WARM UP PROJECT (TASK 2)

Question:

The word cloud for Top 5 news category:

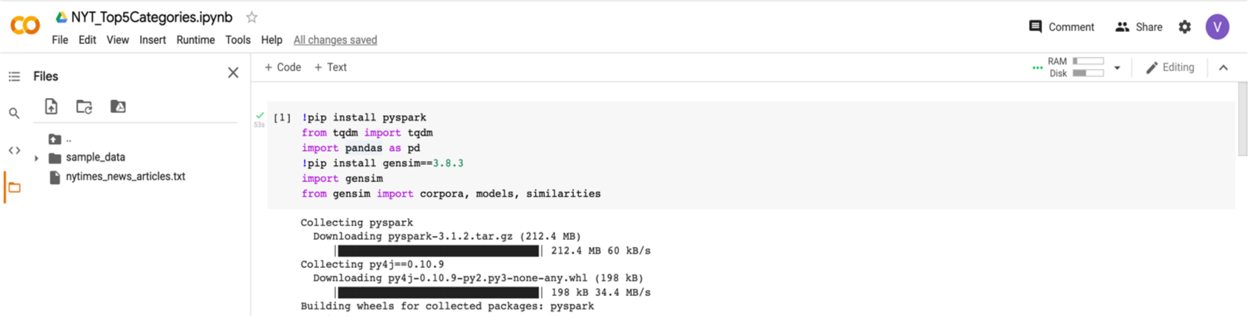
- Category that has the most of news articles
- URL contains the category which the new belong

For example:-

- -/us/politics/
- -/world/
- -/sports/
- -/arts/
- -and so on

Step 1:

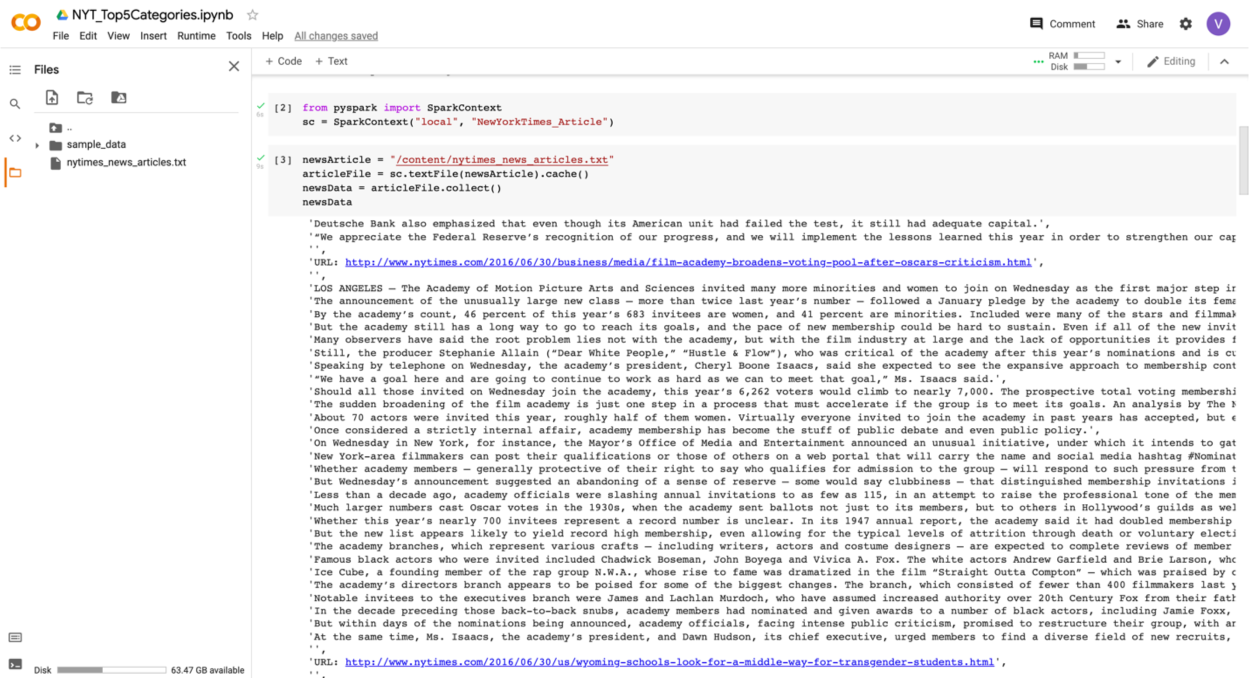
1. Install and import all libraries
   I have used pyspark, tqdm, pandas, genism



```
[1] !pip install pyspark
    from tqdm import tqdm
    import pandas as pd
    !pip install gensim==3.8.3
    import gensim
    from gensim import corpora, models, similarities

    Collecting pyspark
      Downloading pyspark-3.1.2.tar.gz (212.4 MB)
         |████████████████████████████████| 212.4 MB 60 kB/s
    Collecting py4j==0.10.9
      Downloading py4j-0.10.9-py2.py3-none-any.whl (198 kB)
         |████████████████████████████████| 198 kB 34.4 MB/s
    Building wheels for collected packages: pyspark
```

