**HOW TO RUN:**

Just upload the ipynb file into Google Colab and ctrl+f9 to run all the cells

**SIDENOTE BEFORE READING APPROACH SECTION:**

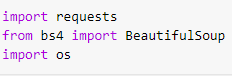
**LINKS 37 AND 50 IN THE OUTPUT DATASTRUCTURES ARE SHOWING 404 ERROR SO I LEFT BLANK FOR THEM IN THE FILE**

**APPROACH TO SOLVE PROBEM:**

**Task 1: To create a function to extract article text and also its title and store it in a text document**

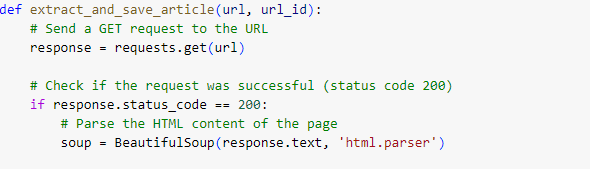
Import Necessary Libraries:

1. Requests: send http requests to the url and retrieve page content
2. BeautifulSoup: used for navigating and extracting the required info from webpage
3. Os: for storing the results of the extraction to a text file



Function definition: needs to have url and url\_id

After checking the request and its status, I used beautifulsoup to parse through the document



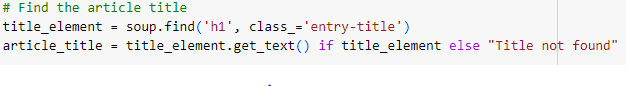
Now in order find the title and text details I went to said website to find its details

First: Right click -> Inspect

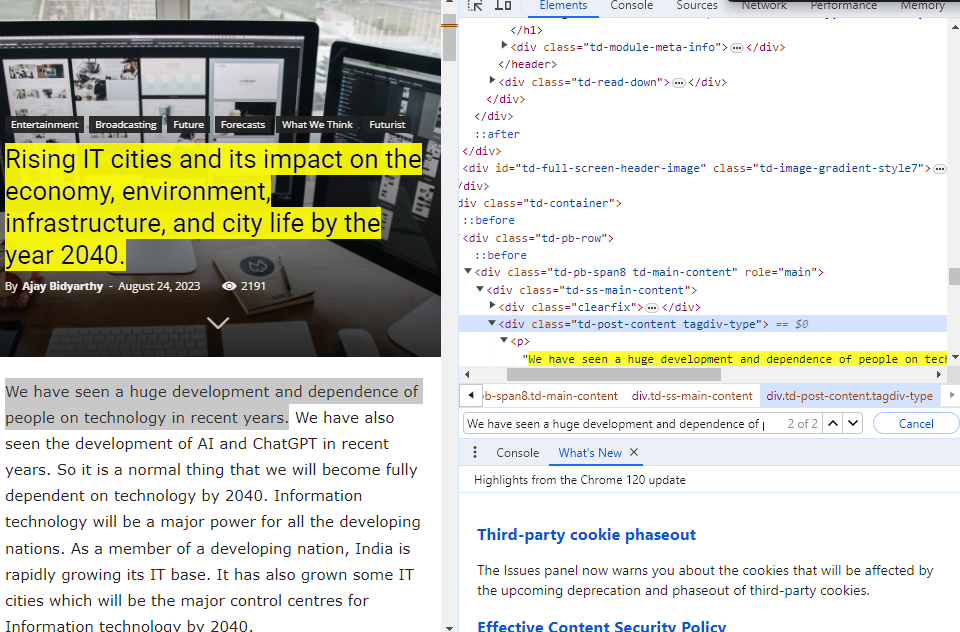
Second: In the side cltrl+f to search through file, copy paste the title and find



Since I know the title is stored as header as h1 tag and in class = entry-title, I implement them for the search and extraction

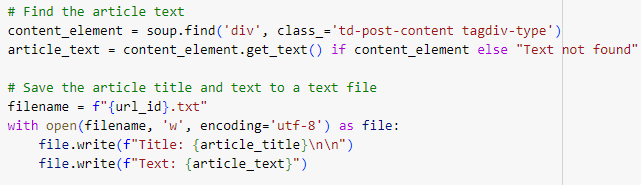


Same way I did for the content:



I choose the div tag for this all the text content is stored in that div tag and its respective class = td-post-content-tagdiv-type

And later store them the text file



I created addition functions similar to this function , when the webpage had different class name for h1 or div tag.

