ANISH ADNANI

D17B / 01

NLP

(1)

EXPERIMENT 10

Aim: Exploratory data analysis of a given text (word cloud) using Python and colab notebook

Theory:

What is exploratory data analysis?

Exploratory data analysis or (EDA) is understanding the datasets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data in order to apply machine learning. Plotting in EDA consists of histograms, box plots, scatter plot and many more. It often takes much time to explore the data. Though the process of EDA, we can ask to define problem statement or definition on our dataset which is very important. EDA is primarily used to see when data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationship between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

- why is exploratory data analysis is important?

The main purpose of EDA is to help look at data before making any assumption. It can help identify obvious errors, as well as better understands patterns within the data, detects outliers or anomalous events, find interesting relations among the variables

EDA also helps stakeholders by continuing they are asking the right questions. EDA can help answer questions about standard deviations categorical variables, and confidence intervals

- How to perform exploratory data analysis?

This is one such question that everyone is keen on knowing the answer. well, the answer it depends on the dataset that you are working. There is no single method or common method in order to perform EDA, whereas in there is some common methods and plots that would be used in the EDA process

- Word cloud:

word cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. significant textual data points can be highlighted using a word cloud. word clouds are widely used for analyzing data from social networking websites.

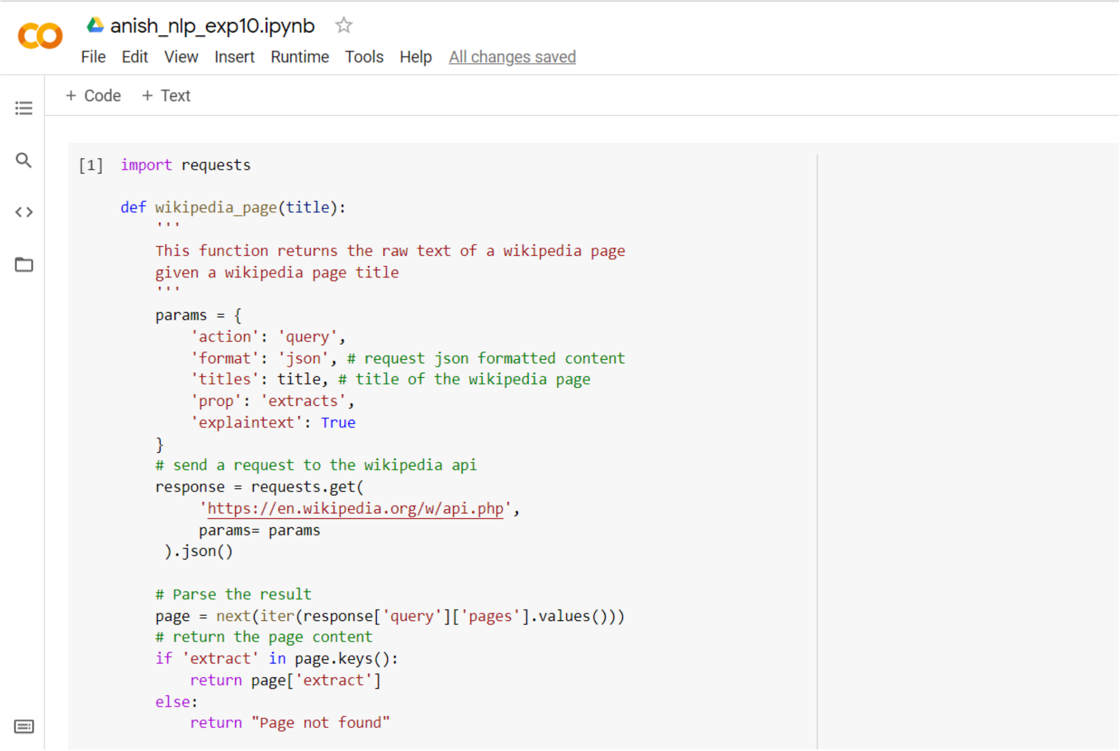
- Advantage of word cloud
- Analyzing customer and employee feedback
- Identifying new SEO keywords to target

