Trier University   Trier, 14.11.2022
Department IV – Computer Sciences
Prof. Dr.-Ing. Ralf Schenkel (H 539, schenkel@uni-trier.de)

Assignment 3 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, 21.11.2022

Note that the meetings on December 19, 2022 and January 2, 2023 will take place in the following Zoom meeting: `https://uni-trier.zoom.us/j/83637205106?pwd=RnowWVR6SUNJeHQ3MWZyeUtOV3Z3dz09` (Meeting-ID: **836 3720 5106**, password: **fHqDg15P**). It may be necessary to login with your university account before accessing the meeting.

Task 1:   Information Needs   3+6+6 Points

In this task we examine how search engines can satisfy different information needs.

(a) Assume that a user searches for the keyword „queen". Which information needs could be behind this query, i.e., what could the user actually be searching? Name at least three different information needs.

(b) Search for the keyword „queen"with two arbitrary search engines and document the top 10 results. Assign each result to one of your information needs from the previous subtask or to the need „others" if it does not match any. Now aggregate the numbers – which search engine yields how many results for which information need? Is the assignment always unique?

(c) How could one improve result quality, i.e, how could one help the search engine to yield many results for one specific information need? Give, for each of the three information needs listed above, a corresponding modified query, and check for one search engine how the results change.

Task 2:   Precision and Recall   5+5+5 Points

(a) Construct an information need that has to do with computer science (e.g., important pioneers in computer science research) and a corresponding query (e.g., „computer science pioneers") – make sure to pick not this information need and query. Determine the top 10 results for this query with an arbitrary search engine. Hand in a screenshot where the results are readable.

(b) Assess each result with respect to its relevance to your information need. Determine the precision of the result.

(c) Why is it impossible to determine the recall with these 10 documents? What could one do to determine a good approximation for the complete set of relevant results for this query, which could be used to compare the recall of two (or more) search engines?

Task 3:   Additional evaluation metrics   4+6+5 Points

(a) Research what the evaluation measures F_β (also known as F-Measure or F-Score) and *fallout* are. Give a formal definition for both and state your source. Explain with your own words the impact of the parameter β and if

- $F_\beta = 0$

- $F_\beta = 1$

- $Fallout = 0$

- $Fallout = 1$

are possible and, if so, what they mean.

(b) We consider a document collection D with documents d_1 to d_{30} and a query q. An expert has determined the documents in D relevant to q:

$$d_2, d_4, d_5, d_6, d_9, d_{13}, d_{14}, d_{18}, d_{26}$$

Three IR systems S_1, S_2 and S_3 are fed with the collection D and the query q and yield the following result sets:

- $S_1 : d_2, d_3, d_6, d_7, d_9, d_{13}, d_{14}, d_{16}, d_{17}, d_{24}$

- $S_2 : d_1, d_5, d_6, d_9, d_{10}, d_{13}, d_{14}, d_{18}, d_{21}, d_{22}, d_{24}, d_{26}, d_{27}$

- $S_3 : d_2, d_3, d_6, d_{10}, d_{13}, d_{14}, d_{18}, d_{26}$

Determine, for each system, $Recall$, $Precision$, $Fallout$ and F_1. Which system is better than another or maybe even the absolute best? Give a reason for your statement.

(c) Explain with a short text

- an IR system that yields the optimal value for recall (1) for *every* query

- an IR system that yields the optimal value for fallout (0) for *every* query

Both systems should be as simple as possible, and it must be possible to implement them. Show for both systems what the values for two other metrics would be. Try to give upper and lower bounds for these metrics.

These tasks will be discussed on November 28, 2022.

General remarks:

- The tutorial group takes place on Mondays in the regular meeting in F55 at 12:15.

- The first meeting of the tutorial group will be on November 14, 2022.

- To be admitted to the final oral exam, you need to acquire at least 50% of the points in the assignments. In addition, you need to present at least one solution of a task in a convincing way during the tutorial.

- It is preferred to submit in groups of size up to two (but not larger); in that case, only one submission is sufficient for the whole group. Write the names of all group members on your solutions.

- Solutions must be handed in before the deadline

  - in Moodle (`https://moodle.uni-trier.de/`, course `DL-IR-2022`) as as a PDF or, if submitting multiple files, as an archive (.zip or comparable).

  Submissions that arrive after the deadline will not be considered. The name of at least one group member should occur in the file name of your submission. If you want to modify a previously uploaded solution, just re-upload your solution.

- Graded versions of your submissions will be returned in Moodle until the following tutorial.

- Announcements regarding the lecture **and** the tutorial group will be done in the StudIP course for the lecture.