PROJECT REPORT: INVEST INDIA

Project Objective: A report on how the importance of Indian Economy over time has changed in lieu of the Governments priorities through an analysis of PM Modi’s speeches and the number of times the word ‘Economy’ or ‘Economic’ is used.

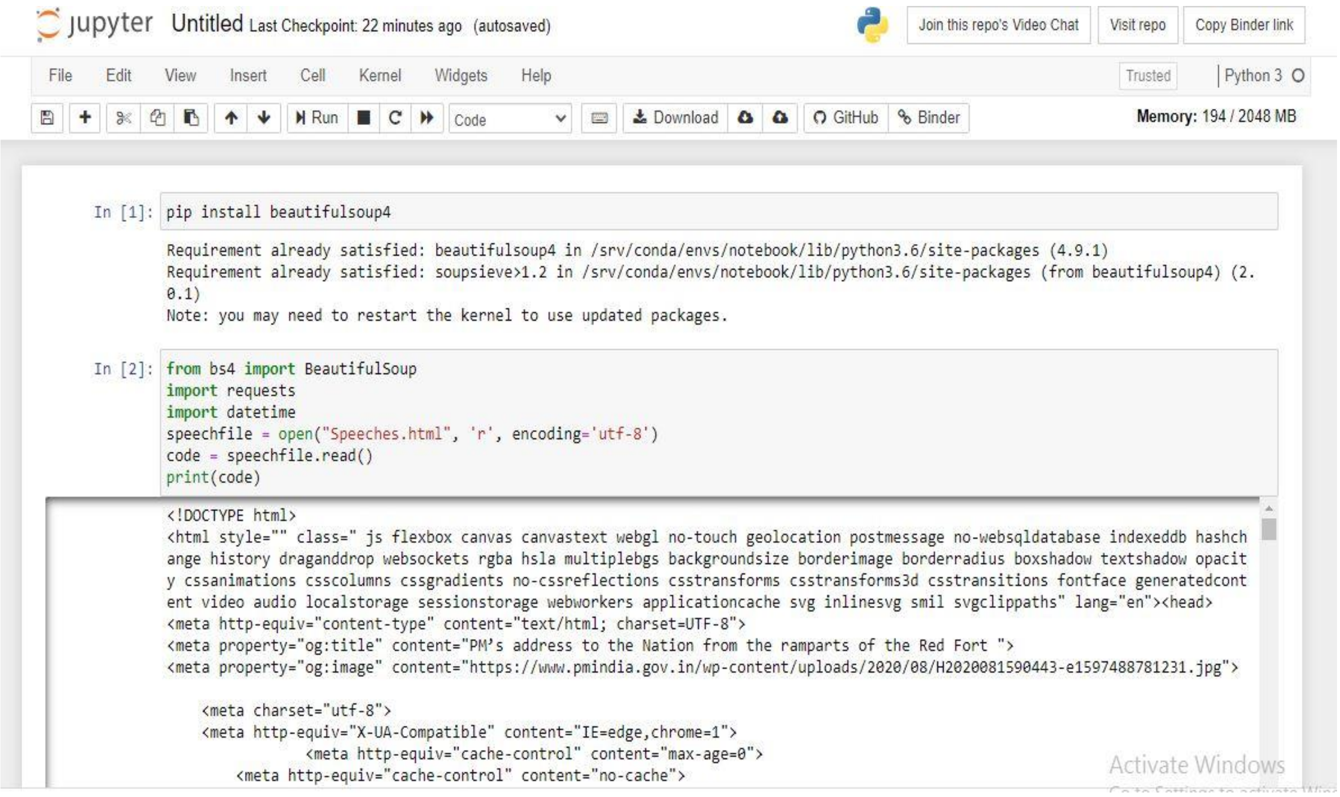
Brief of methods used: Web scraping, BeautifulSoup4 module in Python, Jupyter Notebook environment for execution

The Idea/Algorithm:

1. A public webpage https://www.pmindia.gov.in/en/tag/pmspeech/?query containing the speeches of PM Modi was used.
2. Web scraping of HTML source code of main page provided link to speech page and date of speech.
3. Web scraping was used to extract only the actual text of the speech made. Done for last 50 speeches.
4. The number of times ‘Economy’ or ‘Economic’ was used in each speech was counted and stored along with the date of the speech.
5. A simple point plot was done with speech date on the x axis and ‘Economy’ word count on the y axis. The graph was analysed.

Step by Step code used with Explanation:

1. Python module beautifulsoup4 used for web scraping was installed. The necessary libraries were imported. The html page of the speeches downloaded beforehand was loaded and opened in read mode.