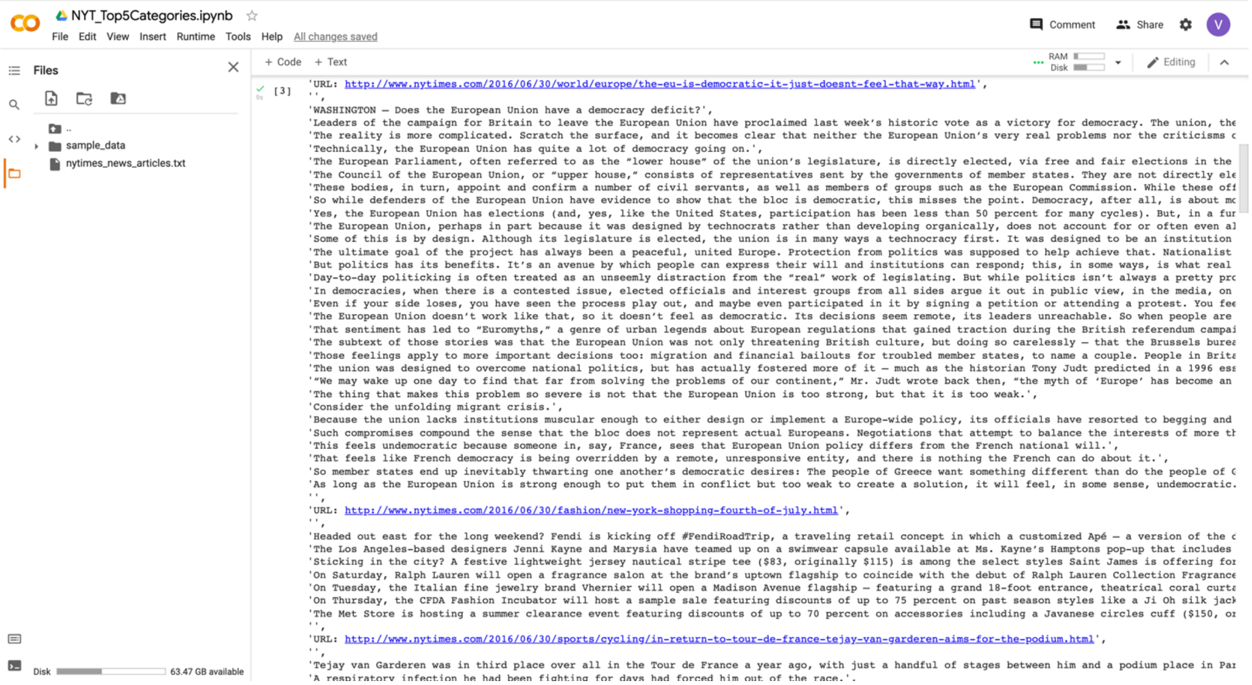
Step 2:

