ISyE 6740 – Spring 2022
Project Report

Team Member Names: Prateek Shukla (one member)

Project Title: Hypothesis testing – general association between topics and sentiment in the CNN-Daily mail articles

Problem Statement

There is often speculation among fellow citizens about news media being associated with certain sentiments toward certain topics or sets of topics. These topics could be certain groups of politicians, countries, or national issues such as globalization, climate change, immigration, etc.

This project aims to do the following:

1: Label each article with two attributes –

- A 'Topic' or topics based on what it talks about the most by comparing it with a predefined list of topics. This project did not depend on the predefined topic of the articles by CNN to be unbiased in this project.
- A 'Sentiment Label' i.e. Positive, Negative or Neutral based on rule-based sentiment analysis of each article

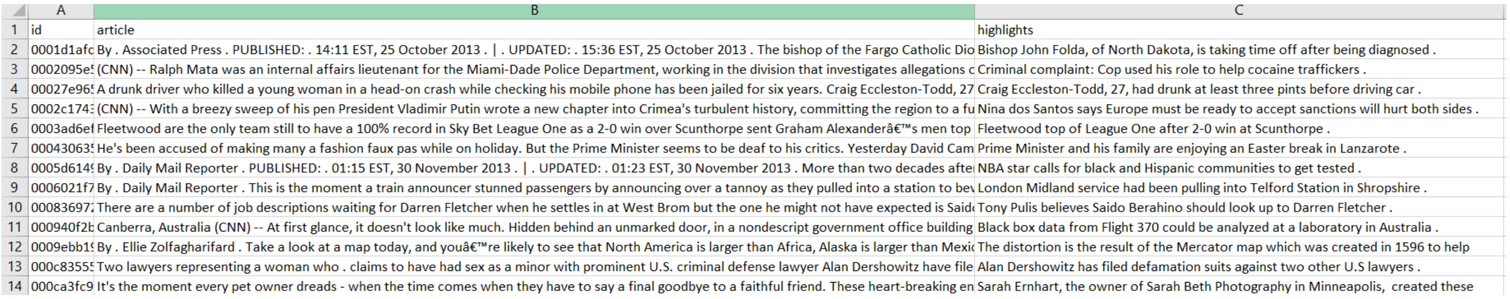
2: Group/Clusters the articles by performing unsupervised spectral clustering based on 'edges' or 'connections' between two articles sharing common topic/s

3: Test the hypothesis of an association between clusters of topics and sentiments in the CNN-daily mail articles using the clusters created above and the sentiment label of the clusters. The mismatch rate between the sentiment label of the entire cluster/topic and individual articles will be the key metric to determine if there is indeed an association thus testing the hypothesis.

Data Source

The core data sources for this project will be the following –
1: CNN-DailyMail News Text Summarization dataset on Kaggle: Source - https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail

This data source consists of over 250k news articles written by journalists at CNN and Dailymail. The data seems to be updated through 2022. This was generated using a web crawler as mentioned on the Kaggle website. Here is a snippet showing a glimpse of the data.



	A	B	C
1	id	article	highlights
2	0001d1afc	By . Associated Press . PUBLISHED: . 14:11 EST, 25 October 2013 . | . UPDATED: . 15:36 EST, 25 October 2013 . The bishop of the Fargo Catholic Dio	Bishop John Folda, of North Dakota, is taking time off after being diagnosed .
3	0002095e5	(CNN) -- Ralph Mata was an internal affairs lieutenant for the Miami-Dade Police Department, working in the division that investigates allegations c	Criminal complaint: Cop used his role to help cocaine traffickers .
4	00027e965	A drunk driver who killed a young woman in a head-on crash while checking his mobile phone has been jailed for six years. Craig Eccleston-Todd, 27	Craig Eccleston-Todd, 27, had drunk at least three pints before driving car .
5	0002c1743	(CNN) -- With a breezy sweep of his pen President Vladimir Putin wrote a new chapter into Crimea's turbulent history, committing the region to a fu	Nina dos Santos says Europe must be ready to accept sanctions will hurt both sides .
6	0003ad6ef	Fleetwood are the only team still to have a 100% record in Sky Bet League One as a 2-0 win over Scunthorpe sent Graham Alexanderâ€™s men top	Fleetwood top of League One after 2-0 win at Scunthorpe .
7	000430635	He's been accused of making many a fashion faux pas while on holiday. But the Prime Minister seems to be deaf to his critics. Yesterday David Cam	Prime Minister and his family are enjoying an Easter break in Lanzarote .
8	0005d6149	By . Daily Mail Reporter . PUBLISHED: . 01:15 EST, 30 November 2013 . | . UPDATED: . 01:23 EST, 30 November 2013 . More than two decades afte	NBA star calls for black and Hispanic communities to get tested .
9	0006021f7	By . Daily Mail Reporter . This is the moment a train announcer stunned passengers by announcing over a tannoy as they pulled into a station to bev	London Midland service had been pulling into Telford Station in Shropshire .
10	000836972	There are a number of job descriptions waiting for Darren Fletcher when he settles in at West Brom but the one he might not have expected is Said	Tony Pulis believes Saido Berahino should look up to Darren Fletcher .
11	000940f2b	Canberra, Australia (CNN) -- At first glance, it doesn't look like much. Hidden behind an unmarked door, in a nondescript government office building	Black box data from Flight 370 could be analyzed at a laboratory in Australia .
12	0009ebb19	By . Ellie Zolfagharifard . Take a look at a map today, and youâ€™re likely to see that North America is larger than Africa, Alaska is larger than Mexic	The distortion is the result of the Mercator map which was created in 1596 to help
13	000c83555	Two lawyers representing a woman who . claims to have had sex as a minor with prominent U.S. criminal defense lawyer Alan Dershowitz have file	Alan Dershowitz has filed defamation suits against two other U.S lawyers .
14	000ca3fc9	It's the moment every pet owner dreads - when the time comes when they have to say a final goodbye to a faithful friend. These heart-breaking en	Sarah Ernhart, the owner of Sarah Beth Photography in Minneapolis, created these

Figure 1- Snippet of Original CNN Dailymail Articles Data

Due to computing time limitations ((realized during the analysis), only 5% of the original dataset was used for the analysis thus totaling 14148 articles.