In [1]:
```
pip install beautifulsoup4
```

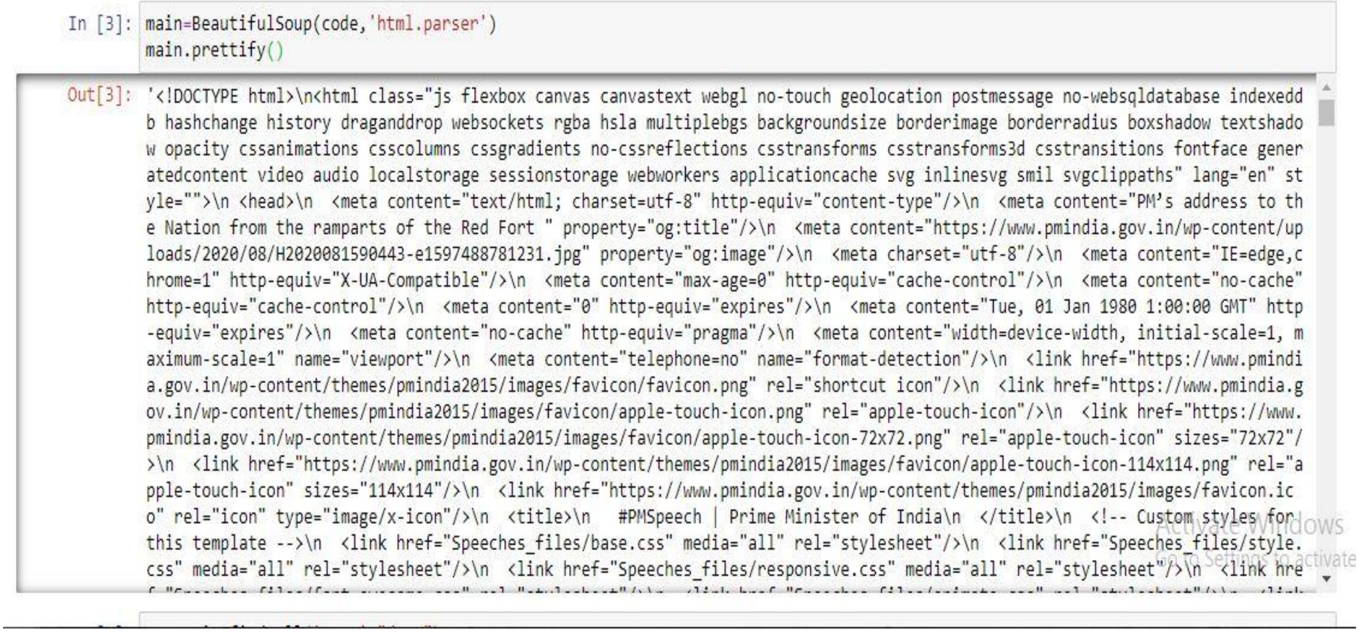
```
Requirement already satisfied: beautifulsoup4 in /srv/conda/envs/notebook/lib/python3.6/site-packages (4.9.1)
Requirement already satisfied: soupsieve>1.2 in /srv/conda/envs/notebook/lib/python3.6/site-packages (from beautifulsoup4) (2.
0.1)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
from bs4 import BeautifulSoup
import requests
import datetime
speechfile = open("Speeches.html", 'r', encoding='utf-8')
code = speechfile.read()
print(code)
```

```
<!DOCTYPE html>
<html style="" class=" js flexbox canvas canvastext webgl no-touch geolocation postmessage no-websqldatabase indexeddb hashch
ange history draganddrop websockets rgba hsla multiplebgs backgroundsize borderimage borderradius boxshadow textshadow opacit
y cssanimations csscolumns cssgradients no-cssreflections csstransforms csstransforms3d csstransitions fontface generatedcont
ent video audio localstorage sessionstorage webworkers applicationcache svg inlinesvg smil svgclippaths" lang="en"><head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<meta property="og:title" content="PM’s address to the Nation from the ramparts of the Red Fort ">
<meta property="og:image" content="https://www.pmindia.gov.in/wp-content/uploads/2020/08/H2020081590443-e1597488781231.jpg">

    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
                <meta http-equiv="cache-control" content="max-age=0">
        <meta http-equiv="cache-control" content="no-cache">
```