1. Import Sparkcontext and read the textfile (news article)
   'content/nytimes.new_articles.txt' → sc.textFile(newsarticle) using sparkcontext



```
[2] from pyspark import SparkContext
    sc = SparkContext("local", "NewYorkTimes_Article")

[3] newsArticle = "/content/nytimes_news_articles.txt"
    articleFile = sc.textFile(newsArticle).cache()
    newsData = articleFile.collect()
    newsData

     'Deutsche Bank also emphasized that even though its American unit had failed the test, it still had adequate capital.',
     '“We appreciate the Federal Reserve’s recognition of our progress, and we will implement the lessons learned this year in order to strengthen our cap
     '',
     'URL: http://www.nytimes.com/2016/06/30/business/media/film-academy-broadens-voting-pool-after-oscars-criticism.html',
     '',
     'LOS ANGELES — The Academy of Motion Picture Arts and Sciences invited many more minorities and women to join on Wednesday as the first major step ir
     'The announcement of the unusually large new class — more than twice last year’s number — followed a January pledge by the academy to double its fema
     'By the academy’s count, 46 percent of this year’s 683 invitees are women, and 41 percent are minorities. Included were many of the stars and filmmak
     'But the academy still has a long way to go to reach its goals, and the pace of new membership could be hard to sustain. Even if all of the new invit
     'Many observers have said the root problem lies not with the academy, but with the film industry at large and the lack of opportunities it provides f
     'Still, the producer Stephanie Allain (“Dear White People,” “Hustle & Flow”), who was critical of the academy after this year’s nominations and is cu
     'Speaking by telephone on Wednesday, the academy’s president, Cheryl Boone Isaacs, said she expected to see the expansive approach to membership cont
     '“We have a goal here and are going to continue to work as hard as we can to meet that goal,” Ms. Isaacs said.',
     'Should all those invited on Wednesday join the academy, this year’s 6,262 voters would climb to nearly 7,000. The prospective total voting membershi
     'The sudden broadening of the film academy is just one step in a process that must accelerate if the group is to meet its goals. An analysis by The N
     'About 70 actors were invited this year, roughly half of them women. Virtually everyone invited to join the academy in past years has accepted, but e
     'Once considered a strictly internal affair, academy membership has become the stuff of public debate and even public policy.',
     'On Wednesday in New York, for instance, the Mayor’s Office of Media and Entertainment announced an unusual initiative, under which it intends to gat
     'New York-area filmmakers can post their qualifications or those of others on a web portal that will carry the name and social media hashtag #Nominat
     'Whether academy members — generally protective of their right to say who qualifies for admission to the group — will respond to such pressure from t
     'But Wednesday’s announcement suggested an abandoning of a sense of reserve — some would say clubbiness — that distinguished membership invitations i
     'Less than a decade ago, academy officials were slashing annual invitations to as few as 115, in an attempt to raise the professional tone of the mem
     'Much larger numbers cast Oscar votes in the 1930s, when the academy sent ballots not just to its members, but to others in Hollywood’s guilds as wel
     'Whether this year’s nearly 700 invitees represent a record number is unclear. In its 1947 annual report, the academy said it had doubled membership
     'But the new list appears likely to yield record high membership, even allowing for the typical levels of attrition through death or voluntary electi
     'The academy branches, which represent various crafts — including writers, actors and costume designers — are expected to complete reviews of member
     'Famous black actors who were invited included Chadwick Boseman, John Boyega and Vivica A. Fox. The white actors Andrew Garfield and Brie Larson, who
     'Ice Cube, a founding member of the rap group N.W.A., whose rise to fame was dramatized in the film “Straight Outta Compton” — which was praised by c
     'The academy’s directors branch appears to be poised for some of the biggest changes. The branch, which consisted of fewer than 400 filmmakers last y
     'Notable invitees to the executives branch were James and Lachlan Murdoch, who have assumed increased authority over 20th Century Fox from their fath
     'In the decade preceding those back-to-back snubs, academy members had nominated and given awards to a number of black actors, including Jamie Foxx,
     'But within days of the nominations being announced, academy officials, facing intense public criticism, promised to restructure their group, with ar
     'At the same time, Ms. Isaacs, the academy’s president, and Dawn Hudson, its chief executive, urged members to find a diverse field of new recruits,
     '',
     'URL: http://www.nytimes.com/2016/06/30/us/wyoming-schools-look-for-a-middle-way-for-transgender-students.html',
     '',
```