- Disadvantage of word cloud.
- word clouds are not perfect for every situation
- Data should be optimized for context

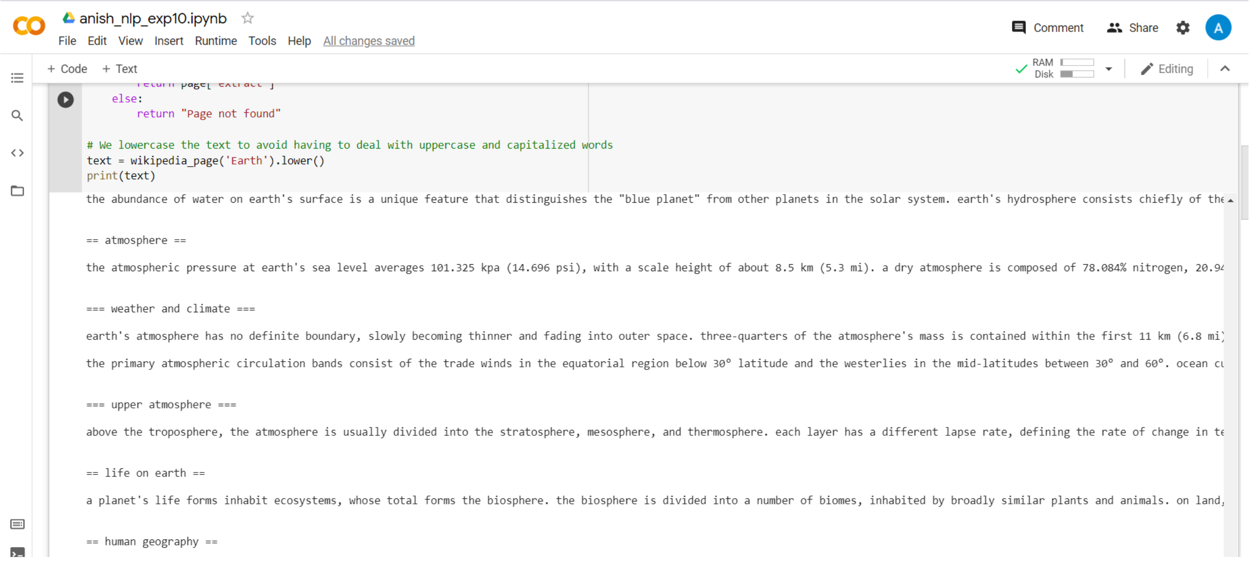
Conclusion:

The exploratory data analysis of a given text using python and colab notebook has been implemented successfully.