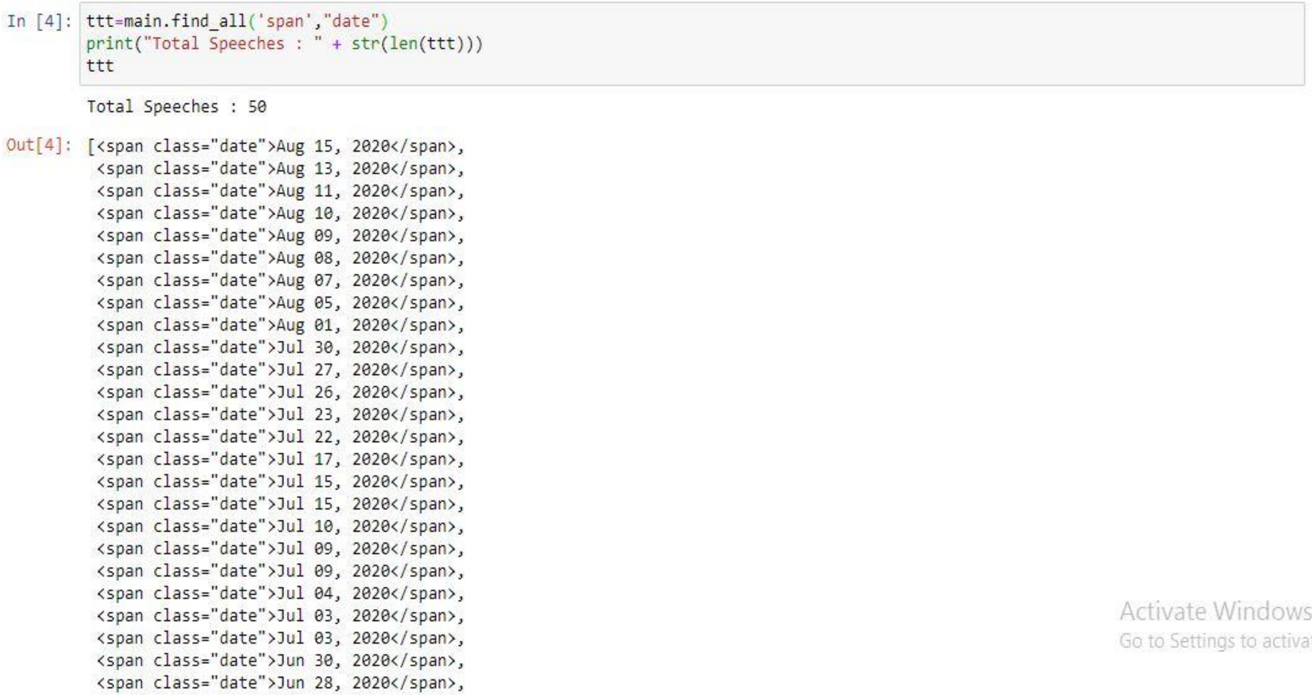
2. A BeautifulSoup object was created and the HTML code for the main page was loaded. The BeautifulSoup objects helps us parse the code. We then display the code in a structured way.



```
In [3]: main=BeautifulSoup(code,'html.parser')
        main.prettify()
```

```
Out[3]: '<!DOCTYPE html>\n<html class="js flexbox canvas canvastext webgl no-touch geolocation postmessage no-websqldatabase indexedd
        b hashchange history draganddrop websockets rgba hsla multiplebgs backgroundsize borderimage borderradius boxshadow textshado
        w opacity cssanimations csscolumns cssgradients no-cssreflections csstransforms csstransforms3d csstransitions fontface gener
        atedcontent video audio localstorage sessionstorage webworkers applicationcache svg inlinesvg smil svgclippaths" lang="en" st
        yle="">\n <head>\n  <meta content="text/html; charset=utf-8" http-equiv="content-type"/>\n  <meta content="PM’s address to th
        e Nation from the ramparts of the Red Fort " property="og:title"/>\n  <meta content="https://www.pmindia.gov.in/wp-content/up
        loads/2020/08/H2020081590443-e1597488781231.jpg" property="og:image"/>\n  <meta charset="utf-8"/>\n  <meta content="IE=edge,c
        hrome=1" http-equiv="X-UA-Compatible"/>\n  <meta content="max-age=0" http-equiv="cache-control"/>\n  <meta content="no-cache"
        http-equiv="cache-control"/>\n  <meta content="0" http-equiv="expires"/>\n  <meta content="Tue, 01 Jan 1980 1:00:00 GMT" http
        -equiv="expires"/>\n  <meta content="no-cache" http-equiv="pragma"/>\n  <meta content="width=device-width, initial-scale=1, m
        aximum-scale=1" name="viewport"/>\n  <meta content="telephone=no" name="format-detection"/>\n  <link href="https://www.pmindi
        a.gov.in/wp-content/themes/pmindia2015/images/favicon/favicon.png" rel="shortcut icon"/>\n  <link href="https://www.pmindia.g
        ov.in/wp-content/themes/pmindia2015/images/favicon/apple-touch-icon.png" rel="apple-touch-icon"/>\n  <link href="https://www.
        pmindia.gov.in/wp-content/themes/pmindia2015/images/favicon/apple-touch-icon-72x72.png" rel="apple-touch-icon" sizes="72x72"/
        >\n  <link href="https://www.pmindia.gov.in/wp-content/themes/pmindia2015/images/favicon/apple-touch-icon-114x114.png" rel="a
        pple-touch-icon" sizes="114x114"/>\n  <link href="https://www.pmindia.gov.in/wp-content/themes/pmindia2015/images/favicon.ic
        o" rel="icon" type="image/x-icon"/>\n  <title>\n   #PMSpeech | Prime Minister of India\n  </title>\n  <!-- Custom styles for
        this template -->\n  <link href="Speeches_files/base.css" media="all" rel="stylesheet"/>\n  <link href="Speeches_files/style.
        css" media="all" rel="stylesheet"/>\n  <link href="Speeches_files/responsive.css" media="all" rel="stylesheet"/>\n  <link hre
```

3. BeautifulSoup function **find_all()** was used to find the dates for the speeches in the BeautifulSoup object using the tag ‘span’ and stored in variable **ttt**. We see here that the total speeches are 50, as is required.