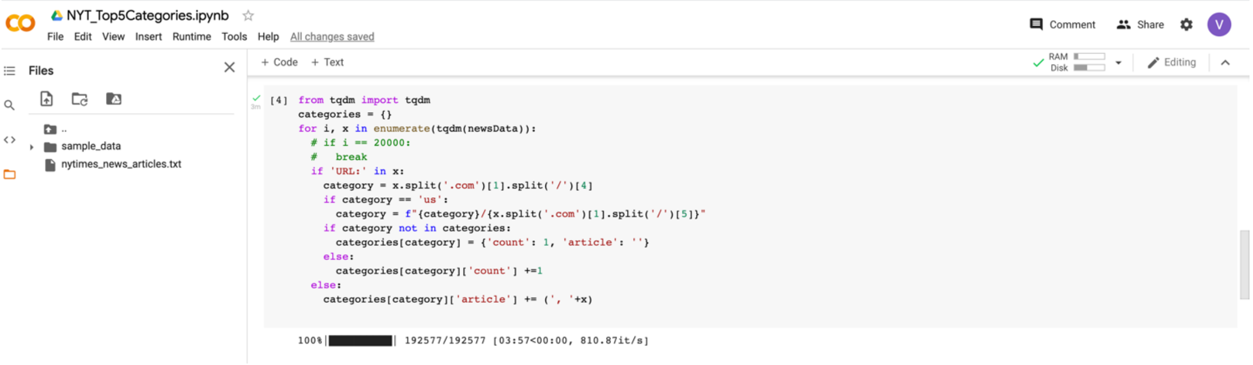
2. Ouptut of newsData collection



```
[3] 'URL: http://www.nytimes.com/2016/06/30/world/europe/the-eu-is-democratic-it-just-doesnt-feel-that-way.html',
    '',
    'WASHINGTON — Does the European Union have a democracy deficit?',
    'Leaders of the campaign for Britain to leave the European Union have proclaimed last week’s historic vote as a victory for democracy. The union, the
    'The reality is more complicated. Scratch the surface, and it becomes clear that neither the European Union’s very real problems nor the criticisms c
    'Technically, the European Union has quite a lot of democracy going on.',
    'The European Parliament, often referred to as the “lower house” of the union’s legislature, is directly elected, via free and fair elections in the
    'The Council of the European Union, or “upper house,” consists of representatives sent by the governments of member states. They are not directly ele
    'These bodies, in turn, appoint and confirm a number of civil servants, as well as members of groups such as the European Commission. While these offi
    'So while defenders of the European Union have evidence to show that the bloc is democratic, this misses the point. Democracy, after all, is about mc
    'Yes, the European Union has elections (and, yes, like the United States, participation has been less than 50 percent for many cycles). But, in a fur
    'The European Union, perhaps in part because it was designed by technocrats rather than developing organically, does not account for or often even al
    'Some of this is by design. Although its legislature is elected, the union is in many ways a technocracy first. It was designed to be an institution
    'The ultimate goal of the project has always been a peaceful, united Europe. Protection from politics was supposed to help achieve that. Nationalist
    'But politics has its benefits. It’s an avenue by which people can express their will and institutions can respond; this, in some ways, is what real
    'Day-to-day politicking is often treated as an unseemly distraction from the “real” work of legislating. But while politics isn’t always a pretty prc
    'In democracies, when there is a contested issue, elected officials and interest groups from all sides argue it out in public view, in the media, on
    'Even if your side loses, you have seen the process play out, and maybe even participated in it by signing a petition or attending a protest. You fee
    'The European Union doesn’t work like that, so it doesn’t feel as democratic. Its decisions seem remote, its leaders unreachable. So when people are
    'That sentiment has led to “Euromyths,” a genre of urban legends about European regulations that gained traction during the British referendum campai
    'The subtext of those stories was that the European Union was not only threatening British culture, but doing so carelessly — that the Brussels burea
    'Those feelings apply to more important decisions too: migration and financial bailouts for troubled member states, to name a couple. People in Brita
    'The union was designed to overcome national politics, but has actually fostered more of it — much as the historian Tony Judt predicted in a 1996 ess
    '“We may wake up one day to find that far from solving the problems of our continent,” Mr. Judt wrote back then, “the myth of ‘Europe’ has become an
    'The thing that makes this problem so severe is not that the European Union is too strong, but that it is too weak.',
    'Consider the unfolding migrant crisis.',
    'Because the union lacks institutions muscular enough to either design or implement a Europe-wide policy, its officials have resorted to begging and
    'Such compromises compound the sense that the bloc does not represent actual Europeans. Negotiations that attempt to balance the interests of more th
    'This feels undemocratic because someone in, say, France, sees that European Union policy differs from the French national will.',
    'That feels like French democracy is being overridden by a remote, unresponsive entity, and there is nothing the French can do about it.',
    'So member states end up inevitably thwarting one another’s democratic desires: The people of Greece want something different than do the people of G
    'As long as the European Union is strong enough to put them in conflict but too weak to create a solution, it will feel, in some sense, undemocratic.
    '',
    'URL: http://www.nytimes.com/2016/06/30/fashion/new-york-shopping-fourth-of-july.html',
    '',
    'Headed out east for the long weekend? Fendi is kicking off #FendiRoadTrip, a traveling retail concept in which a customized Apé — a version of the c
    'The Los Angeles-based designers Jenni Kayne and Marysia have teamed up on a swimwear capsule available at Ms. Kayne’s Hamptons pop-up that includes
    'Sticking in the city? A festive lightweight jersey nautical stripe tee ($83, originally $115) is among the select styles Saint James is offering for
    'On Saturday, Ralph Lauren will open a fragrance salon at the brand’s uptown flagship to coincide with the debut of Ralph Lauren Collection Fragrance
    'On Tuesday, the Italian fine jewelry brand Vhernier will open a Madison Avenue flagship — featuring a grand 18-foot entrance, theatrical coral curta
    'On Thursday, the CFDA Fashion Incubator will host a sample sale featuring discounts of up to 75 percent on past season styles like a Ji Oh silk jack
    'The Met Store is hosting a summer clearance event featuring discounts of up to 70 percent on accessories including a Javanese circles cuff ($150, or
    '',
    'URL: http://www.nytimes.com/2016/06/30/sports/cycling/in-return-to-tour-de-france-tejay-van-garderen-aims-for-the-podium.html',
    '',
    'Tejay van Garderen was in third place over all in the Tour de France a year ago, with just a handful of stages between him and a podium place in Par
    'A respiratory infection he had been fighting for days had forced him out of the race.',
```

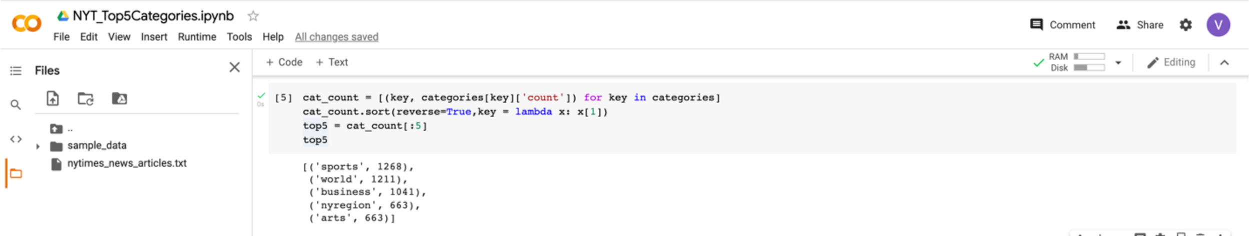
Step 3: TOP 5 Catogories

1. Extracting Top 5 categories based on the URL string
2. For each value in newsdata → If ‘URL: ’ is present then split the url with ‘/’ position 4 to extract words
3. Based on this 4th position value assigned the dictionary values in categories as word count and its corresponding article


