ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Information Retrieval
Winter 2019/2020

Prof. Dr.-Ing. Klaus Berberich
Telefon: 06 81 58 67-243
klaus.berberich@htwsaar.de

Programming Assignment 4

The programming assignment will be discussed on **February 3**. To obtain bonus points, you have to submit your solution via Moodle by **February 1 at 12:00 (noon)**. Please submit your solution, consisting of source code files and possibly libraries, **as one zip archive**. Teams of up to three students are allowed.

4.1 Command-Line Interface (1 Point)

In the previous programming assignments, you have programmed your own little search engine. In this exercise, we want to make it a bit more usable. Implement a command-line interface that allows users to use your search engine. The interface should greet users with a prompt where they can enter their queries. Once a query has been entered, the top-10 results should be printed to the standard output. For each result, its rank, its title, its URL, and its relevance score should be included in the result presentation.

4.2 Improve the Search Engine (1 Point)

During the course of the semester, we have seen various ideas how better results can be obtained. In this exercise, you are asked to try to improve your search engine over its current state. Some ideas that you can try:

1. Boost documents whose titles contain query terms
2. Boost documents that have been published more recently
3. Make use of other meta data that is available in the document collection (e.g., whether the article was shown on the title page or which categories it has been assigned to)

You are welcome to try own ideas. For the submission, please provide a short description of your idea and include the code that you used to implement it.

4.3 **Evaluation** (1 Point)

We now want to evaluate our search engine and your improved version (if you solved **4.2**). Determine the top-5 results for one or both systems for the queries listed below and pool them. Following that, assess the relevance of the documents using the following grades:

0 : Irrelevant

1 : Relevant

2 : Highly Relevant

The queries that you should consider are:

Q1 `summer olympics opening ceremony`

Q2 `alpine disaster kaprun`

Q3 `concorde crash paris`

Compute for each query and each system the value of nDCG@5 and also report the mean nDCG@5 over the three queries per system. For the submission, please include your relevance assessments in a CSV files with the columns query (e.g., Q1), document identifier, and relevance assessment (i.e., 0-2). Also submit a text file with the per-query nDCG@5 scores and the mean values for each system.