INFORMATION RETRIEVAL

HOMEWORK EXERCISES L11. QUERIES AND SESSIONS

SUZAN VERBERNE 2022

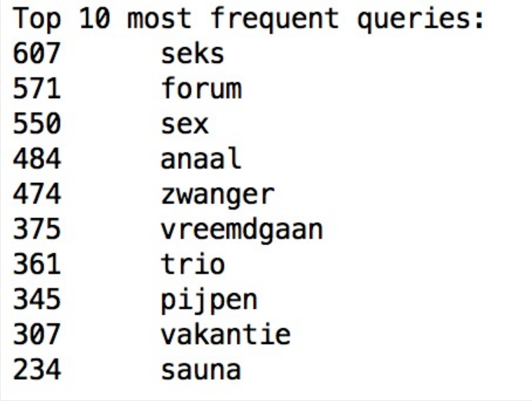
Universiteit Leiden

EXERCISE 1

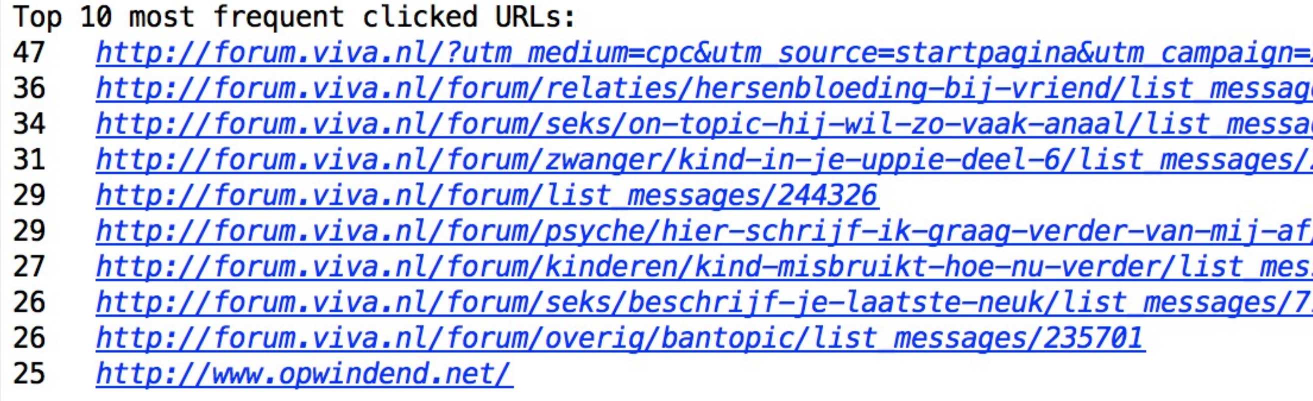
- Extract and show the following basic statistics from the data:
  a. The number of unique queries
  b. The top-10 most frequent queries
  c. The top-10 most clicked URLs

EXERCISE 1 - SOLUTION

- 1a. number of unique queries:
  - only event_type = “q”: 21559; all: 21810
- 1b. top-10 most frequent queries and 1c. top-10 most clicked URLs



```
Top 10 most frequent queries:
607       seks
571       forum
550       sex
484       anaal
474       zwanger
375       vreemdgaan
361       trio
345       pijpen
307       vakantie
234       sauna
```



```
Top 10 most frequent clicked URLs:
47    http://forum.viva.nl/?utm_medium=cpc&utm_source=startpagina&utm_campaign=
36    http://forum.viva.nl/forum/relaties/hersenbloeding-bij-vriend/list_messag
34    http://forum.viva.nl/forum/seks/on-topic-hij-wil-zo-vaak-anaal/list_messa
31    http://forum.viva.nl/forum/zwanger/kind-in-je-uppie-deel-6/list_messages/
29    http://forum.viva.nl/forum/list_messages/244326
29    http://forum.viva.nl/forum/psyche/hier-schrijf-ik-graag-verder-van-mij-af
27    http://forum.viva.nl/forum/kinderen/kind-misbruikt-hoe-nu-verder/list_mes
26    http://forum.viva.nl/forum/seks/beschrijf-je-laatste-neuk/list_messages/7
26    http://forum.viva.nl/forum/overig/bantopic/list_messages/235701
25    http://www.opwindend.net/
```