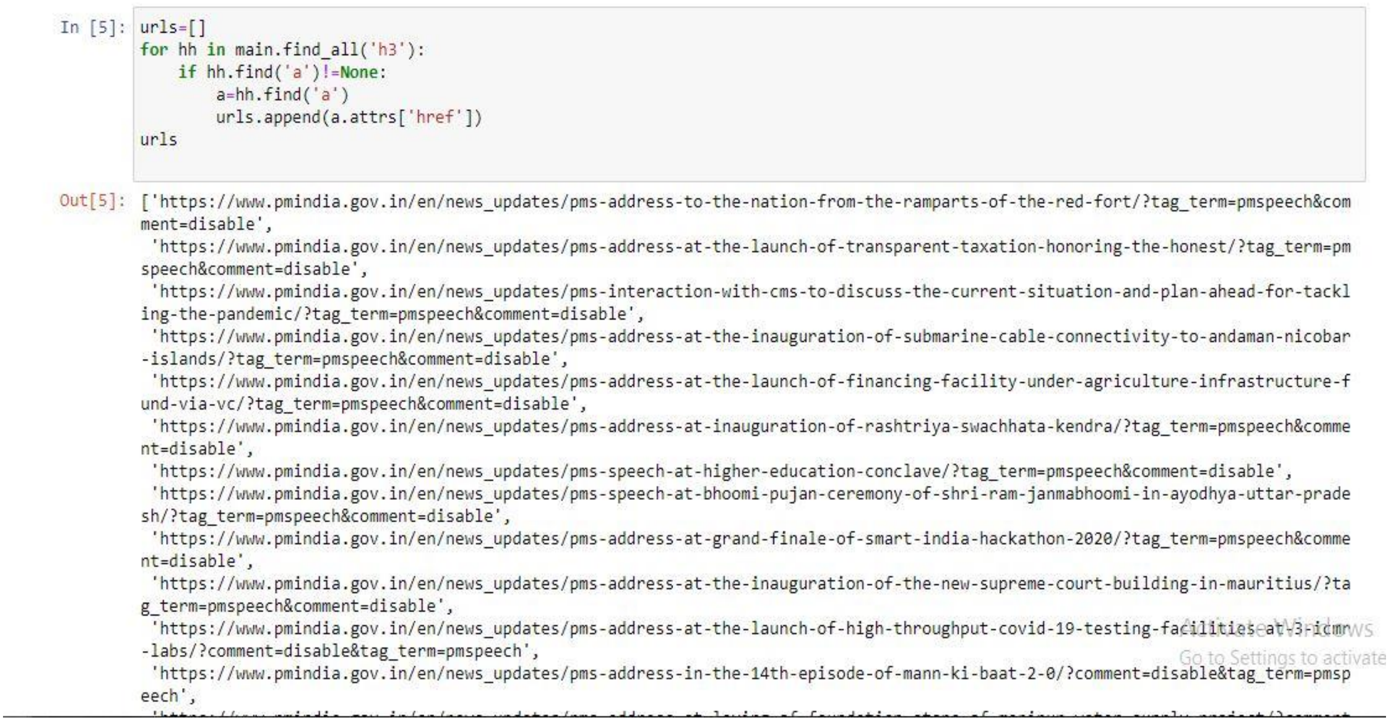


```
In [4]: ttt=main.find_all('span',"date")
        print("Total Speeches : " + str(len(ttt)))
        ttt

        Total Speeches : 50

Out[4]: [<span class="date">Aug 15, 2020</span>,
         <span class="date">Aug 13, 2020</span>,
         <span class="date">Aug 11, 2020</span>,
         <span class="date">Aug 10, 2020</span>,
         <span class="date">Aug 09, 2020</span>,
         <span class="date">Aug 08, 2020</span>,
         <span class="date">Aug 07, 2020</span>,
         <span class="date">Aug 05, 2020</span>,
         <span class="date">Aug 01, 2020</span>,
         <span class="date">Jul 30, 2020</span>,
         <span class="date">Jul 27, 2020</span>,
         <span class="date">Jul 26, 2020</span>,
         <span class="date">Jul 23, 2020</span>,
         <span class="date">Jul 22, 2020</span>,
         <span class="date">Jul 17, 2020</span>,
         <span class="date">Jul 15, 2020</span>,
         <span class="date">Jul 15, 2020</span>,
         <span class="date">Jul 10, 2020</span>,
         <span class="date">Jul 09, 2020</span>,
         <span class="date">Jul 09, 2020</span>,
         <span class="date">Jul 04, 2020</span>,
         <span class="date">Jul 03, 2020</span>,
         <span class="date">Jul 03, 2020</span>,
         <span class="date">Jun 30, 2020</span>,
         <span class="date">Jun 28, 2020</span>,
```