2: Alternative to 1 (if needed): Source – https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning

This project did not need to use Data Source 2 as the data in source 1 was sufficient.

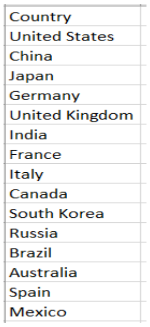
3: Sentiment Source –
All of the following data sources could be used as an English sentiment dictionary/lexicon or dataset for words with a sentiment label to generate a sentiment label for the article.

- https://nlp.stanford.edu/sentiment/code.html
- https://www.cs.jhu.edu/~mdredze/datasets/sentiment/
- http://jmcauley.ucsd.edu/data/amazon/

Although, a more structured approach is by using existing widely used lexicon-based python packages such as Textblob, VADER, SentiWordNet, etc. for sentiment analysis. Textblob package was used in this project.

4: List of topics –
The following lists of topics were initially created and planned to be tested in this project although due to computing time limitations (realized during the analysis), only 1 and 2 were finall. Datasets for these lists of topics

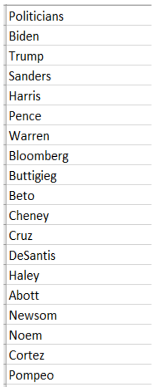
1) 15 biggest countries in the world economy-wise - https://www.investopedia.com/insights/worlds-top-economies/



Country
United States
China
Japan
Germany
United Kingdom
India
France
Italy
Canada
South Korea
Russia
Brazil
Australia
Spain
Mexico

Figure 2- Snippet of Countries Dataset

2) A few famous Politicians in the United States (most of them being future or past presidential candidates)
   https://today.yougov.com/ratings/politics/popularity/politicians/all



Politicians
Biden
Trump
Sanders
Harris
Pence
Warren
Bloomberg
Buttigieg
Beto
Cheney
Cruz
DeSantis
Haley
Abott
Newsom
Noem
Cortez
Pompeo

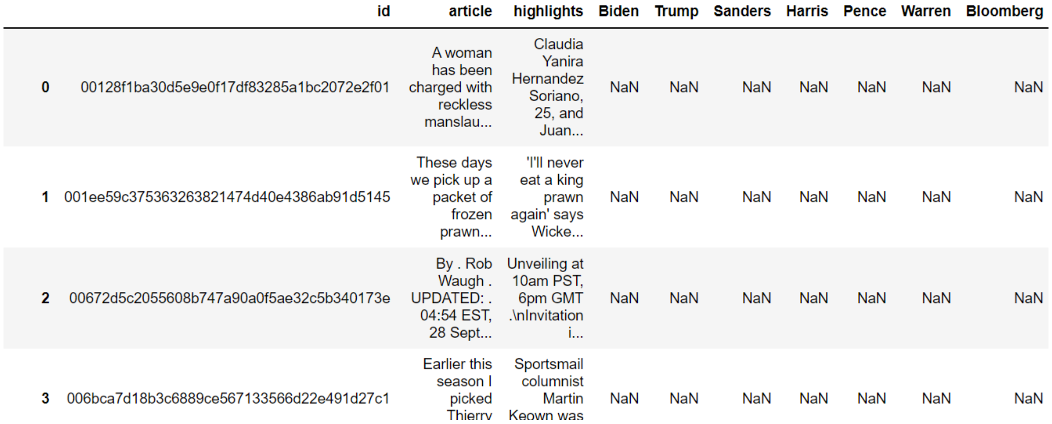
Figure 3- Snippet of Politicians Dataset

3) A few controversial political issues
4) 50 states in the US

Methodology

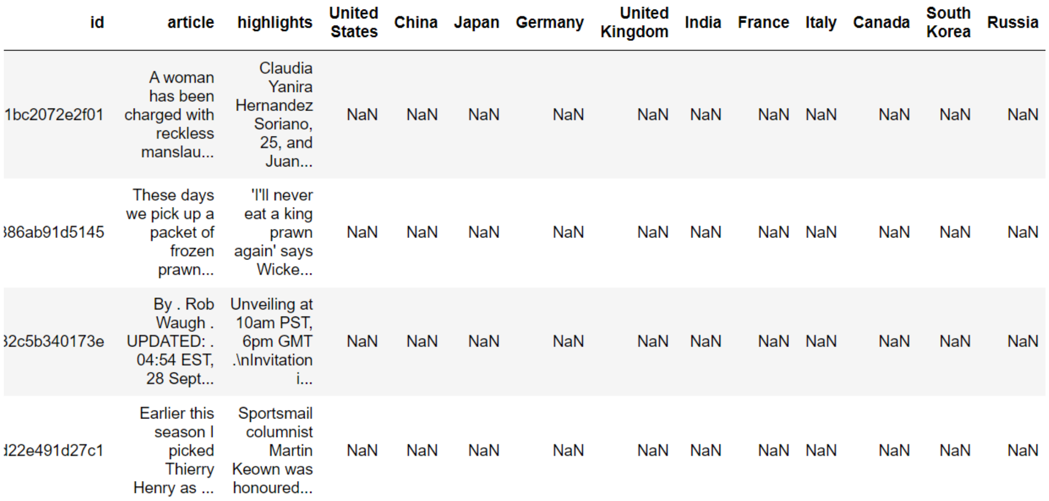
The following major steps will be undertaken to complete the analysis –