NYT_Top5Categories.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
sample_data
nytimes_news_articles.txt
+ Code + Text
RAM Disk
Editing

```
[4] from tqdm import tqdm
    categories = {}
    for i, x in enumerate(tqdm(newsData)):
      # if i == 20000:
      #   break
      if 'URL:' in x:
        category = x.split('.com')[1].split('/')[4]
        if category == 'us':
          category = f"{category}/{x.split('.com')[1].split('/')[5]}"
        if category not in categories:
          categories[category] = {'count': 1, 'article': ''}
        else:
          categories[category]['count'] +=1
      else:
        categories[category]['article'] += (', '+x)

    100%|██████████| 192577/192577 [03:57<00:00, 810.87it/s]
```

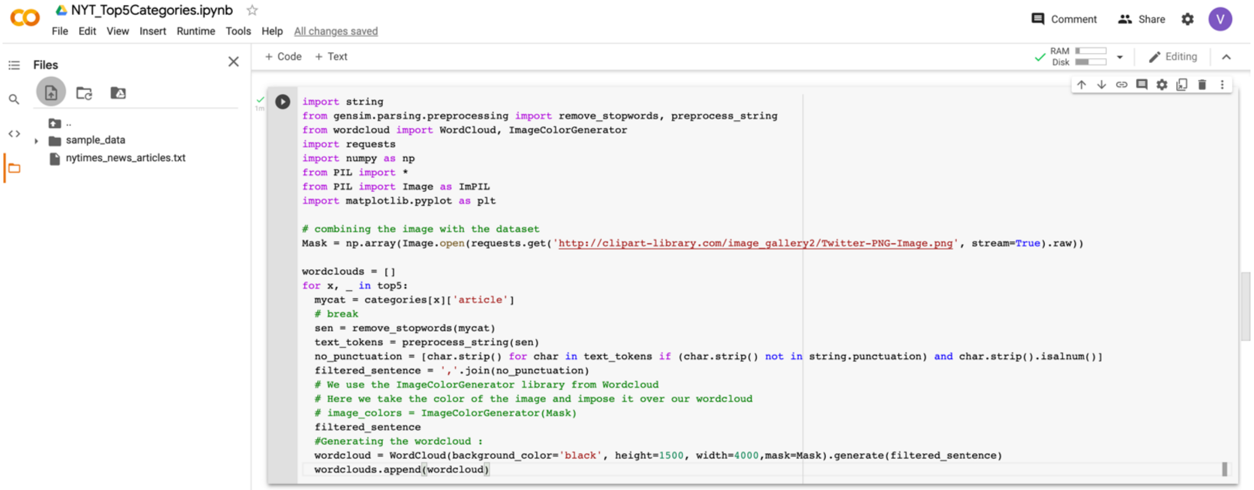
4. With this key-value pair (‘count’, ‘article’), with the help of reverse sort I got Top 5 categories which have highest value of frequency count.
5. I got ‘sports’ → 1268 , ‘world’ → 1211 , ‘business’ → 1041 , ‘nyregion’ → 663, ‘arts’ → 663


NYT_Top5Categories.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
Files
sample_data
nytimes_news_articles.txt
+ Code + Text
RAM Disk
Editing

```
[5] cat_count = [(key, categories[key]['count']) for key in categories]
    cat_count.sort(reverse=True,key = lambda x: x[1])
    top5 = cat_count[:5]
    top5

    [('sports', 1268),
     ('world', 1211),
     ('business', 1041),
     ('nyregion', 663),
     ('arts', 663)]
```

