INFORMATION RETRIEVAL

HOMEWORK EXERCISES L11. QUERIES AND SESSIONS

SUZAN VERBERNE 2022

Universiteit Leiden

PRELIMINARIES

- The assignment on Brightspace has as attachment a sample of a query log of the search functionality on the Viva forum (https://forum.viva.nl/, Dutch)
- Download this file viva_search_correct.100K.txt.zip and unzip it

EXERCISE 1

- Extract and show the following basic statistics from the data:
  - The number of unique queries
  - The top-10 most frequent queries
  - The top-10 most clicked URLs

EXERCISE 2

- Create a matrix with as rows each of the top-10 most frequent queries, as columns the clicked URLs occurring in the data for the top-10 queries, and as cell values the click count for the query on the URL
  - Hint: don't use the top-10 URLs but all clicked URLs for the top-10 queries (number of columns is much higher than 10)
- For each query pair in the top-10 queries, calculate the cosine similarity between the queries using the matrix of click counts
  - Hint: one row in the matrix is a vector representing one query information
- Show a 10-by-10 matrix with the top-10 queries as rows and columns and a cosine similarity score in each cell.
  - Hint: the diagonal will be all 1s.