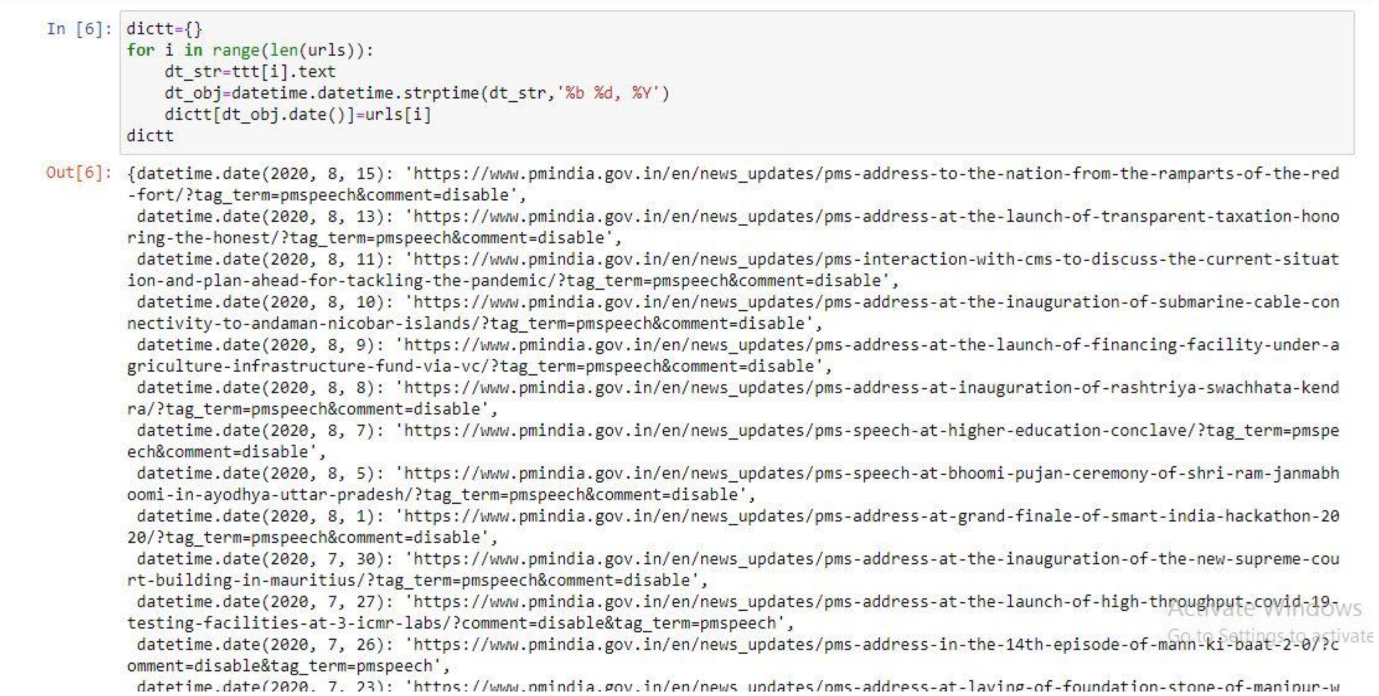
4. The suitable tags in the HTML script for the speech URL’s were determined from careful observation of the script, then URL’s were extracted from the script using the **find_all()** function. The URL’s had the tag ‘a’ under the tag ‘h3’ for speech descriptions. The URL’s were stored in the list variable **urls**.



```
In [5]: urls=[]
        for hh in main.find_all('h3'):
            if hh.find('a')!=None:
                a=hh.find('a')
                urls.append(a.attrs['href'])
        urls
```

```
Out[5]: ['https://www.pmindia.gov.in/en/news_updates/pms-address-to-the-nation-from-the-ramparts-of-the-red-fort/?tag_term=pmspeech&com
        ment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-transparent-taxation-honoring-the-honest/?tag_term=pm
        speech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-interaction-with-cms-to-discuss-the-current-situation-and-plan-ahead-for-tackl
        ing-the-pandemic/?tag_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-inauguration-of-submarine-cable-connectivity-to-andaman-nicobar
        -islands/?tag_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-financing-facility-under-agriculture-infrastructure-f
        und-via-vc/?tag_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-inauguration-of-rashtriya-swachhata-kendra/?tag_term=pmspeech&comme
        nt=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-speech-at-higher-education-conclave/?tag_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-speech-at-bhoomi-pujan-ceremony-of-shri-ram-janmabhoomi-in-ayodhya-uttar-prade
        sh/?tag_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-grand-finale-of-smart-india-hackathon-2020/?tag_term=pmspeech&comme
        nt=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-inauguration-of-the-new-supreme-court-building-in-mauritius/?ta
        g_term=pmspeech&comment=disable',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-high-throughput-covid-19-testing-facilities-at-3-icmr
        -labs/?comment=disable&tag_term=pmspeech',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-in-the-14th-episode-of-mann-ki-baat-2-0/?comment=disable&tag_term=pmsp
        eech',
         'https://www.pmindia.gov.in/en/news_updates/pms-address-at-laying-of-foundation-stone-of-manipur-water-supply-project/?comment
```