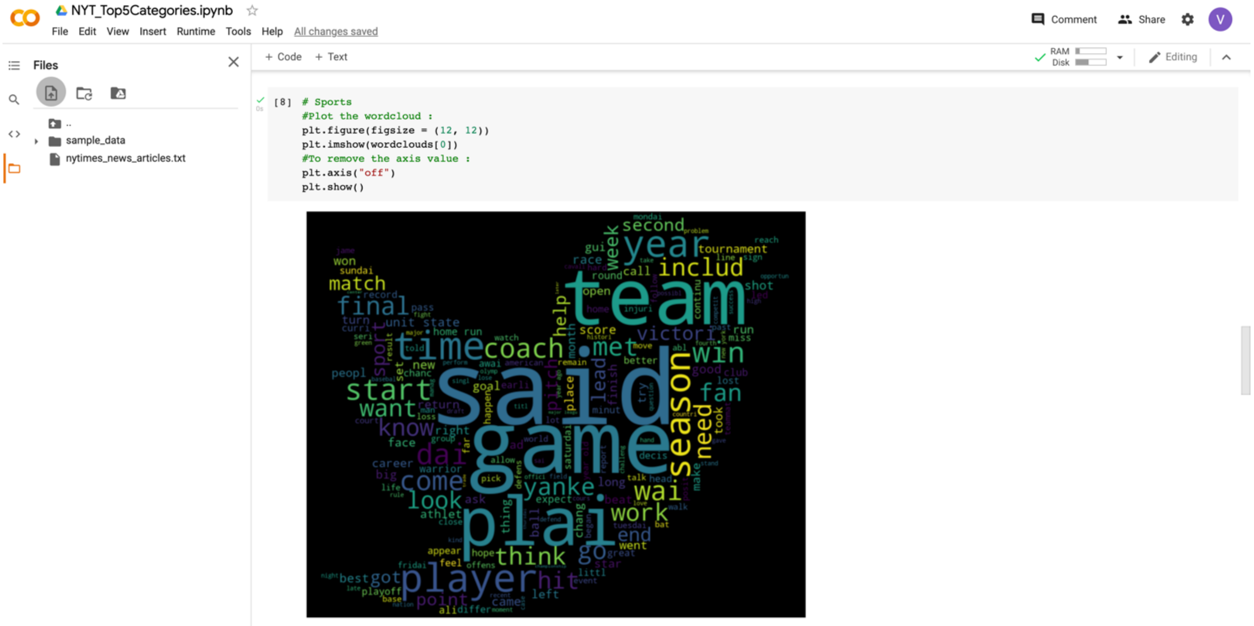
Step 4: WordCloud of Top 5 Categories Articles

1. Generated word cloud based on the top 5 category article using matplotlib and wordcloud
2. Preprocessed the text file → cleaned the text by removing stopwords and getting tokenized words
3. Removing special characters and punctuations
4. Plot Word Cloud


NYT_Top5Categories.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
V
Files
+ Code + Text
RAM
Disk
Editing
..
sample_data
nytimes_news_articles.txt

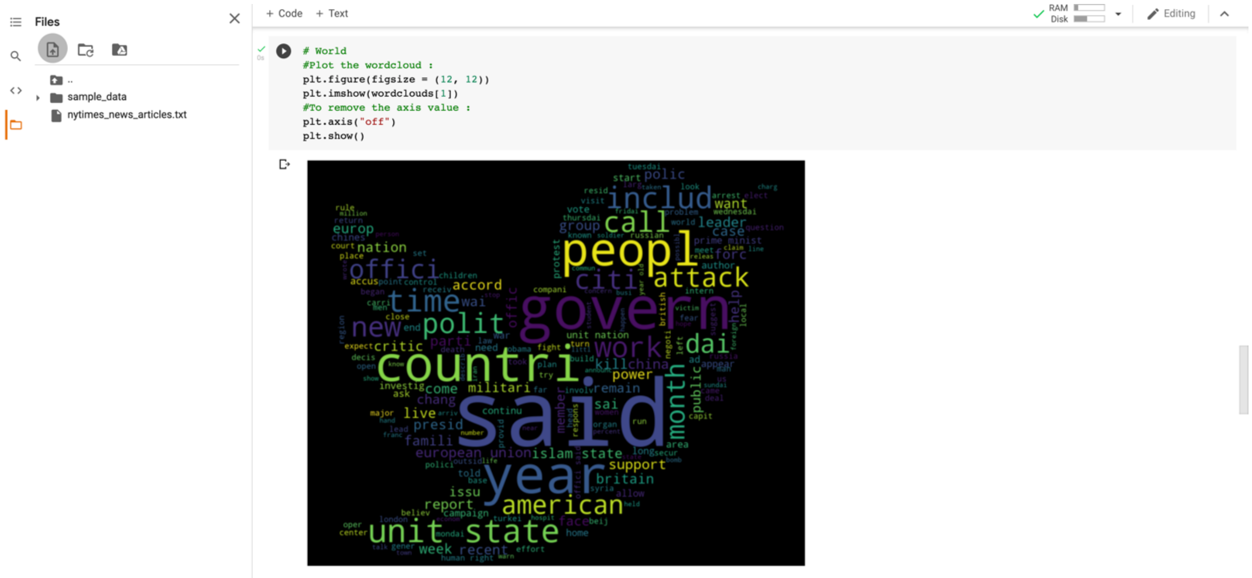
```
import string
from gensim.parsing.preprocessing import remove_stopwords, preprocess_string
from wordcloud import WordCloud, ImageColorGenerator
import requests
import numpy as np
from PIL import *
from PIL import Image as ImPIL
import matplotlib.pyplot as plt

# combining the image with the dataset
Mask = np.array(Image.open(requests.get('http://clipart-library.com/image_gallery2/Twitter-PNG-Image.png', stream=True).raw))

wordclouds = []
for x, _ in top5:
  mycat = categories[x]['article']
  # break
  sen = remove_stopwords(mycat)
  text_tokens = preprocess_string(sen)
  no_punctuation = [char.strip() for char in text_tokens if (char.strip() not in string.punctuation) and char.strip().isalnum()]
  filtered_sentence = ','.join(no_punctuation)
  # We use the ImageColorGenerator library from Wordcloud
  # Here we take the color of the image and impose it over our wordcloud
  # image_colors = ImageColorGenerator(Mask)
  filtered_sentence
  #Generating the wordcloud :
  wordcloud = WordCloud(background_color='black', height=1500, width=4000,mask=Mask).generate(filtered_sentence)
  wordclouds.append(wordcloud)
```