Word cloud in google colab

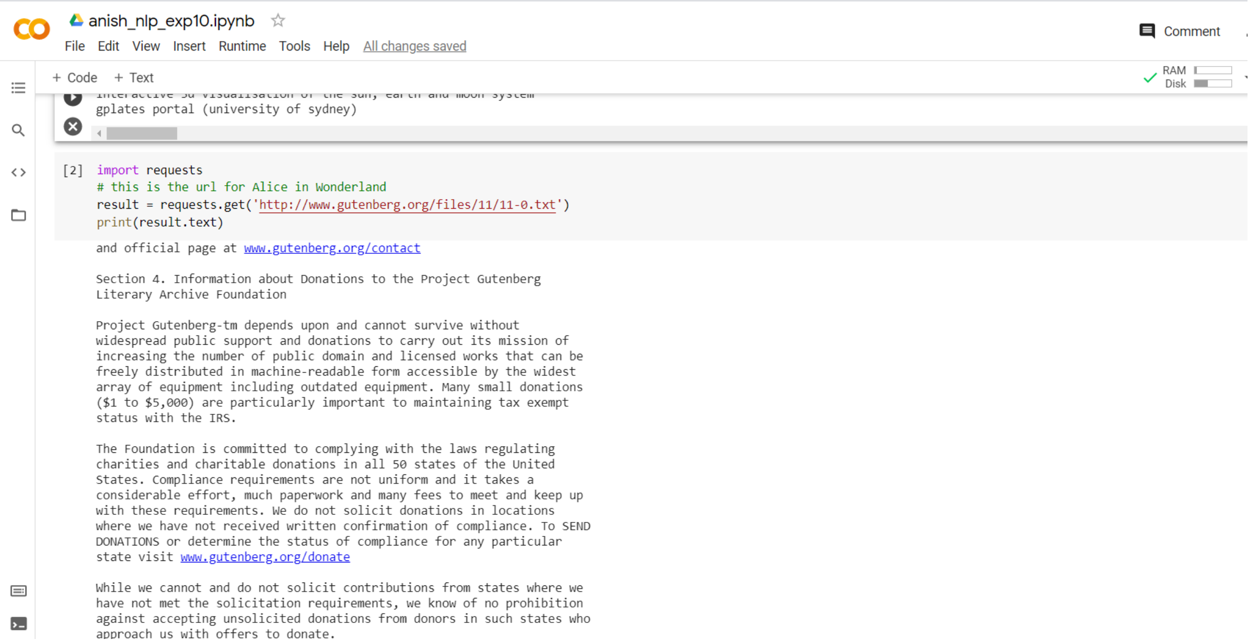


```
[1] import requests

    def wikipedia_page(title):
        '''
        This function returns the raw text of a wikipedia page
        given a wikipedia page title
        '''
        params = {
            'action': 'query',
            'format': 'json', # request json formatted content
            'titles': title, # title of the wikipedia page
            'prop': 'extracts',
            'explaintext': True
        }
        # send a request to the wikipedia api
        response = requests.get(
             'https://en.wikipedia.org/w/api.php',
             params= params
         ).json()

        # Parse the result
        page = next(iter(response['query']['pages'].values()))
        # return the page content
        if 'extract' in page.keys():
            return page['extract']
        else:
            return "Page not found"
```



```
        return page['extract']
    else:
        return "Page not found"

# We lowercase the text to avoid having to deal with uppercase and capitalized words
text = wikipedia_page('Earth').lower()
print(text)
```

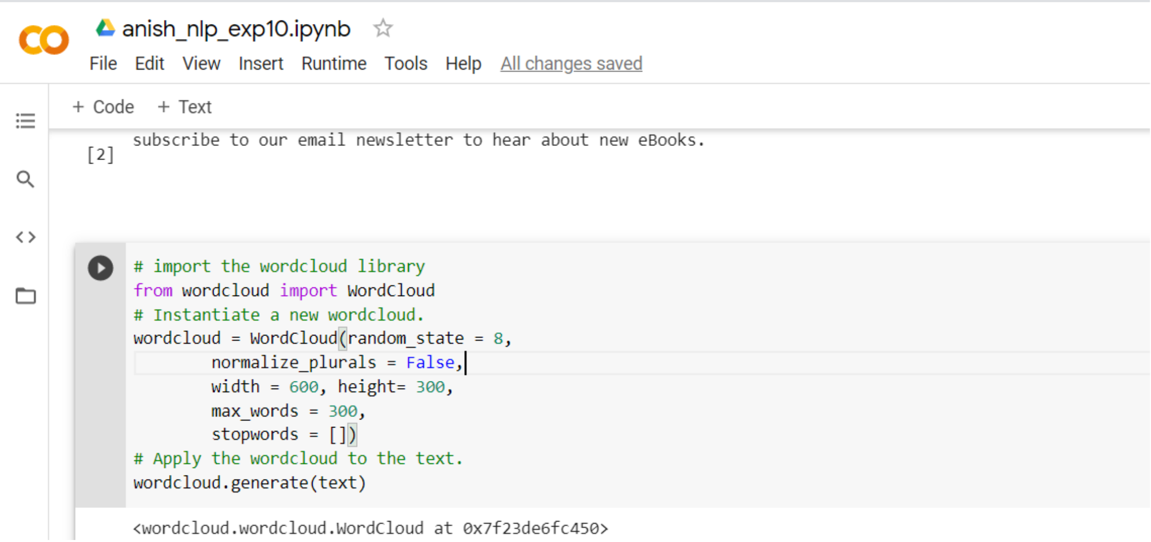
```
the abundance of water on earth's surface is a unique feature that distinguishes the "blue planet" from other planets in the solar system. earth's hydrosphere consists chiefly of the


== atmosphere ==

the atmospheric pressure at earth's sea level averages 101.325 kpa (14.696 psi), with a scale height of about 8.5 km (5.3 mi). a dry atmosphere is composed of 78.084% nitrogen, 20.94


=== weather and climate ===

earth's atmosphere has no definite boundary, slowly becoming thinner and fading into outer space. three-quarters of the atmosphere's mass is contained within the first 11 km (6.8 mi)

the primary atmospheric circulation bands consist of the trade winds in the equatorial region below 30° latitude and the westerlies in the mid-latitudes between 30° and 60°. ocean cu


=== upper atmosphere ===

above the troposphere, the atmosphere is usually divided into the stratosphere, mesosphere, and thermosphere. each layer has a different lapse rate, defining the rate of change in te


== life on earth ==

a planet's life forms inhabit ecosystems, whose total forms the biosphere. the biosphere is divided into a number of biomes, inhabited by broadly similar plants and animals. on land,


== human geography ==
```



```
interactive 3d visualisation of the sun, earth and moon system
gplates portal (university of sydney)
```

```
[2] import requests
    # this is the url for Alice in Wonderland
    result = requests.get('http://www.gutenberg.org/files/11/11-0.txt')
    print(result.text)
```