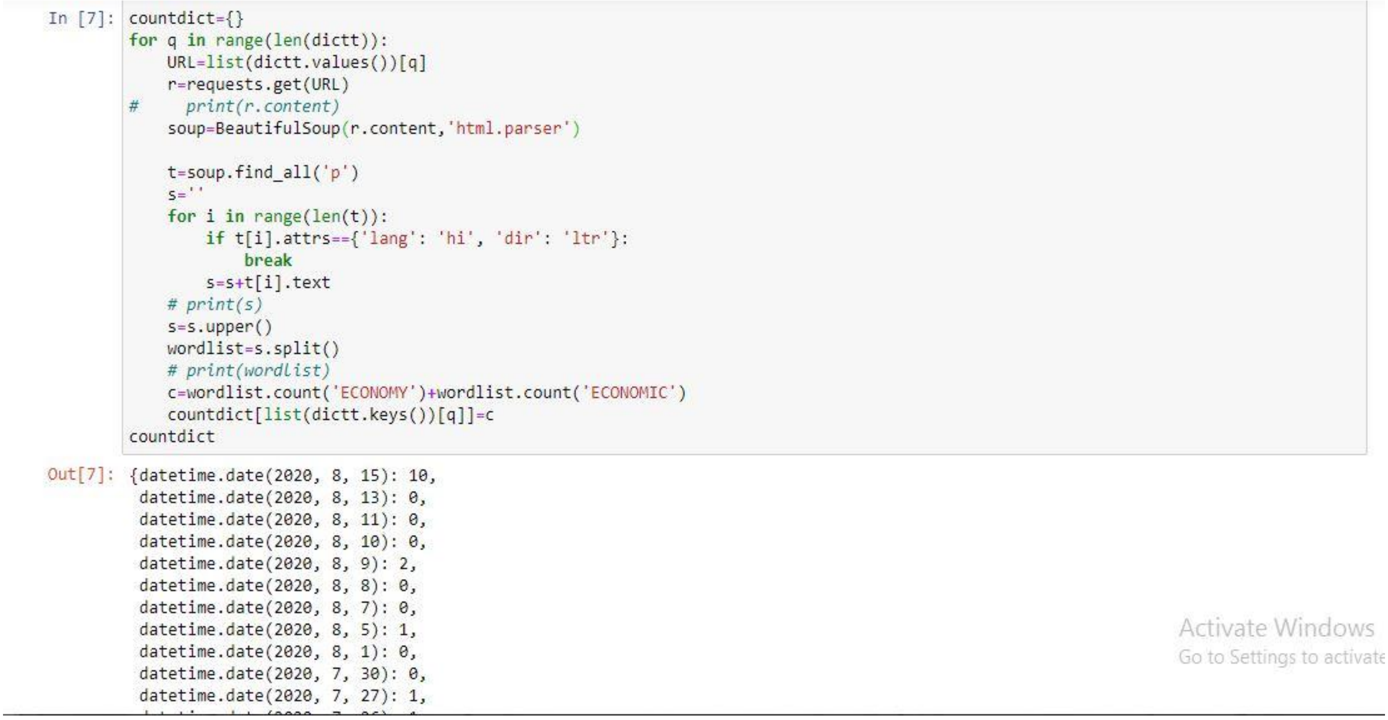
5. A dictionary object **dictt** was created to store the dates of the speeches with their corresponding speech URL’s so that they could be accessed easily to extract speech text. Time extracted previously in text format was converted Python recognizable datetime form.



```
In [6]: dictt={}
        for i in range(len(urls)):
            dt_str=ttt[i].text
            dt_obj=datetime.datetime.strptime(dt_str,'%b %d, %Y')
            dictt[dt_obj.date()]=urls[i]
        dictt
```

```
Out[6]: {datetime.date(2020, 8, 15): 'https://www.pmindia.gov.in/en/news_updates/pms-address-to-the-nation-from-the-ramparts-of-the-red
        -fort/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 13): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-transparent-taxation-hono
        ring-the-honest/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 11): 'https://www.pmindia.gov.in/en/news_updates/pms-interaction-with-cms-to-discuss-the-current-situat
        ion-and-plan-ahead-for-tackling-the-pandemic/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 10): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-inauguration-of-submarine-cable-con
        nectivity-to-andaman-nicobar-islands/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 9): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-financing-facility-under-a
        griculture-infrastructure-fund-via-vc/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 8): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-inauguration-of-rashtriya-swachhata-kend
        ra/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 7): 'https://www.pmindia.gov.in/en/news_updates/pms-speech-at-higher-education-conclave/?tag_term=pmspe
        ech&comment=disable',
         datetime.date(2020, 8, 5): 'https://www.pmindia.gov.in/en/news_updates/pms-speech-at-bhoomi-pujan-ceremony-of-shri-ram-janmabh
        oomi-in-ayodhya-uttar-pradesh/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 8, 1): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-grand-finale-of-smart-india-hackathon-20
        20/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 7, 30): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-inauguration-of-the-new-supreme-cou
        rt-building-in-mauritius/?tag_term=pmspeech&comment=disable',
         datetime.date(2020, 7, 27): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-the-launch-of-high-throughput-covid-19-
        testing-facilities-at-3-icmr-labs/?comment=disable&tag_term=pmspeech',
         datetime.date(2020, 7, 26): 'https://www.pmindia.gov.in/en/news_updates/pms-address-in-the-14th-episode-of-mann-ki-baat-2-0/?c
        omment=disable&tag_term=pmspeech',
         datetime.date(2020, 7, 23): 'https://www.pmindia.gov.in/en/news_updates/pms-address-at-laying-of-foundation-stone-of-manipur-w
```

