An Automated Method for the Characterization of Search Engine Results

Elizabeth McShane

MONTANA STATE UNIVERSITY BILLINGS

Project Overview

According to a 2012 Pew Internet survey, 91% of online adults use some form of web search[1]. While search engine optimization studies are commonly employed by companies to gauge their visibility in search results, few studies have been done to characterize results from the user's perspective. Since previous research has relied on manual review of search results[2,3] we set out to develop and test a set of software tools to gather and analyze search engine results automatically.

Objectives

Inspired by a lack of automated methodology, we set out to develop a program that would facilitate a variety of search engine-based research. In addition to developing a toolset that could utilize a website categorization API, we also wanted to make sure the program was flexible enough for a wide range of potential research applications. For instance, we want to be able to customize the way websites could be categorized based on future research goals.

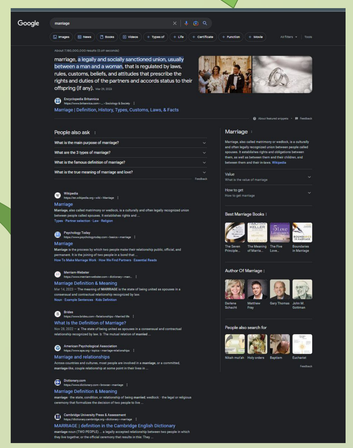
Thus, we set out to develop a modular program that contained three major parts:

1. A web crawler that would take a list of search terms and apply them to different search engines and return the results requested. For our trial run, we obtained the first 10 organic results[4].
2. Python-based algorithms that would take the data and prepare them for the web categorization API.
3. Python-based algorithm that would send a list of domains to the API and receive the proper categorizations and store the results in a database for later analysis.

Methodology

```
horseshoe
cabinet
supplementary
willpower
hallway
parallel
consolidate
garlic
racism
ideal
principle
exclude
particular
anniversary
inquiry
automatic
ruin
marriage
reject
demonstration
```

Selinium is used to search for each term and return results from each engine

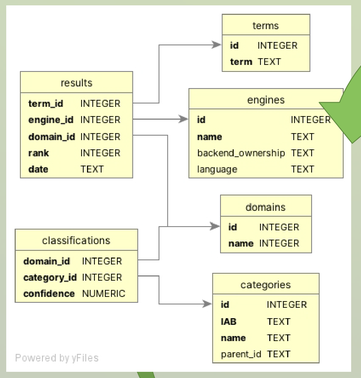


Python and pandas are used to clean up search result urls, leaving the domains. These domains are stored in the database.

	rank	term	searchengine	url	stripLink
0	1	horseshoe	google	https://en.wikipedia.org/wiki/Horseshoe	wikipedia.org
1	2	horseshoe	google	https://www.caesars.com/horseshoe	caesars.com
2	3	horseshoe	google	https://www.derbymuseum.org/Blog/Article/52/Th...	derbymuseum.org
3	4	horseshoe	google	https://thegoodluckgiftshop.com/horseshoe-luck...	thegoodluckgiftshop.com
4	5	horseshoe	google	https://bouldercountyopenspace.org/i/history/h...	bouldercountyopenspace.org
...	...	...	...	...	...
1685	9	uncertainty	bing	https://sciencing.com/how-to-calculate-uncerta...	sciencing.com
1686	10	uncertainty	bing	https://www.helpguide.org/articles/anxiety/dea...	helpguide.org
1687	11	uncertainty	bing	https://hbr.org/2021/04/6-strategies-for-leadi...	hbr.org
1688	12	uncertainty	bing	https://greatergood.berkeley.edu/article/item/...	berkeley.edu
1689	13	uncertainty	bing	https://ejje.weblio.jp/content/uncertainty	weblio.jp

The domains are given to the API for categorization using IAB taxonomy

Python and pandas are used to take the data received from the API and store them in a database for future analysis.



Analysis Possibilities:

- Comparing results within a given search engine
- Comparing results among search engines
- Determine:
  - Does a given search engine return consistent results?
  - Do different search engines return similar results?
  - Is there a pattern to the type (category) of websites returned?
  - Do results change over time?

Future Development Goals

1. Develop more tests to ensure urls are processed properly
2. Continue sending urls to API for classification
3. Develop experimental models
4. Create a GUI
5. Continue to develop web scraping capabilities
6. Continue to develop analysis program
7. Add code to remove already-classified domains before submission to API

References

1. Purcell, Kristen, Joanna Brenner, and Lee Rainie. "Main Findings." *Pew Research Center: Internet, Science & Tech* (blog), March 9, 2012. https://www.pewresearch.org/internet/2012/03/09/main-findings-11/.
2. Hawking, David, Nick Craswell, Peter Richard Bailey, and Kathleen Griffiths. "Measuring Search Engine Quality." *Information Retrieval*, 2001. https://doi.org/10.1023/A:1011468107287.
3. Anuyah, Oghenemaro, Ashlee Milton, Michael Green, and Maria Soledad Pera. "An Empirical Analysis of Search Engines' Response to Web Search Queries Associated with the Classroom Setting." *Aslib Journal of Information Management* 72, no. 1 (January 1, 2020): 88–111. https://doi.org/10.1108/AJIM-06-2019-0143.
4. Southern, Matt G. "Over 25% of People Click the First Google Search Result." Search Engine Journal, July 15, 2020. https://www.searchenginejournal.com/google-first-page-clicks/374516/.
5. IAB Tech Lab. Content Taxonomy. https://iabtechlab.com/standards/content-taxonomy/

Special Thanks

John Pannell: mentorship, guidance and patience.
Taylor Branson: Selinium processing and test data.