Web Page Analysis Crawler.

This is web scraper and crawler to analysis the keyword which classifies the web page. I used Beautiful soup, which is a html parser and parse all the html into text. Than I filter all text is visible on page and analyses on it.

Requirements:

Python3.x
BeautifulSoup
HTMLParser
Request

Installation:

pip install BS4
pip install requests

Algorithm:

Step1: Parse the web page using HtmlParser BeautifulSoup.
Step2: Get all the text visible on page into a list. The list would contain all text from all tags expect ['style', 'script', '[document]'] and Comments
Step3: Filter the list in a standard format.
Step4: delete the words which are less important using the file containing unwanted word list.
Step5: Get text from the 'title' tag and repeat step 3 and 4.
Step6: Sort the list. And get moss repeated words from the list and add to title list and return final list.

How to run:

Go to in the webcrawler directory.
Run: python3.5 Scraper.py [url] on terminal

Example: cd webcrawler
python3.5 Scraper.py https://www.brightedge.com/

Output: List of words are: ['marketing', 'customers', 'seo', 'brightedge', 'enterprise', 'company', 'content', 'platform']

python3.5 Scraper.py http://www.amazon.com/Cuisinart-CPT-122-Compact-2-Slice Toaster/dp/B009GQ034C/ref=sr_1_1?s=kitchen&ie=UTF8&qid=1431620315&sr=1-1&keywords=toaster

Output:
List of words are: ['cpt 122 ', 'kitchen ', 'amazon com ', 'compact ', 'plastic ', 'slice ', 'stars ', 'cuisinart ', 'white ', 'dining ', 'toaster ']

** Please find document about function and methods in the program source file…