Foundations of Information Retrieval

Refresher

Q&A session: 29.01.2015
Questions to: Helge Holzmann, holzmann@L3S.de

1. Information Retrieval

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

   D1:
   "World War Z is a 2013 British-American film starring Brad Pitt."

   D2:
   "Brad Pitt, an American actor and film producer, wildly varies his film choices."

   a. 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.
   b. Which search results can be obtained from this index for the following queries?

   Q1 = *Brad Pitt*
   Q2 = *American actor*

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

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

   Explain the results!

2. A collection of documents contains 10 documents that are relevant for a query q. For this query, the search engines S1 and S2 return the following relevant (R) and non-relevant (N) documents:

   S1: *NNNRR NNRRR*
   S2: *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.

2. Query Optimization and Tolerant Retrieval

1. A document collection with 50,000 documents contains weather forecasts. Given is the following query:

   (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	10,000

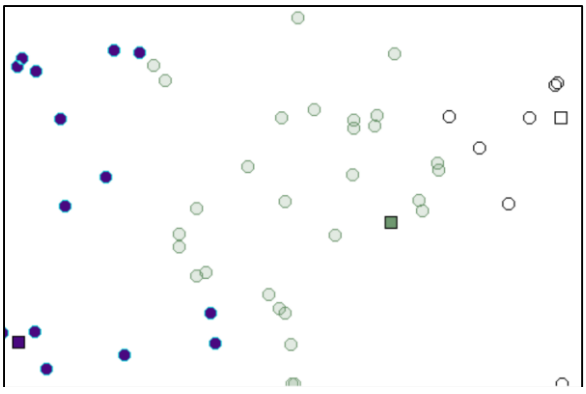
   Describe a possible term distribution for which the order you proposed is not optimal.

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

3. Compute the Levenshtein distance and the bigram based similarity between the terms '*Lucky*' and '*Duck*'.

3. Text Classification and Clustering

1. The figure below shows a state of the k-means algorithm with *k=3*. The squares represent centroids and circles represent the data points. The color encoding corresponds to the current cluster assignment.
   c. What phase of the algorithm has just finished and what phase is going to follow next?
   d. Sketch the changes that will be performed by the *k*-means algorithm in the next step.



2. Given is a model of a Naive-Bayes-Classifier with two classes C1 and C2:

P(C1)	0.4
P(C2)	0.5
P(*world* | C1)	0.7
P(*world* | C2)	0.6
P(tx|C1), tx ≠ *world*	0.1
P(tx|C2), tx ≠ *world*	0.1

Classify the following document using this classification model.

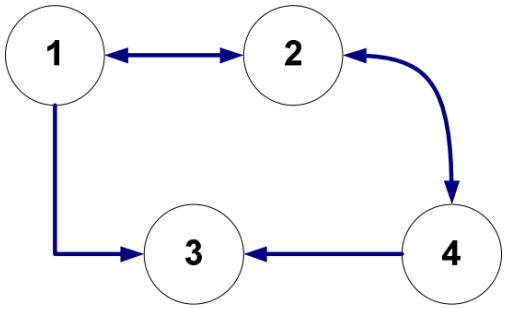
"World War Z was chosen to open the 35th Moscow International Film Festival."

4. Link analysis with PageRank

1. Given is the PageRank formula:

$$\vec{x}^{k+1} = (1 - c)\vec{x}^{k}A + \frac{c}{N}\vec{e}$$

and the following graph:



a. Create the link matrix *A'* with teleportation for this graph. Use the teleportation probability of 10%.
b. $\vec{e}$ is $\vec{1}$. In $\vec{x}_0$ all random surfers are on node 3. Compute the vector $\vec{x}$ for the first four iterations of the PageRank formula (*k* = 0..3) for this graph. Round to 5 decimal places!