Running the Webpage

To run the webpage open the index.html file with your browser. You will then need to put in the csv file generated by the scrapers (madagascar-business.csv) and a favorites csv file (favorites.csv). You can enter data in any or all of the fields Service, Location, and Favorites Only, and you will get relevant results. You will then be able to favorite or unfavorite each business. Once you are done hit save favorites and you can use the downloaded file next time you use the webpage.

Python Library to Install for Scraping

- Python (Version 3.0+)
- Beautiful soup - pip install bs4
- Selenium - pip install -u selenium
- ScraperApi - set up an account on Scraperapi's website https://www.scraperapi.com/

Main File

- Check chromedriver path, should be `'chromedriver.exe'` for Windows or `'chromedriver-mac.exe'` for Mac
- Put in scraperapi key
- If you are running MacOS, you may need to enable the executable file to run in your system's security / privacy settings
- May only want to run one scraper at a time since most of them take a very long time to run

Running Scrapers

Initialize each scraper passing in **keywords** parameter and possibly other parameters. See **main.py** for an example.

WebScraper Base Class

This class is a template for web scrapers. All web scrapers are subclasses of this base class and inherit its methods and class variables. All methods and class variables are accessible from sub classes.

Variables

- **keywords:** all relevant keywords that help subclasses find relevant businesses
- **csv_columns:** header of columns in CSV file
- **business_list:** meant to be a list of dictionaries, contains all information about businesses. Every entry in list is another business

```
Ex. [{name: “Madagascar Business”,
          Search_term: “furniture”,
          Service: “furniture”,
          Location: “Madagascar”,
          Phone: “888 888 8888”
          Email: “furniture@email.com”
          Website: “exampleWebsite.com” }]
```

- **filename:** filename of CSV file that will contain all businesses at the end of scraping
- **url_list:** contains all URLs to scrape when web scraper is parsing

Methods

- **set_keywords:** use in case you want to update keywords
- **write_data_to_csv:** writes data from business list to CSV file which is saved under filename
- **append_data_to_csv:** appends data from business list to CSV file

Google Map Scraper

Link: https://www.google.com/maps
Runtime: 30 min - 1 hour
Required Parameters:

- **keywords**: which keywords to put in the search bar

Optional Parameters:

- **chrome_driver_path:** `'chromedriver.exe'` for Windows or `'chromedriver-mac.exe'` for Mac
- **wait_time**: How long to wait for the page to load. Increase if you get an error

- **coordinates**: Latitude, longitude coordinates of where to search. May need to adjust south or north if looking for businesses on these parts of Madagascar. You may also want to find the location on google maps and put in the coordinates.
- **zoom**: Zoom level of where to search. The scraper will actually zoom in once before scraping so be sure to account for that. Find appropriate zoom level on Google Maps and put into this parameter

Why Run Again?

- Can run again at different coordinates or zoom level to get different results
- Can run again with different keywords that will be different searches on Google Maps
- Can run again with the same parameters to capture businesses recently added to Google Maps

Future Improvement:

- Currently not all businesses on each page of results are collected since you need to scroll down to load the businesses lower on the list, so adding a scroll on each page should result in finding more businesses.

Madagascar Business Directory Scraper

Link: http://www.madayp.com/
Runtime: 0.5~1hr
Required Parameters: Scraperapi key: set up an account on Scraperapi's website https://www.scraperapi.com/ and paste the api key into the code

How Does It Work:

- Goes through each relevant business category on the website and finds all urls pertaining to the individual pages for each business
- Visits each through the Scraperapi and extracts information about each business and saves in a list
- Iterates through list and compiles business dictionary
- Writes business dictionary to CSV file

Why Run again?:

- To gather businesses that were newly added to the website, but this will not happen very often

- To scrape businesses from a category that was previously not included, you should add the corresponding category name from the website in self.search terms.
- In general, it is not recommended to run this scraper frequently because it requires a fair amount of Scraperapi calls.

Notes:

- Frequent connection issues during running, exceptions are dealt with but this might result in runs with less businesses scraped than expected
- One run of the scraper could take anywhere from 500 to 1000 Scraperapi calls. Each account gets 1000 free Scraperapi calls per month

LinkedIn Scraper

Link: https://www.linkedin.com/

Runtime: 2-2.5 hours

Required Parameters: keywords to search for relevant linkedin profiles

How Does It Work:

- Uses Selenium to download LinkedIn Email Finder chrome extension each time chrome driver opens
- Logs into a LinkedIn account and the chrome extension account so that the user can use the chrome extension to gather contact information from a specific profile
- Searches on google for all the possible linkedin profiles associated with the keywords
- Uses Beautiful Soup to filter out the unnecessary profiles by scraping the location of that company or where that person is located, as well as, scraping the google search link description
- Scrapes the title and location given on the linkedin profile and stores it in the business dictionary
- Writes business dictionary to csv file

Why Run again?

- To gather information from recently added profiles to businesses or people
- To find Linkedin profiles associated with new keywords added

How to use the chrome extension:

- Open up a linkedin profile and then click the chrome extension icon on the top right corner of the browser
- A small window will pop up and it will show the name of the linkedin profile you are currently on.
- Click the plus button right of the profile name on the chrome extension window and the extension will generate the correct contact information for that linkedin profile

Notes:

- The chrome extension can gather 400 contacts every 24hrs
- Linkedin will sometimes require security verification. This will especially happen if the program gets run outside of Colorado. You can access the code sent to this email 'hastingsj890@gmail.com' and the password is 'email1234#'. Once the code is inputted, rerun the code and linkedin will no longer ask for a security verification

Yellow Pages Scraper 1

Link: http://www.business-yellowpages.com/madagascar/page-4
Runtime: 1-2 minutes
Required Parameters:

- **keywords**: which keywords to filter businesses by

How Does It Work:

- Uses Beautiful Soup to get business information from all search results
- Filters business dictionary by keywords to find only relevant businesses

Yellow Pages Scraper 2

Link:
https://www.yellowpagesofafrica.com/companies/madagascar/building-materials/start-1/
Runtime: ~2-3 hours
Required Parameters:

- **keywords**: which keywords to filter business categories by

How Does It Work:

- Uses Beautiful Soup to find all URLs pertaining to relevant keywords. Must do because URL format does not change predictably with search bar.
- Composes a list of different result pages for one business category based on how many search results come up
- Iterates through list and compiles business dictionary
- Writes business dictionary to CSV file

Notes:

- Will have to change location of ChromeDriver in parse() method before running

Future Improvements:

- Contact information is not being split into phone, email, website because not all businesses have all those or have multiple of one so it is hard to split

GEM Website

Link: http://www.gem-madagascar.com/membres
Runtime: 5-10 minutes
Keywords: No keywords available
How does it work:

- Obtains a list of member business and associated company URLs found on Groupment de Enterprises de Madagascar
- Goes through each member URL and parses information such as name, location, phone and fax, and official member website
- Write business dictionary to CSV file

Notes:

- Unfortunately, most businesses don't seem to have keywords or descriptors that would make parsing based on them worthwhile
- It appears that this list of businesses is not updated too frequently, so running this scraper once should be sufficient for short to medium term searches

Why Run again?:

- To gather businesses that were newly added to the website, but this will not happen very often