6. A for loop was run to access one by one the URL's stored as values in the dictionary **dictt**. Inside each loop, a BeautifulSoup object was created to parse the HTML script of each of the speech URL's. After determining the tags for the speech body, the main speech body was extracted and stored in a separate String variable **s.** The number of times the word 'Economy' or 'Economic' appeared in the String **s** was counted and stored in a separate dictionary **countdict** with each key being the speech date and its corresponding value being the count of 'Economy' or 'Economic' for that speech. The final **countdict** dictionary after loop finished executing had date-count pairs for all 50 speeches.

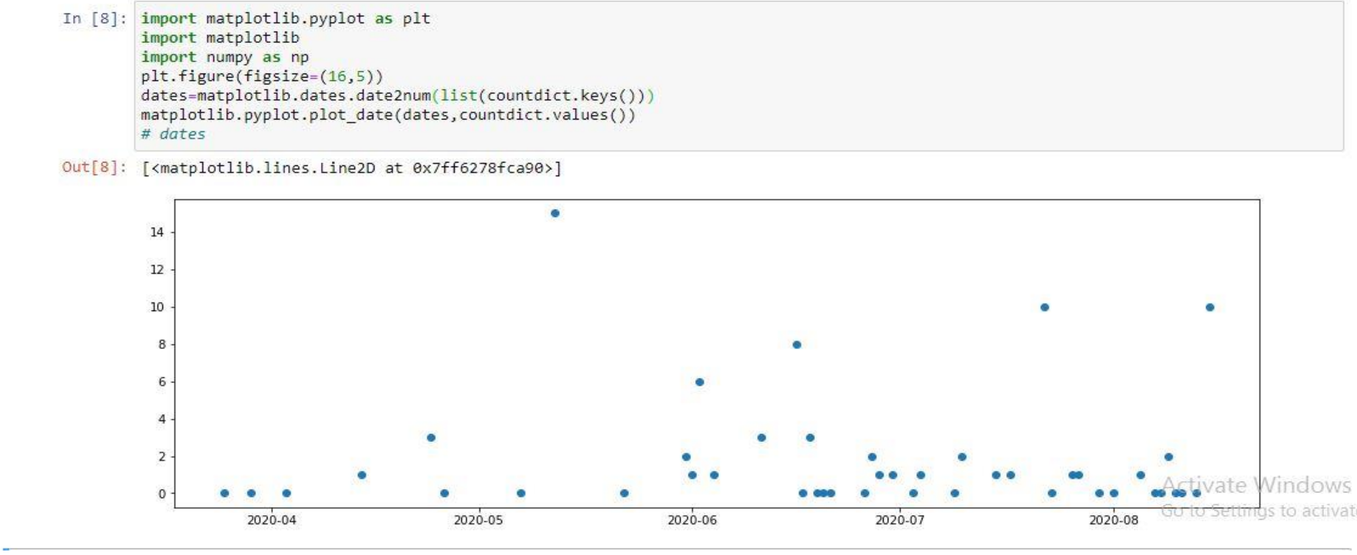


```
In [7]: countdict={}
        for q in range(len(dictt)):
            URL=list(dictt.values())[q]
            r=requests.get(URL)
        #     print(r.content)
            soup=BeautifulSoup(r.content,'html.parser')

            t=soup.find_all('p')
            s=''
            for i in range(len(t)):
                if t[i].attrs=={'lang': 'hi', 'dir': 'ltr'}:
                    break
                s=s+t[i].text
            # print(s)
            s=s.upper()
            wordlist=s.split()
            # print(wordlist)
            c=wordlist.count('ECONOMY')+wordlist.count('ECONOMIC')
            countdict[list(dictt.keys())[q]]=c
        countdict

Out[7]: {datetime.date(2020, 8, 15): 10,
         datetime.date(2020, 8, 13): 0,
         datetime.date(2020, 8, 11): 0,
         datetime.date(2020, 8, 10): 0,
         datetime.date(2020, 8, 9): 2,
         datetime.date(2020, 8, 8): 0,
         datetime.date(2020, 8, 7): 0,
         datetime.date(2020, 8, 5): 1,
         datetime.date(2020, 8, 1): 0,
         datetime.date(2020, 7, 30): 0,
         datetime.date(2020, 7, 27): 1,
```