WordCloud of SPORTS



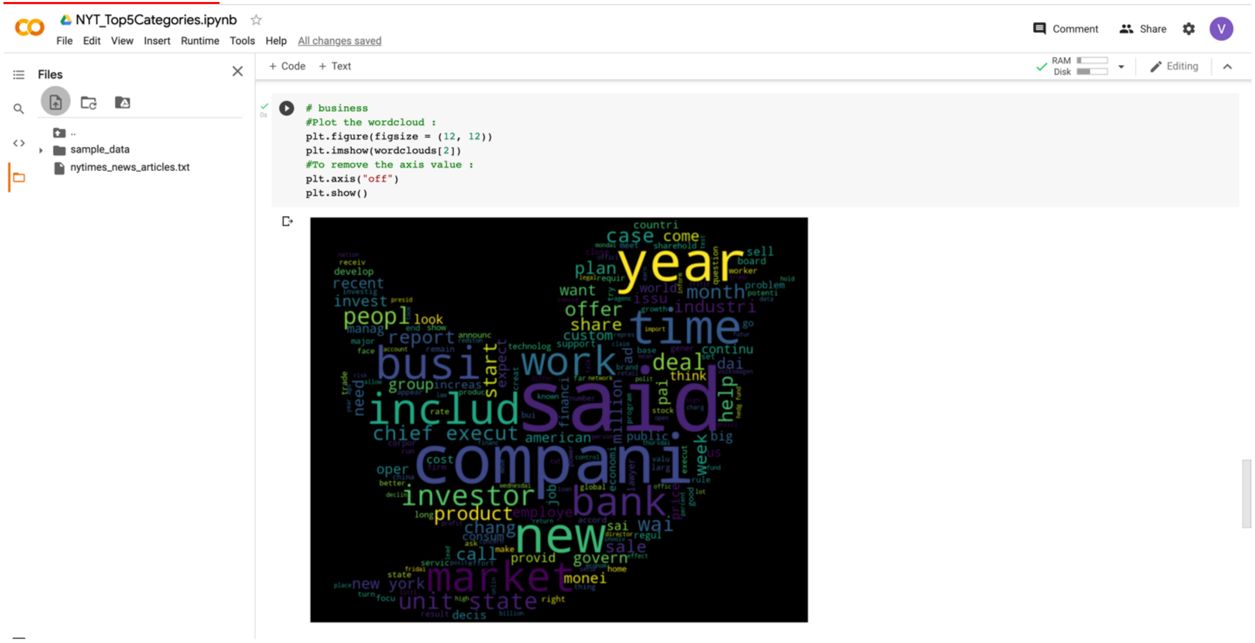
```
[8] # Sports
    #Plot the wordcloud :
    plt.figure(figsize = (12, 12))
    plt.imshow(wordclouds[0])
    #To remove the axis value :
    plt.axis("off")
    plt.show()
```