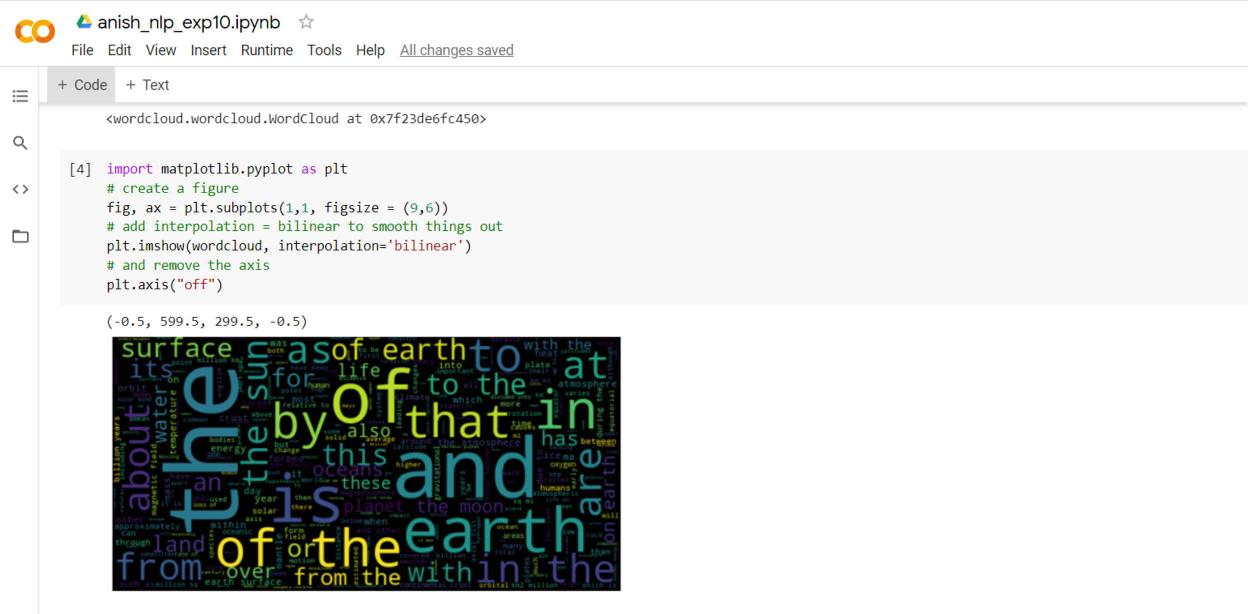
```
and official page at www.gutenberg.org/contact

Section 4. Information about Donations to the Project Gutenberg
Literary Archive Foundation

Project Gutenberg-tm depends upon and cannot survive without
widespread public support and donations to carry out its mission of
increasing the number of public domain and licensed works that can be
freely distributed in machine-readable form accessible by the widest
array of equipment including outdated equipment. Many small donations
($1 to $5,000) are particularly important to maintaining tax exempt
status with the IRS.

The Foundation is committed to complying with the laws regulating
charities and charitable donations in all 50 states of the United
States. Compliance requirements are not uniform and it takes a
considerable effort, much paperwork and many fees to meet and keep up
with these requirements. We do not solicit donations in locations
where we have not received written confirmation of compliance. To SEND
DONATIONS or determine the status of compliance for any particular
state visit www.gutenberg.org/donate

While we cannot and do not solicit contributions from states where we
have not met the solicitation requirements, we know of no prohibition
against accepting unsolicited donations from donors in such states who
approach us with offers to donate.
```



```
[2] subscribe to our email newsletter to hear about new eBooks.
```

```python
# import the wordcloud library
from wordcloud import WordCloud
# Instantiate a new wordcloud.
wordcloud = WordCloud(random_state = 8,
        normalize_plurals = False,
        width = 600, height= 300,
        max_words = 300,
        stopwords = [])
# Apply the wordcloud to the text.
wordcloud.generate(text)
```

```
<wordcloud.wordcloud.WordCloud at 0x7f23de6fc450>
```