- Select a list of topics from the lists of topics mentioned in the 'Data Source' section.
    - As mentioned in the data sources section, two lists of topics were selected. This list of topics were created using Excel manually on random basis from the web sources and read in the form of pandas series ('countries', 'politicians')using pd.read_excel function. Later these series were also converted into 1D arrays ('countries_updated', 'politicians_updated')for the purpose of further analysis. These
- Read the CNN excel into the form of pandas dataframe and also a numpy matrix for further analysis
- Edit articles dataframe 'data_politicians' and 'data_countries' to include indicator columns for individual topics in each list. Set a minimum threshold on the word count for setting the topic of an article.



	id	article	highlights	Biden	Trump	Sanders	Harris	Pence	Warren	Bloomberg
0	00128f1ba30d5e9e0f17df83285a1bc2072e2f01	A woman has been charged with reckless manslau...	Claudia Yanira Hernandez Soriano, 25, and Juan...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	001ee59c375363263821474d40e4386ab91d5145	These days we pick up a packet of frozen prawn...	'I'll never eat a king prawn again' says Wicke...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	00672d5c2055608b747a90a0f5ae32c5b340173e	By . Rob Waugh . UPDATED: . 04:54 EST, 28 Sept...	Unveiling at 10am PST, 6pm GMT .\nInvitation i...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	006bca7d18b3c6889ce567133566d22e491d27c1	Earlier this season I picked Thierry	Sportsmail columnist Martin Keown was	NaN	NaN	NaN	NaN	NaN	NaN	NaN

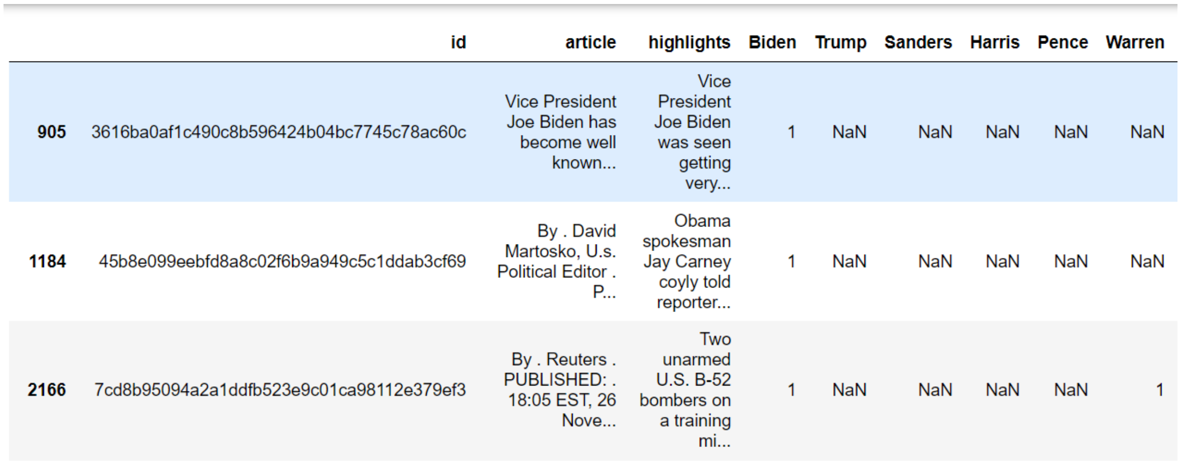
Figure 4 - snippet showing a first few rows with a few indicator columns for data_politicians



id	article	highlights	United States	China	Japan	Germany	United Kingdom	India	France	Italy	Canada	South Korea	Russia
1bc2072e2f01	A woman has been charged with reckless manslau...	Claudia Yanira Hernandez Soriano, 25, and Juan...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
86ab91d5145	These days we pick up a packet of frozen prawn...	'I'll never eat a king prawn again' says Wicke...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2c5b340173e	By . Rob Waugh . UPDATED: . 04:54 EST, 28 Sept...	Unveiling at 10am PST, 6pm GMT .\nInvitation i...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
22e491d27c1	Earlier this season I picked Thierry Henry as ...	Sportsmail columnist Martin Keown was honoured...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

