rezepte*

Explain the resulting similarity scores!

2. A collection of documents contains 250 documents from which 10 are relevant for a given query. For this query, search engines S_1 and S_2 return the following lists (left -> right) of the relevant (R) and not relevant (N) search results:

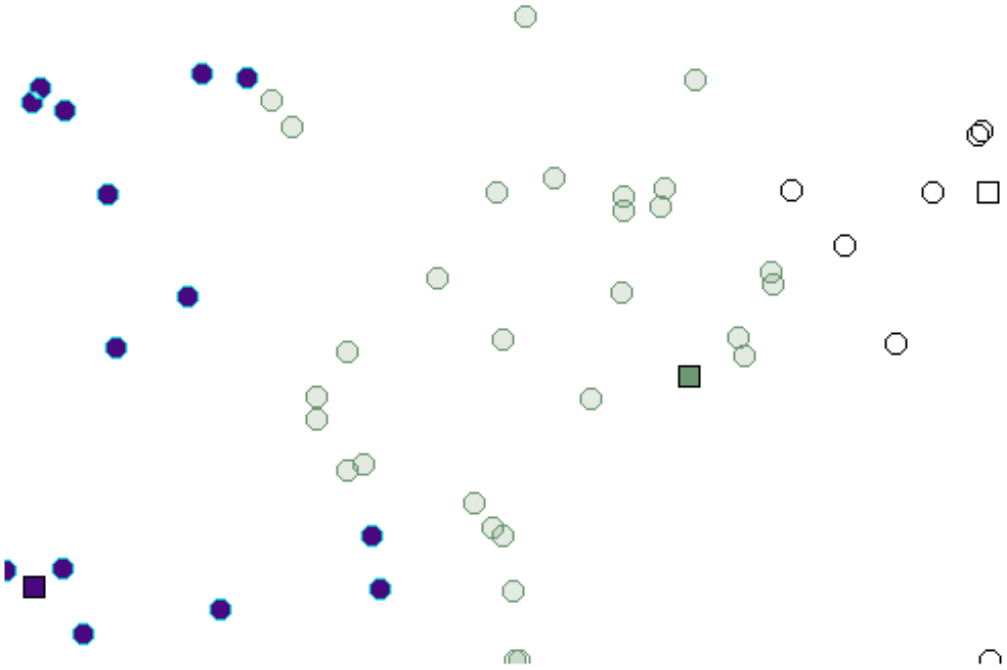
S_1: RRRNN RRNNN
S_2: NNRRR RRNNN

Create a precision-recall diagram for the both search results. Compare the quality of the both results based on the interpolated precision at 40% recall. [**10 P.**]

II. Text Classification and Clustering

1. The figure below describes an intermediate state of the *k*-means algorithm with *k*=3. In this 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 the following Naïve Bayes classifier with two classes C_1 and C_2 and the following parameters: **[5 P.]**

$P(C_1)$	0.3
$P(C_2)$	0.2
$P(t_y\|C_1)$	0.1
$P(\text{“ice”}\|C_2)$	0.4
$P(\text{“age”}\|C_2)$	0.2
$P(t_x\|C_2)$, $t_x \neq$ ”ice”, $t_x \neq$ ”age”	0.1

Classify the following document:

D_1:
Ice Age 4 – Voll verschoben ist ein amerikanischer computeranimierter Actionfilm.

III. Query Optimization and Tolerant Retrieval

1. A document collection with 50,000 documents contains weather forecasts. Given is the following query: **[5 P.]**

(spring) AND (sun OR wind) AND NOT (rain OR thunderstorm)

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

Term	**DF**
spring	25,000
sun	15,000
wind	30,000
rain	2,000
thunderstorm	11,000

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

2. Describe a trigram index structure. Given a wildcard query S*warzeneg*er (Schwarzenegger). For this query, create queries for a trigram index and a permuterm index. **[5 P.]**

3. Compute the Levenshtein distance and the bigram based similarity between the terms „**Lucky**“ and „**Duck**“. **[5 P.]**