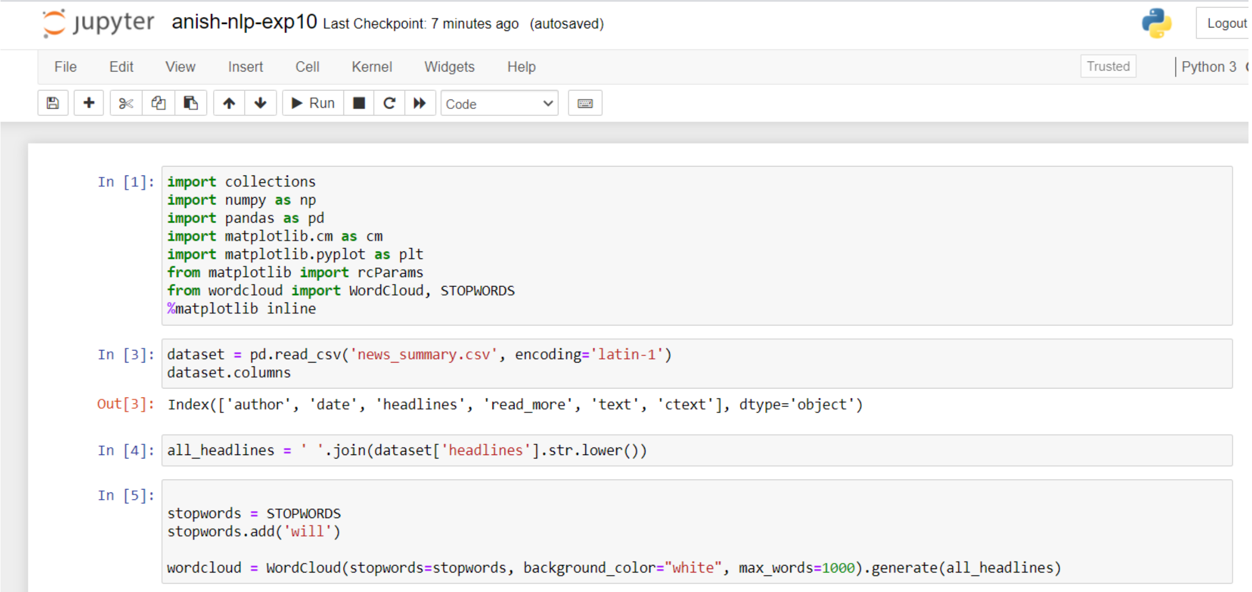


```
<wordcloud.wordcloud.WordCloud at 0x7f23de6fc450>
```

```python
[4] import matplotlib.pyplot as plt
    # create a figure
    fig, ax = plt.subplots(1,1, figsize = (9,6))
    # add interpolation = bilinear to smooth things out
    plt.imshow(wordcloud, interpolation='bilinear')
    # and remove the axis
    plt.axis("off")
```

```
(-0.5, 599.5, 299.5, -0.5)
```