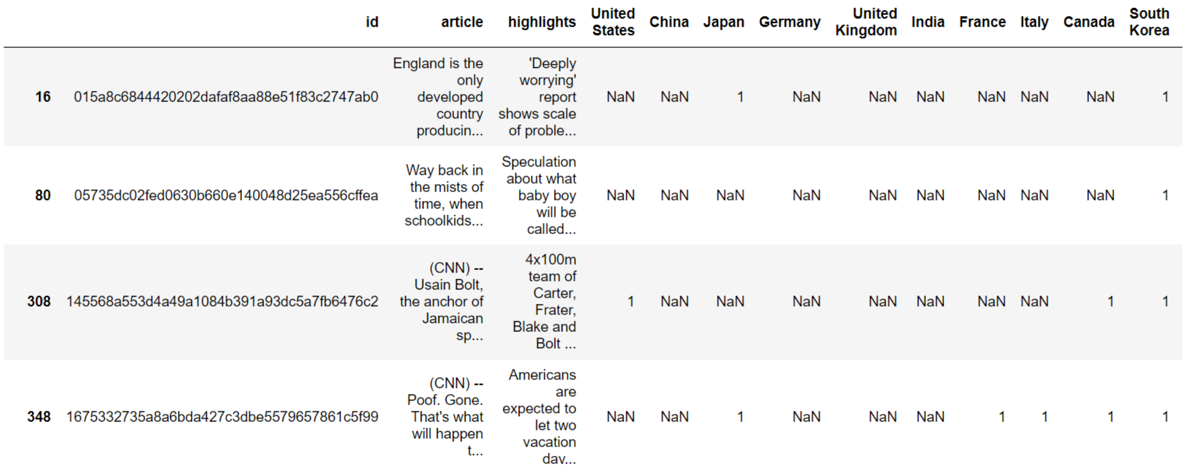
Figure 5 - snippet showing a first few rows with a few indicator columns for data_countries

- Topic Allocation: Review each article i.e. row from the 'data_countries' and 'data_politicians' dataframes for topic/s allocation
    - A loop is ran through each article of 'data countries' dataframe identifying the number of occurrences of each topic (by using another nested loop running through 'countries_updated' array) and the value of corresponding indicator variable for the article/row is allocated '1' if the number of occurrences exceed or equal a preset threshold. Similar procedure is performed for the 'data_politicians' dataframe based on 'politicians' array. For countries, the threshold is set to 1 and for politicians, the threshold is set to 2.
    - Regular expressions package 'Re' and function 'findall' is used to identify matches with a 'IGNORECASE' flag to ignore the case while matching.
    - The original dataframes are edited to reflect the updated values of indicator variables. **An important observation** is that an article could have multiple topics within a list of topics. For example, an article could be related to Joe Biden and Ted Cruz at the same time.



	id	article	highlights	Biden	Trump	Sanders	Harris	Pence	Warren
905	3616ba0af1c490c8b596424b04bc7745c78ac60c	Vice President Joe Biden has become well known...	Vice President Joe Biden was seen getting very...	1	NaN	NaN	NaN	NaN	NaN
1184	45b8e099eebfd8a8c02f6b9a949c5c1ddab3cf69	By . David Martosko, U.s. Political Editor . P...	Obama spokesman Jay Carney coyly told reporter...	1	NaN	NaN	NaN	NaN	NaN
2166	7cd8b95094a2a1ddfb523e9c01ca98112e379ef3	By . Reuters . PUBLISHED: . 18:05 EST, 26 Nove...	Two unarmed U.S. B-52 bombers on a training mi...	1	NaN	NaN	NaN	NaN	1

Figure 6 - Snippet showing example of Articles with Biden as topic and sometime mutiple topics



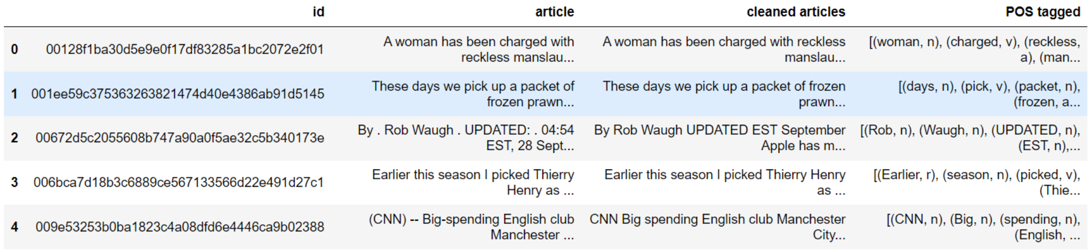
	id	article	highlights	United States	China	Japan	Germany	United Kingdom	India	France	Italy	Canada	South Korea
16	015a8c6844420202dafaf8aa88e51f83c2747ab0	England is the only developed country producin...	'Deeply worrying' report shows scale of proble...	NaN	NaN	1	NaN	NaN	NaN	NaN	NaN	NaN	1
80	05735dc02fed0630b660e140048d25ea556cffea	Way back in the mists of time, when schoolkids...	Speculation about what baby boy will be called...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
308	145568a553d4a49a1084b391a93dc5a7fb6476c2	(CNN) -- Usain Bolt, the anchor of Jamaican sp...	4x100m team of Carter, Frater, Blake and Bolt ...	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	1
348	1675332735a8a6bda427c3dbe5579657861c5f99	(CNN) -- Poof. Gone. That's what will happen t...	Americans are expected to let two vacation dav...	NaN	NaN	1	NaN	NaN	NaN	1	1	1	1

Figure 7- Snippet showing example of Articles with South Korea as topic and sometime mutiple topics

- Sentiment Analysis:
    - A separate 'articles' dataframe is created from any of the dateframes 'data_countries' or 'data_politicians' containing just 'articles' and 'id' columns. This is to be used for further cleaning and manipulation of article text for sentiment analysis. **An important point** to note is that the project uses a rule based sentiment analysis approach extracting sentiments from the significant words based on a lexicon since the there is no training articles data with sentiment label available to be able to use a machine learning based approach.

    - Data preprocessing for sentiment analysis – (Reference - https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/?)

