**Search Audit Documentation**

# **Usage**

* Run: python3 search\_audit.py
* Ouput: search audit results written to data/ folder, one output for each jurisdiction and search query type (includes: search query, domain, domain base, domain suffix, location)
* Best to use / set up Amazon [EC2 instance](https://aws.amazon.com/ec2/) or some variation to run in the background because search queries
* Steps to set up Amazon EC2:
  + Contact Metin to set up an Amazon EC2 account (or use shared EC2 instance for this project so everyone has the same access)
  + Metin will provide you with a key
  + Save that key in your folder with the name [name.pem]
  + X
  + X
  + X
  + X
* mkdir data/ on terminal if the folder data/ does not exist. The current program writes files to the data folder

**Search\_audit.py**

* Gather canonical names of locations (see documentation from [WebSearcher, localization](https://github.com/gitronald/WebSearcher#localization)) and modify the variable “locs” with the name of the locations mapped to the canonical name
  + How to gather canonical names:
  + X
* Code explanation e.g. line q[0] = re.sub(r”\bcalifornia\b”, state, q[0]) : some queries mentioned cities, for example the query could be: “eviction in California”. We are of course not interested in users from Florida who mention California in their queries. To solve this problem the code substitutes these cities with the states mentioned in locs: e.g. the new query would become “eviction in florida”. If you have any future queries where cities need to be substituted, add them to the loc.items list. Added loc.items are not processed/taken into account/removed?
* Modify the variable “csv\_files” to include all of the csv files with the queries you are planning to run, for example eviction.csv, domestic\_violence.csv, etc.
* Note: Included timer to force the program to not send hundreds of queries in succession; if you receive a time-out issue, it is most likely due to issues with running queries too quickly via Google. Currently, using time.sleep() to time out; can toggle with the values as needed.
* Note: mkdir data/ on terminal if the folder data/ does not exist. The current program writes files to the data folder.
* Note: to run this program in the background (so you can leave your laptop while the program runs) use command: nohup python3 search\_audit.py &
  + To force quit: ps -ef | grep "python3" (gives you PID)
  + kill -9 PID (where PID is the number associated with the command)

**R studio**

* Change the directory to your directory
* Queries that are not useful were removed. You can either add non-useful queries to line 20: *remove\_queries <- c( )* or delete them from the csv.
* Replace the names of the files
* Replace the names of the locations
* Lines 243 - 268: add or remove non-US url suffixes
* Change title, subtitle, x and y coordinates descriptions

# **General pseudocode / logic**

1. For each csv file, run the search query on each specific location
2. Modify any city or state-name specific query and replace it with the location (e.g. “Eviction San Francisco” would be modified to “Eviction Tallahassee” if running Tallahassee queries)
3. Run the query using the web searcher engine. If the query is unsuccessful, time out for a few minutes and re-run the query.
4. Gather all of the related search queries related to that location and write to a CSV file

After, run a general data analysis that examines things such as top search results, domain suffixes, etc.

# **EC2 appendix**

Replace [ubuntu@ec2-34-220-227-191.us-west-2.compute.amazonaws.com](mailto:ubuntu@ec2-34-220-227-191.us-west-2.compute.amazonaws.com) with associated EC2 instance

**SSH into Amazon AWS:**

ssh -i ~/.ssh/pem\_file.pem ubuntu@ec2-34-220-227-191.us-west-2.compute.amazonaws.com

**SCP to AWS instance:**

scp -i ~/.ssh/pem\_file.pem ~/path/to/file ec2-user@ec2-3-15-182-227.us-east-2.compute.amazonaws.com:~/

**SCP from AWS instance**

scp -r -i ~/.ssh/pem\_file.pem [ubuntu@ec2-34-220-227-191.us-west-2.compute.amazonaws.com](mailto:ubuntu@ec2-34-220-227-191.us-west-2.compute.amazonaws.com):~/data/path/to/file ~/Documents/legal\_design\_lab/