Steven Krenn, Wendy Wei

Instruction

To create an index.html file with all of the keywords scraped from a live website

Objective

The purpose of this tool is to help clients to obtain patent keywords by scraping the internet from a live website and creating a text file with all of the patent data, that is then parsed and then keywords are outputted to a index.html file

Methodology

We first used the requests and beautifulSoup libraries on this project. The requests library requests a website and calls down all of it’s source html code. BeautifulSoup parses the html file in to a more readable format. We used BeautifulSoup to output all of the hyper links of the file. We then processed the list in to a workable format for iteration. We then iterated over the list and wrote the contents of each patent website’s text to a master file. After we have the master output.txt file, we parse the text file and create a dictionary of keywords and how many times they show up in the output.text file. We then sort dictionary from highest frequency of occurrence to lowest. We then output that sorted dictionary to the index.html file for output.

Scope

This tool can be applied to any other patent files which has a similar format of current input file. It can be applied to a different search of 15 links from the input website.