7. Module **matplotlib** for plotting was imported. Dates present as keys in dictionary **countdict** was converted to Python timestamp. Matplotlib function **plot_date()** was used to plot the counts (present as values in dictionary **countdict**) against the speech dates.

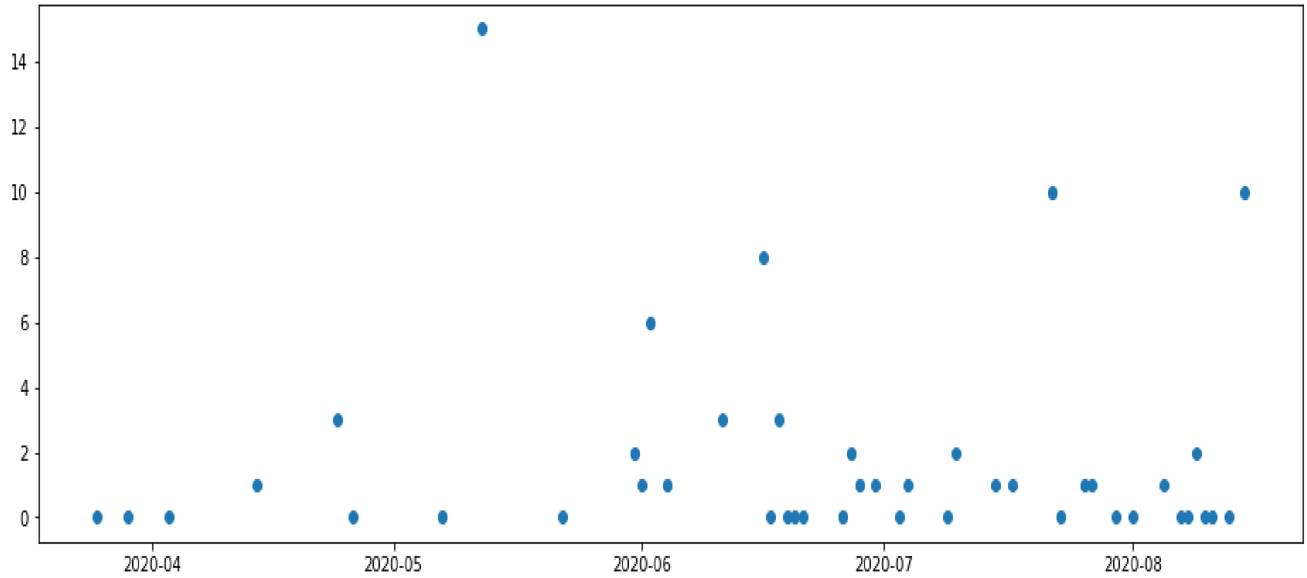


```
In [8]: import matplotlib.pyplot as plt
        import matplotlib
        import numpy as np
        plt.figure(figsize=(16,5))
        dates=matplotlib.dates.date2num(list(countdict.keys()))
        matplotlib.pyplot.plot_date(dates,countdict.values())
        # dates

Out[8]: [<matplotlib.lines.Line2D at 0x7ff6278fca90>]
```

Observations:

The final plot was obtained is shown below:



1. We see that the count for ‘Economy’ is quite low in the months of April and May. In one speech it was used 3 times and in another once. Another speech, which stands out, has a count of >14, which means economy was especially stressed on in this particular speech by the PM.
2. The count is slightly greater in the month of June, being used 1-2 times in 3 speeches and >4 times in 2 speeches.
3. For the months of July and August, the counts are comparatively much higher. We see more speeches with counts in the range 2-4 in the month of July, although none of the counts are especially high, which means only passing references were made to the Economy. In the month of August, we have approximately 4 speeches with a count of 2-4. We have 2 speeches with counts of 10, which signifies the Economy was specially addressed by the PM in these 2 speeches.

Conclusion:

From April to August, we see an upward rise in the number of references to the Economy made by the PM in his speeches, with a significant rise in the months of July and August. This makes sense as those 2 months correspond with the lifting of the lockdown and a special emphasis on rebuilding the Economy of the nation and communicating the measures being taken to the citizens.