Assignment 2

Information Retrieval and Text Mining 18/19

2018-11-13; to be submitted 2018-11-27

Roman Klinger, Johannes Erwerle

- **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 (or from elsewhere) 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.
- **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 (3 points)

According to cosine similarity, which document d_i is most relevant to the given query q? Use the log term frequency weight ($1 + \log_{10}(\text{tf})$, if $\text{tf} > 0$) as the weight for terms, as discussed in the lecture. What are the values for each comparison? Explain your solution and provide similarity measures for all query document pairs.

q algorithm intersection

d_1 intersection algorithm for two documents is efficient

d_2 intersection algorithm

d_3 algorithm

(note: cosine similarity will be discussed on the 20th of November)

Task 2 (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 3 (3 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? How would you distribute the different merge steps? Which advantages would your solution have and which disadvantages can occur?

Task 4 (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 4.1

Compute the coefficients k and b.

Subtask 4.2

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

Task 5 (3 points)

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

Task 6 (3 points)

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

1111010000111111010101110000

Task 7 (1 points)

Describe in your own words: What is the advantage of the k-gram index vs. the permuterm index for handling wildcard queries?

In general, would you take into account properties of the language for the decision which of the two approaches you prefer? Please explain!

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` and `german.dic` 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). Note that the Tweets are a mixture of German and English. Find a way to decide automatically if you correct to German or to English.

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

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? Provide examples that show differences in the results with and without spelling correction.