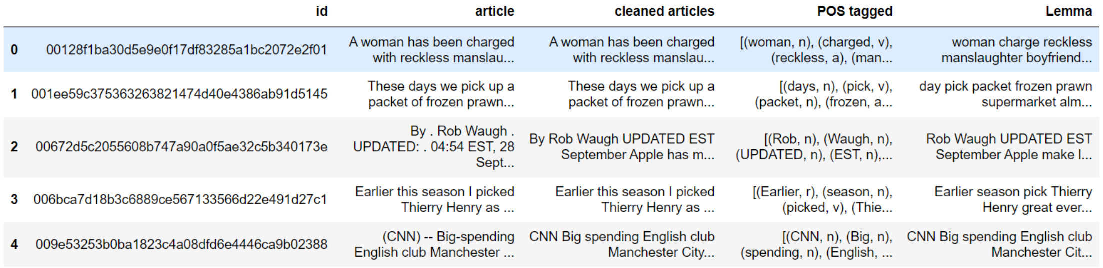
        - Text cleaning 'article' column: function 'sub' within 're' package is used to remove all special characters and numericals leaving the alphabets
        - Tokenization of text – breaking the text into tokens i.e. smaller pieces at word level using word_tokenize() function within nltk package
        - POS Tagging – Parts of speech (POS) tagging is performed to convert each token i.e. word into a tuple of the form (word,tag). The tag represents the context of the word and will be used for Lemmetization step below. NLTK pos_tag function is used for this.



	id	article	cleaned articles	POS tagged
0	00128f1ba30d5e9e0f17df83285a1bc2072e2f01	A woman has been charged with reckless manslau...	A woman has been charged with reckless manslau...	[(woman, n), (charged, v), (reckless, a), (man...
1	001ee59c375363263821474d40e4386ab91d5145	These days we pick up a packet of frozen prawn...	These days we pick up a packet of frozen prawn...	[(days, n), (pick, v), (packet, n), (frozen, a...
2	00672d5c2055608b747a90a0f5ae32c5b340173e	By . Rob Waugh . UPDATED: . 04:54 EST, 28 Sept...	By Rob Waugh UPDATED EST September Apple has m...	[(Rob, n), (Waugh, n), (UPDATED, n), (EST, n),...
3	006bca7d18b3c6889ce567133566d22e491d27c1	Earlier this season I picked Thierry Henry as ...	Earlier this season I picked Thierry Henry as ...	[(Earlier, r), (season, n), (picked, v), (Thie...
4	009e53253b0ba1823c4a08dfd6e4446ca9b02388	(CNN) -- Big-spending English club Manchester ...	CNN Big spending English club Manchester City...	[(CNN, n), (Big, n), (spending, n), (English, ...

Figure 8 - A snippet showing cleaned articles and POS tagged columns

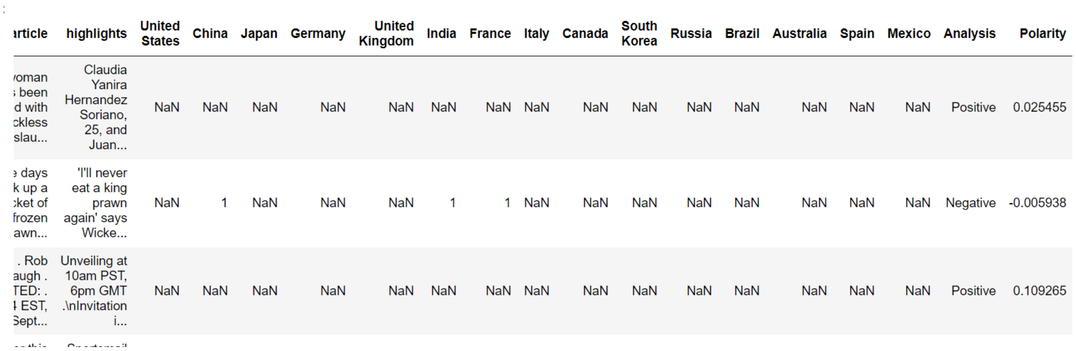
        - Stopwords Removal – Stopwords are words like ‘I’ , ‘me’, ‘was’, ‘his’ that carry very little useful information. This is performed using an if condition to only select words not in stopwords.words(‘english’) list
        - Lemmetization – The process of lemmetization involves extracting the root words (called Lemmas) from words which will carry lexical meaning or a sentiment. For example, ‘glance’ from ‘glanced’. This is performed on the tuples obtained in the POS tagging step using lemmatize function.



	id	article	cleaned articles	POS tagged	Lemma
0	00128f1ba30d5e9e0f17df83285a1bc2072e2f01	A woman has been charged with reckless manslau...	A woman has been charged with reckless manslau...	[(woman, n), (charged, v), (reckless, a), (man...	woman charge reckless manslaughter boyfriend...
1	001ee59c375363263821474d40e4386ab91d5145	These days we pick up a packet of frozen prawn...	These days we pick up a packet of frozen prawn...	[(days, n), (pick, v), (packet, n), (frozen, a...	day pick packet frozen prawn supermarket alm...
2	00672d5c2055608b747a90a0f5ae32c5b340173e	By . Rob Waugh . UPDATED: . 04:54 EST, 28 Sept...	By Rob Waugh UPDATED EST September Apple has m...	[(Rob, n), (Waugh, n), (UPDATED, n), (EST, n),...	Rob Waugh UPDATED EST September Apple make l...
3	006bca7d18b3c6889ce567133566d22e491d27c1	Earlier this season I picked Thierry Henry as ...	Earlier this season I picked Thierry Henry as ...	[(Earlier, r), (season, n), (picked, v), (Thie...	Earlier season pick Thierry Henry great ever...
4	009e53253b0ba1823c4a08dfd6e4446ca9b02388	(CNN) -- Big-spending English club Manchester ...	CNN Big spending English club Manchester City...	[(CNN, n), (Big, n), (spending, n), (English, ...	CNN Big spending English club Manchester Cit...

