Documentation

How To Execute:

1. python main.py [url]
   python main.py http://www.cnn.com/2013/06/10/politics/edwardsnowdenprofile/

2. python main.py
   >> Please enter URL: http://www.cnn.com/2013/06/10/politics/edwardsnowdenprofile/

Output:

Keywords**:** nsa, safeguard, leaks, liberty, privacy, man, government, snowden

Libraries Used:

BeautifulSoup
Urllib3

Description:

This program is designed to take an url as an input and extract keywords from the website and return them as the output. The approach used is as follows:

- Fetch the URL.
- Download the webpage.
- Extract each part of the webpage using BeautifulSoup library.
- Normalize the content/words.
- Processing the content and headers.
- Translate each of the above to Bag of Word representations.
- Remove Stop Words.
- Get top 20 words from content and top 10 words from the headers.
- The words in the header are more important than in the content so they are given 3 times more weight.
- The list of words from content and header is then merged and sorted in reverse order.
- Report top 8 words with maximum frequencies.

Architecture:

1. Main program: Main program takes url from standard input and performs program execution by calling various functions to process the input and print the topics that best describe the article.

2. Parse program: It is responsible for web crawling. Parse program parses the HTML page by BeautifulSoup and prepares the data for the Process.

3. Process program: It is responsible for word countings, and analysis. Text processing is done here including calculating Unigram and removing stop words.

Error Handling:

1. Make sure all the libraries and dependencies are installed before you run the program.
2. Error is appropriately handled for various types of input in case if the url is not properly typed or not specified in the command.
3. While connecting to the URL, if we are unable to get content then it will be handled by try except block.