WordCloud of World



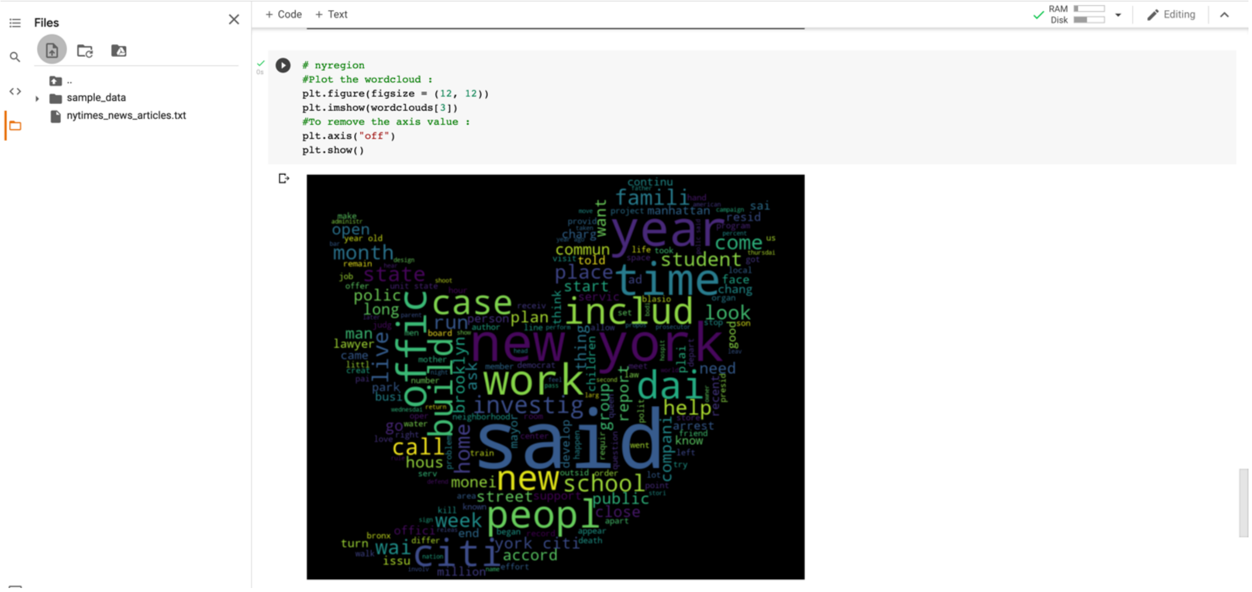
```
# World
#Plot the wordcloud :
plt.figure(figsize = (12, 12))
plt.imshow(wordclouds[1])
#To remove the axis value :
plt.axis("off")
plt.show()
```

WordCloud of Business



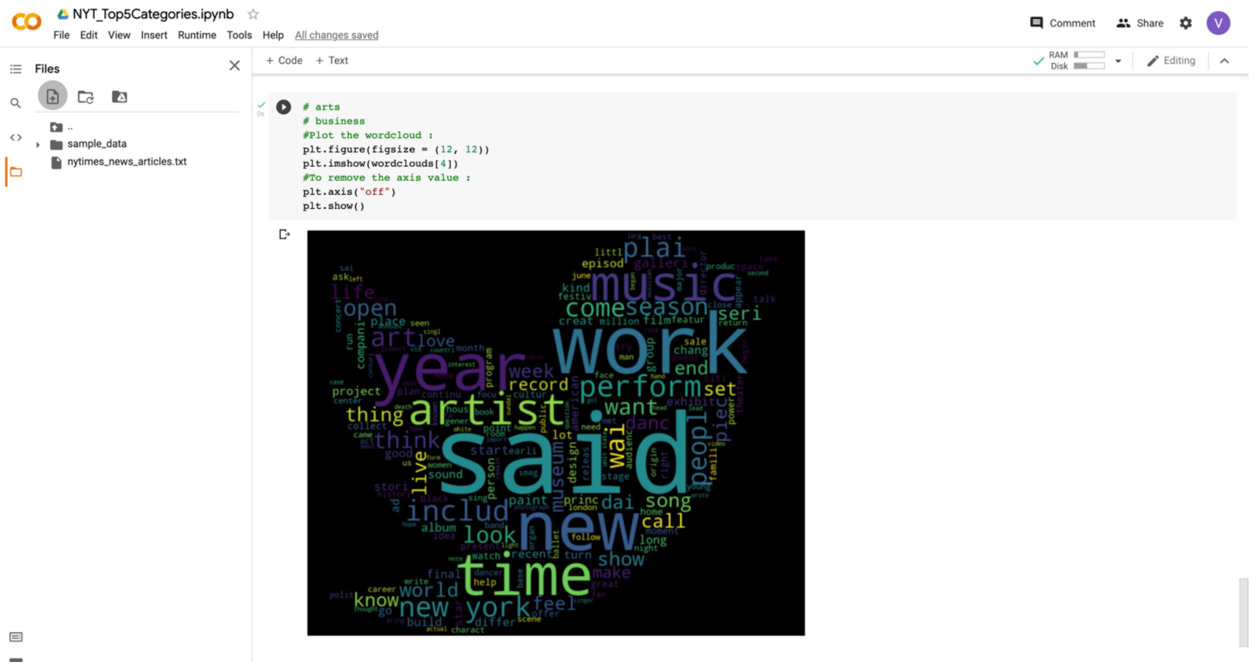
```
# business
#Plot the wordcloud :
plt.figure(figsize = (12, 12))
plt.imshow(wordclouds[2])
#To remove the axis value :
plt.axis("off")
plt.show()
```

WordCloud of nyregion



```
# nyregion
#Plot the wordcloud :
plt.figure(figsize = (12, 12))
plt.imshow(wordclouds[3])
#To remove the axis value :
plt.axis("off")
plt.show()
```

WordCloud of arts



```
# arts
# business
#Plot the wordcloud :
plt.figure(figsize = (12, 12))
plt.imshow(wordclouds[4])
#To remove the axis value :
plt.axis("off")
plt.show()
```