Figure 9 - Snippet showing 'articles' dataframe post lemmetization

    - Next, TextBlob function within ‘textblob’ package is used to calculate the ‘Polarity’ score of each article. If the polarity score is less than 0, the analysis column (which represents sentiment of the articles) is set to ‘Negative’, ‘Positive’ in case of score greater than 0 and ‘Neutral’ in case of score equal to 0. Finally, the ‘polarity’ and ‘analysis’ columns are appended into ‘data_countries’ and ‘data_politicians’ dataframe with same indices



ırticle	highlights	United States	China	Japan	Germany	United Kingdom	India	France	Italy	Canada	South Korea	Russia	Brazil	Australia	Spain	Mexico	Analysis	Polarity
/oman ; been d with ckless slau...	Claudia Yanira Hernandez Soriano, 25, and Juan...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Positive	0.025455
; days k up a :ket of frozen awn...	'I'll never eat a king prawn again' says Wicke...	NaN	1	NaN	NaN	NaN	1	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Negative	-0.005938
. Rob augh . TED: . 1 EST, Sept...	Unveiling at 10am PST, 6pm GMT .\nInvitation i...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Positive	0.109265

Figure 10 - Snippet showing updated data_countries dataframe with polarity and analysis columns

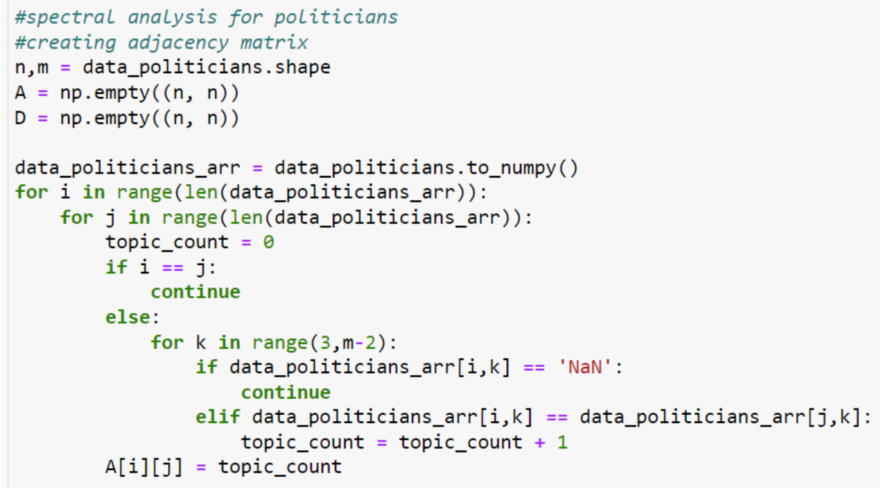
- Spectral Clustering (Reference - https://towardsdatascience.com/spectral-clustering-aba2640c0d5b)

    - The project aims to cluster the article based on common topic/s with weightage given to number of common topics between two articles.
    - **Rationale behind selecting spectral clustering as the modeling techinique** –
        - Project aims to test hypothesis of an association between a **general clusters/sets** of articles with common topics (from a predefined list of topics) AND sentiment for CNN articles. The analysis do not aim to

analyze sentiments around **specific** topics by simple sentiment analysis for each topic as that could be more biased approach.

The general clusters of articles with common topics could be created using an unsupervised learning technique. Since, the articles are connected to each other based on common topic/s with each common topic counted as weight of 1, the project uses these connections/edges to use the articles data as a graph thus needing spectral clustering.

- For spectral clustering analysis, each article is considered to be a node. If the topic of two nodes is the same, we consider there to be an edge/relationship between the two nodes/articles. The adjacency matrix A is created using the 'data_politicians_arr' (a matrix created from the data_politicians dataframe) by running three loops through the data as shown below. Same procedure is performed on the data_countries dataframe. **An important note** – each common topic count as connection between two articles so two articles with 3 common topic will have 3 as values in the adjacency matrix between those two nodes.



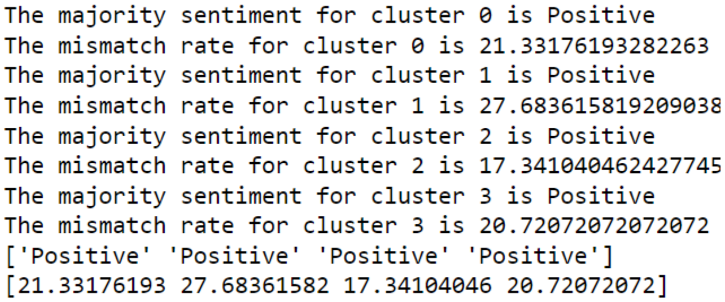
```python
#spectral analysis for politicians
#creating adjacency matrix
n,m = data_politicians.shape
A = np.empty((n, n))
D = np.empty((n, n))

data_politicians_arr = data_politicians.to_numpy()
for i in range(len(data_politicians_arr)):
    for j in range(len(data_politicians_arr)):
        topic_count = 0
        if i == j:
            continue
        else:
            for k in range(3,m-2):
                if data_politicians_arr[i,k] == 'NaN':
                    continue
                elif data_politicians_arr[i,k] == data_politicians_arr[j,k]:
                    topic_count = topic_count + 1
        A[i][j] = topic_count
```

