Assignment 2

Information Retrieval and Text Mining 17/18

2017-11-14; to be submitted 2017-11-28

Roman Klinger, Florian Strohm

- **Deadline:** This assignment will be discussed on 2017-11-30. Submissions are accepted via Ilias until end of day 2017-11-28.
- **Groups:** Working in groups of up to three people is encouraged, up to four people is allowed. More people are not allowed. Copying results from one group to another is not allowed. Changing groups during the term is allowed.
- **Grading:** Passing the assignments is a requirement for participation in the exam in all modules IRTM can be part of. Altogether 80 points need to be reached. There are five assignments with 20 pen & paper points and 10 programming points each. That means, altogether, 150 points can be reached. Explain all solutions.
- **Submission:** First make a group in Ilias, then submit the PDF. Write all group members on each page of the PDF. Only submit *one* PDF file. If you are technically not able to make a group (it seems that happens on Ilias from time to time), do not submit a PDF multiple times by multiple people – only submit it once. Submission for the programming tasks should also be in the same PDF.

Task 1 (2 points)

According to the Jaccard Index, which document d_i is more relevant to the given query q? What are the values for each comparison? Explain your solution.

q algorithm intersection

d_1 the intersection algorithm for two documents is efficient

d_2 intersection of two or more objects is another smaller object

d_3 intersection algorithm

Task 2 (3 points)

According to cosine similarity with tf·idf weights, which of the documents d_1, d_2, d_3 is more relevant to the query “algorithm intersection”? Explain your solution.

Task 3 (3 points)

Answer the following questions about distributed indexing:

- What information does the task description contain that the master gives to a parser?
- What information does the parser report back to the master upon completion of the task?
- What information does the task description contain that the master gives to an inverter?
- What information does the inverter report back to the master upon completion of the task?

Task 4 (2 points)

Explain logarithmic merging in your own words. Include the motivation for this method in your explanation and make clear what advantages this method has in contrast to one auxiliary index and only one index on hard disk.

How could you use use a distributed compute cluster (for instance with map-reduce) in combination with logarithmic merging? Which advantages would your solution have and which disadvantages can occur?

Task 5 (4 points)

Heaps' law is an empirical law.

Assume that you have a collection with the following properties:

dataset	collection size	vocabulary size
subset 1	10M	100K
subset 2	1M	30K

- K means kilo: times 1000
- M means mega: times 1000000
- G means giga: times 1000000000

Subtask 5.1

Compute the coefficients k and b.

Subtask 5.2

Compute the expected vocabulary size for the complete collection (1G tokens).

Task 6 (3 points)

Given a collection that contains only four different words a, b, c, d. The frequency order is $a > b > c > d$. The total number of tokens in the collection is 6000. Assume that *Zipf's law* holds exactly for this collection. What are the frequencies of the four words?

Task 7 (2 points)

Calculate the variable byte code and the gamma code for 217.

Task 8 (1 points)

From the following sequence of γ-coded gaps, reconstruct first the gap sequence and then the postings sequence:
11110100001111101010111000

Programming Task 2 (10 points)

Implement a spelling correction with Levenshtein distance (you can use a library of your choice to calculate the distance between two strings, or you can implement this yourself).

As a list of correct words, you could use the file `english-words` on Ilias, but you can also use a different word list.

Choose between one of the following subtasks:

Subtask 1

Count how often each word in the Tweets from assignment 1 has been misspelled. (with a specific edit distance which you can chose, try to find a good value, explain how you did that)

What are the top ten most often misspelled words and their corrections?

Submit the whole programming code.

Subtask 2

Include the spelling correction in your indexer/query tool from assignment 1. Explain how you did that and submit the whole code again in the PDF submission file. Highlight the new spelling correction part in the submission. Can you show with some example queries that it improves the results?