### Web Scraper README

This README explains the approach taken to design the web scraper. For this, I took a general approach to search recursively starting at a base url to search for a given regular expression. In this way, the script can be more easily generalized to search for other strings such as email addresses, physical addresses, etc. as well as phone numbers by changing the regular expression.

The script for this is regexWebScraper.py. I ran it under Python version 2.7. It can be run simply by invoking:

>>> python regexWebScraper.py

The script has one dependency which is the BeautifulSoup module. The module is where helpful to effectively and efficiently parse large HTML content. You can obtain the module by downloading version 3.2.1 from <http://www.crummy.com/software/BeautifulSoup>.

There is no installation needed as the module BeautifulSoup.py can be moved to the workspace with the web scraper script after unzipping and extracting the zipped file.

The scraper basically does a breadth first search of the links from the baseUrl reading the content of every page and doing a regular expression find for the given expression. If a URL cannot be opened for any reason (i.e. permissions) then it simply skips and continues the search. It keeps a set of visited URLs and suffixes to avoid revisiting any links. After it reads the content of a page, it will pass this content into the BeautfulSoup class which then gives us the ability to easily parse those contents for information such as links which will be visited as long as they have not been already. The matches for the regular expression which is passed in are stored as a set to avoid having any duplicates.