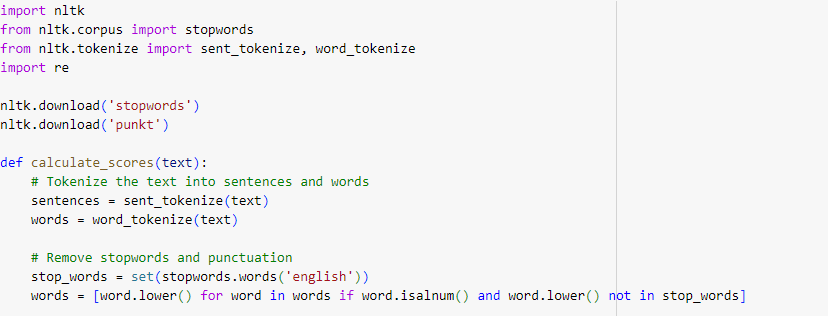
**Task 2: Create a function to perform sentiment analysis and other text metrics on the text file and then print the scores obtained**

Import necessary libraries:

1. Nltk: for text into sentences and words and remove stopwords and punctuations.
2. Re: for pattern matching

Function definition: text file opened in read mode

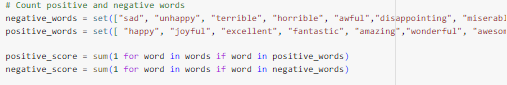
Text is tokenized into sentences and words and the stopwords and punctuations removed



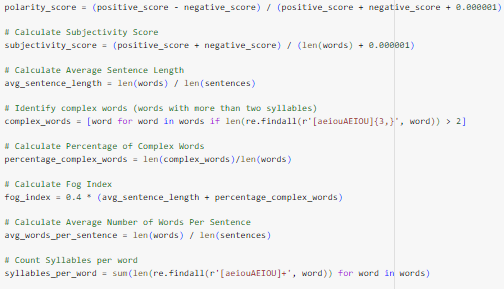
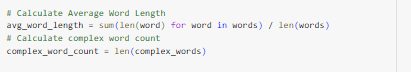
Sentiment Analysis:

Define sets of positive , negative and personal pronouns and count their occurrences in the cleaned text.

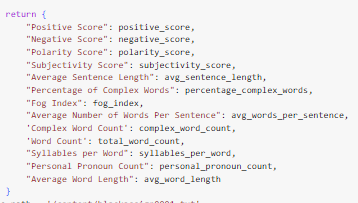




Score calculations: As noted form the Text Analysis.docx …..so no explaining here

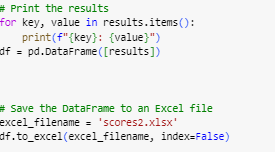
  


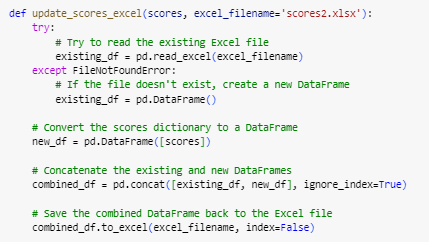
Print the scores:



**Task 3: Create a excel file to store the scores obtained and an update function so keep updating the excel file with new scores from other websites links.**

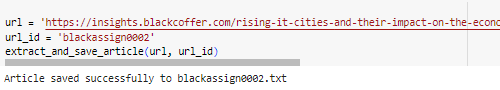
Import pd : for creating dataframe and in then update it with scores found and create a function to concat the new and and old scores into the excel file.





**Finally when I am creating text file , calculate scores and update score into excel it will look like this:**

Create text file:



Calculate scores , print it and then update it excel file