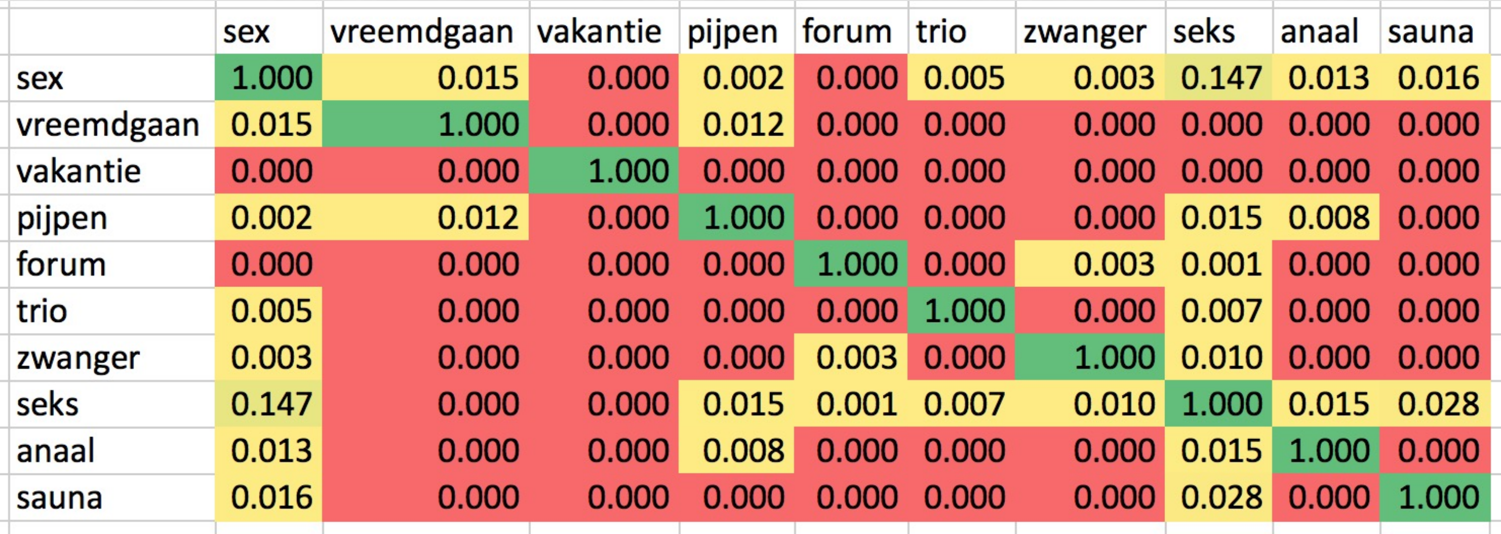
EXERCISE 2

- 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.

EXERCISE 2 - SOLUTION

Cosine similarity between queries

- matrix should have queries as rows and columns and show cosine similarities ([0..1]) between queries



	sex	vreemdgaan	vakantie	pijpen	forum	trio	zwanger	seks	anaal	sauna
sex	1.000	0.015	0.000	0.002	0.000	0.005	0.003	0.147	0.013	0.016
vreemdgaan	0.015	1.000	0.000	0.012	0.000	0.000	0.000	0.000	0.000	0.000
vakantie	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
pijpen	0.002	0.012	0.000	1.000	0.000	0.000	0.000	0.015	0.008	0.000
forum	0.000	0.000	0.000	0.000	1.000	0.000	0.003	0.001	0.000	0.000
trio	0.005	0.000	0.000	0.000	0.000	1.000	0.000	0.007	0.000	0.000
zwanger	0.003	0.000	0.000	0.000	0.003	0.000	1.000	0.010	0.000	0.000
seks	0.147	0.000	0.000	0.015	0.001	0.007	0.010	1.000	0.015	0.028
anaal	0.013	0.000	0.000	0.008	0.000	0.000	0.000	0.015	1.000	0.000
sauna	0.016	0.000	0.000	0.000	0.000	0.000	0.000	0.028	0.000	1.000