Identifying top topics using word cloud



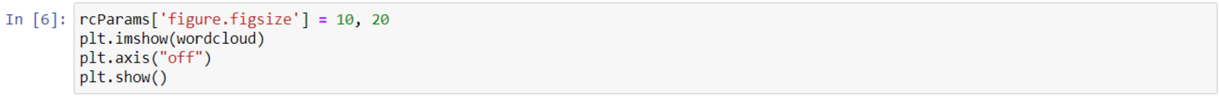
```
In [1]: import collections
        import numpy as np
        import pandas as pd
        import matplotlib.cm as cm
        import matplotlib.pyplot as plt
        from matplotlib import rcParams
        from wordcloud import WordCloud, STOPWORDS
        %matplotlib inline
```

```
In [3]: dataset = pd.read_csv('news_summary.csv', encoding='latin-1')
        dataset.columns
```

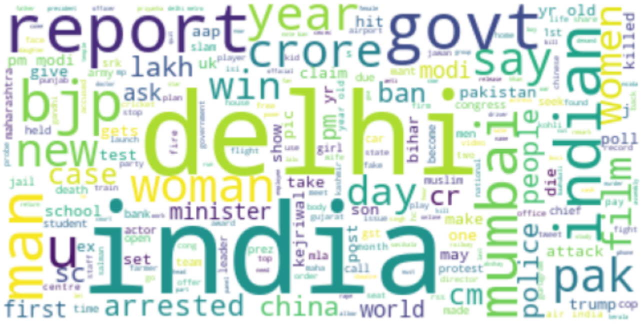
```
Out[3]: Index(['author', 'date', 'headlines', 'read_more', 'text', 'ctext'], dtype='object')
```

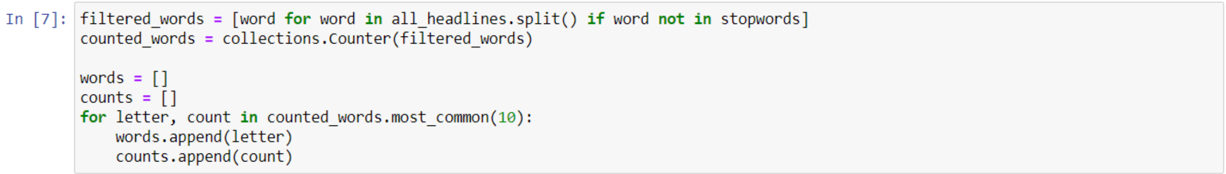
```
In [4]: all_headlines = ' '.join(dataset['headlines'].str.lower())
```

```
In [5]:
        stopwords = STOPWORDS
        stopwords.add('will')

        wordcloud = WordCloud(stopwords=stopwords, background_color="white", max_words=1000).generate(all_headlines)
```

