Categories for the Queries

Nicola Lea Libera

November 2020

1 Procedure for finding queries and matching categories

For the game, we had to collect a list of queries that we thought would be worth disguising. Therefore we looked up different sources of query collections (e.g. log files) and created a collection of our own.

The first thing we started with was the list of queries from Arampatzis that he mentioned in his work.[2] [4] Since the concept of the game included showing websites that matched the queries, sexual content was deleted from this list. Furthermore queries that contained heavy sexual content were not included in our collection at all.

Next, we looked up some forums to inspire ourselves. With the help of those we though of some queries on our own and added those to the list.

Afterwards, we searched through the aol log files which contain anonymous query logs. This increased our list even further.

An article concerning google [3] and topic files of this github repository[1] completed our list.

Categories and their numbers						
	Health	Personal	Knowledge/Education	Law	Crime	Politics
Number of queries per category	197 + 14	166 + 54	43 + 29	28 + 9	69 + 71	20 + 10

Red numbers mean that the category of the query is not unambiguous and could therefore also belong to another category.

Seeing the big gaps that exist between the numbers of some categories it might be wise to combine them into 4 instead of six. It would make sense to combine the categories *Knowledge/Education* and *Politics* and also the categories *Law* and *Crime*.

References

[1] URL: %7B%5Curl%7Bhttps://www.cnet.com/google-amp/news/google-is-giving-data-to-police-based-on-search-keywords-court-docs-show/?__twitter_impression=true&s=09%7D%7D.

[2] Pavlos S. Efraimidis Avi Arampatzis George Drosatos. *Versatile Query Scrambling for Private Web Search.* 2015.

[3] Alfred Ng. *Google is giving data to police based on search keywords, court docs show.* [Online; accessed 09-November-2020]. URL: %7B%5Curl%7Bhttps://www.cnet.com/google-amp/news/google-is-giving-data-to-police-based-on-search-keywords-court-docs-show/?__twitter_impression=true&s=09%7D%7D.

[4] *Seed queries.* [Online; accessed 09-November-2020]. URL: %7B%5Curl%7Bhttp://lethe.nonrelevant.net/datasets/95-seed-queries-v1.0.txt%7D%7D.