Figure 11- Snippet showing code to create adjacency matrix for data_countries

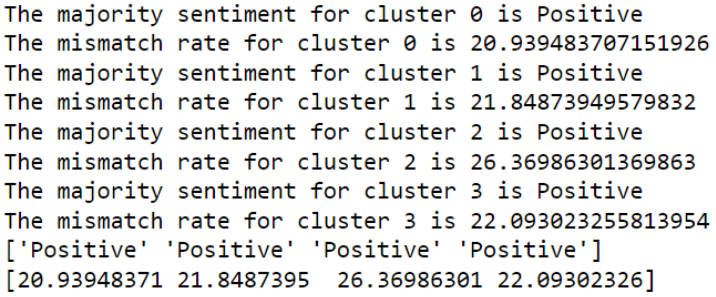
- Degree matrix D is created for 'data_politicians' dataframe using the 'data_politicians_arr' matrix i.e. a diagonal matix with diagonal values representing the number of edges/connections for each node/article. Same procedure is performed on the data_countries dataframe
- Laplacian matrix L is created ➔ L = D – A. Same procedure is performed on the data_countries dataframe
- First, the project performs spectral clustering assuming a smaller number of clusters k = 4. The Kmeans function within sklearn.cluster package to perform k means clustering on the Laplacian matrix L. The k smallest eigenvalues and corresponding eigenvectors are obtained using the eigh function within linalg library of numpy package. Same procedure is performed on the data_countries dataframe
- The 'labels' array is outputted with cluster labels 0,1,2,and 3 for all the articles/nodes. Same procedure is performed on the data_countries dataframe

- The number of articles within each cluster (in increasing order from 0 to 3) for data_politicians and data_countries respectively is given as follows for reference –
  - 13576, 177, 173, 222
  - 12990, 528, 139, 491
- **Important observation** - The above numbers shows that there is indeed one much bigger cluster for both the datasets representing a strong connection between a big set of articles for both of the datasets.
- The majority sentiment label is calculated for each cluster by observing the sentimemt labels of majority number of articles i.e. more than 50%. The sentiment labels are calculated as 'Postive' or 'Negative' or 'Neutral'. Same procedure is performed on the data_countries dataframe
- The mismatch rate for each cluster is calculated by using the formula –
  - Mismatch rate = (number of mistmatches between the sentiment label of each article and the majority label of the cluster/number of the total articles in a cluster) * 100. Same procedure is performed on the data_countries dataframe
- The output of the majority label and mismatch rate analysis is shown below for k = 4



```
The majority sentiment for cluster 0 is Positive
The mismatch rate for cluster 0 is 21.33176193282263
The majority sentiment for cluster 1 is Positive
The mismatch rate for cluster 1 is 27.683615819209038
The majority sentiment for cluster 2 is Positive
The mismatch rate for cluster 2 is 17.341040462427745
The majority sentiment for cluster 3 is Positive
The mismatch rate for cluster 3 is 20.72072072072072
['Positive' 'Positive' 'Positive' 'Positive']
[21.33176193 27.68361582 17.34104046 20.72072072]
```

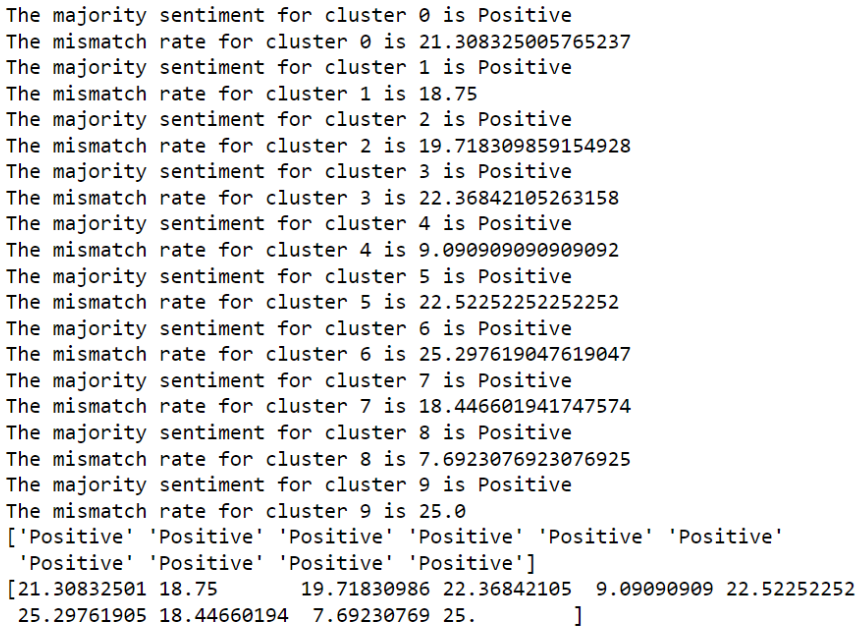
Figure 12- Majority Sentiment Label and Mismatch Rate Output for Politicians dataset for k = 4



```
The majority sentiment for cluster 0 is Positive
The mismatch rate for cluster 0 is 20.939483707151926
The majority sentiment for cluster 1 is Positive
The mismatch rate for cluster 1 is 21.84873949579832
The majority sentiment for cluster 2 is Positive
The mismatch rate for cluster 2 is 26.36986301369863
The majority sentiment for cluster 3 is Positive
The mismatch rate for cluster 3 is 22.093023255813954
['Positive' 'Positive' 'Positive' 'Positive']
[20.93948371 21.8487395  26.36986301 22.09302326]
```