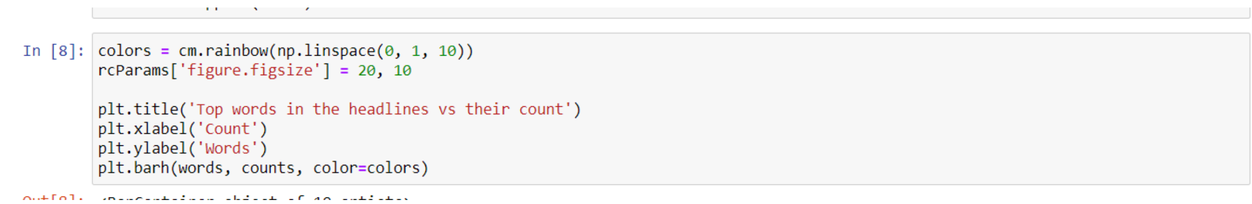


```
In [6]: rcParams['figure.figsize'] = 10, 20
        plt.imshow(wordcloud)
        plt.axis("off")
        plt.show()
```



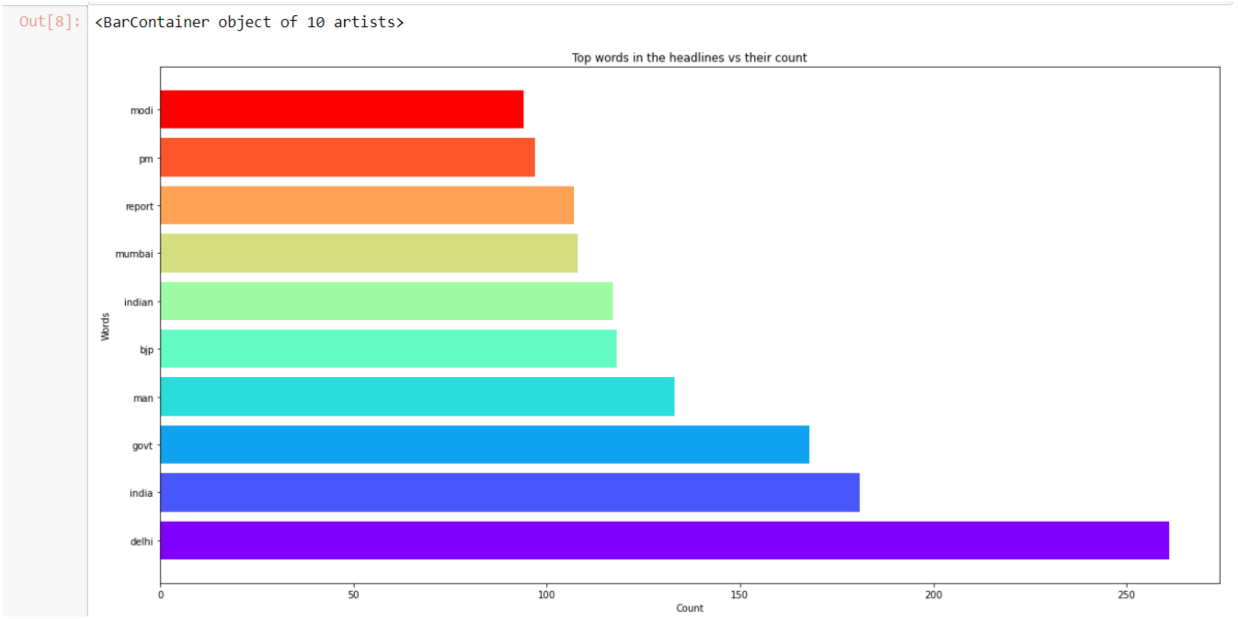


```
In [7]: filtered_words = [word for word in all_headlines.split() if word not in stopwords]
        counted_words = collections.Counter(filtered_words)

        words = []
        counts = []
        for letter, count in counted_words.most_common(10):
            words.append(letter)
            counts.append(count)
```



```
In [8]: colors = cm.rainbow(np.linspace(0, 1, 10))
        rcParams['figure.figsize'] = 20, 10

        plt.title('Top words in the headlines vs their count')
        plt.xlabel('Count')
        plt.ylabel('Words')
        plt.barh(words, counts, color=colors)
```

Out[8]: <BarContainer object of 10 artists>



```
Out[8]: <BarContainer object of 10 artists>
```