The very step, for data extraction & text analysis, is to load the required libraries of BeautifulSoup & requests. Along with this, we’ve also imported the necessary positive words & negative words datasets, for determining which are the positive & negative words respectively.

* We begin with the first URL.
* For the sake of ease & clarity, the whole process has been divided into two broad steps, the DATA EXTRACTION part & the DATA ANALYSIS part.

**DATA EXTRACTION**

* Connect to the website by providing the URL.
* Use the get method to retrieve raw html content from the specified URL.
* Create a text file (1.txt) and open it in the read mode.
* Parse the html contents of the website.
* Using the prettify method on the parsed content, convert it into a nicely formatted Unicode string.
* Search for the specific id (in this case h1) by also mentioning its class, to retrieve the title of the website.
* Repeat the above step to get the content of the website.
* Open the text file (1.txt) created above in write mode, and write the title and content respectively, in the file and save it.

**DATA ANALYSIS**

1. **Removing stopwords from text file**

* Import the nltk library.
* From the corpus package, download the list of stopwords.
* Get the nltk stopword list into a set.
* Open the previously saved text file (1.txt), read the lines and split the lines into individual words.
* Check if any of the individual words present in the text file match the stopwords list.
* If any word in the text file is not in the stopwords list, then that word is stored in a newly created text file (stopwords-removed1.txt).
* This new file stores all the words of the previous text file that are not stopwords.
* Count the number of clean words, i.e. words present in the newly created text file (stopwords-removed1).
* Check if any of the individual words in the text file are present in the positive words list.
* Check if any of the individual words in the text file are present in the negative words list.

1. **Extracting derived variables**

* Calculate the number of positive words in the text file by using ‘len’ function. This gives the positive score.
* Calculate the number of negative words in the text file by using ‘len’ function. This gives the negative score.
* Using the mentioned formulas, calculate the polarity score & subjectivity score
* Count the number of sentences. Open the file and read the lines. If we encounter any of these [. ! ?], then the sentence count Is incremented.
* Count the syllable length & complex words. Syllable count will be incremented if any vowels in the individual word is encountered or if the word ends with ‘e’ or ‘le’ etc. A word is a complex word if the syllable count for that word is greater than 2.

1. **Analysis of readability**

* Using the standard formulas, calculate the average sentence length, percentage of complex words, fog index and average words per sentence.
* Calculate the number of personal pronouns (any word like “I”, “we”, “my”, “ours”, “us”, “We”, “My”)
* Calculate the total number of characters in the text file
* Calculate the average word length