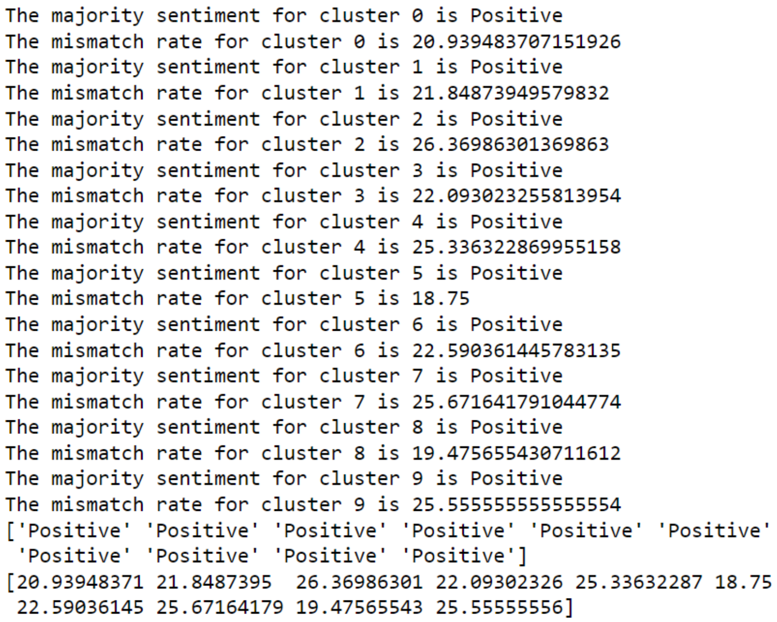
Figure 13- Majority Sentiment Label and Mismatch Rate Output for Countries dataset for k = 4

- The same analysis is performed for both data_countries and data_politicians dataframes assuming k = 10. This helps to find the effect of diluting the clusters or catching further variance on mismatch rate. The number of articles in each clusters is as follows for politicians and countries respectively –
  - 13009, 64, 71, 76, 11, 111, 336, 309, 13, 148
  - 11815, 357, 292, 86, 446, 128, 332, 335, 267, 90
  - The above number shows that there is still one much bigger cluster and the smaller cluster starts diluting based on the value of k
- The output for k=10 is shown below



```
The majority sentiment for cluster 0 is Positive
The mismatch rate for cluster 0 is 21.308325005765237
The majority sentiment for cluster 1 is Positive
The mismatch rate for cluster 1 is 18.75
The majority sentiment for cluster 2 is Positive
The mismatch rate for cluster 2 is 19.718309859154928
The majority sentiment for cluster 3 is Positive
The mismatch rate for cluster 3 is 22.36842105263158
The majority sentiment for cluster 4 is Positive
The mismatch rate for cluster 4 is 9.090909090909092
The majority sentiment for cluster 5 is Positive
The mismatch rate for cluster 5 is 22.52252252252252
The majority sentiment for cluster 6 is Positive
The mismatch rate for cluster 6 is 25.297619047619047
The majority sentiment for cluster 7 is Positive
The mismatch rate for cluster 7 is 18.446601941747574
The majority sentiment for cluster 8 is Positive
The mismatch rate for cluster 8 is 7.6923076923076925
The majority sentiment for cluster 9 is Positive
The mismatch rate for cluster 9 is 25.0
['Positive' 'Positive' 'Positive' 'Positive' 'Positive' 'Positive'
 'Positive' 'Positive' 'Positive' 'Positive']
[21.30832501 18.75       19.71830986 22.36842105  9.09090909 22.52252252
 25.29761905 18.44660194  7.69230769 25.        ]
```

Figure 14 - Majority Sentiment Label and Mismatch Rate Output for Politicians dataset for k = 10



```
The majority sentiment for cluster 0 is Positive
The mismatch rate for cluster 0 is 20.939483707151926
The majority sentiment for cluster 1 is Positive
The mismatch rate for cluster 1 is 21.84873949579832
The majority sentiment for cluster 2 is Positive
The mismatch rate for cluster 2 is 26.36986301369863
The majority sentiment for cluster 3 is Positive
The mismatch rate for cluster 3 is 22.093023255813954
The majority sentiment for cluster 4 is Positive
The mismatch rate for cluster 4 is 25.336322869955158
The majority sentiment for cluster 5 is Positive
The mismatch rate for cluster 5 is 18.75
The majority sentiment for cluster 6 is Positive
The mismatch rate for cluster 6 is 22.590361445783135
The majority sentiment for cluster 7 is Positive
The mismatch rate for cluster 7 is 25.671641791044774
The majority sentiment for cluster 8 is Positive
The mismatch rate for cluster 8 is 19.475655430711612
The majority sentiment for cluster 9 is Positive
The mismatch rate for cluster 9 is 25.555555555555554
['Positive' 'Positive' 'Positive' 'Positive' 'Positive' 'Positive'
 'Positive' 'Positive' 'Positive' 'Positive']
[20.93948371 21.8487395  26.36986301 22.09302326 25.33632287 18.75
 22.59036145 25.67164179 19.47565543 25.55555556]
```

Figure 15 - Majority Sentiment Label and Mismatch Rate Output for Countries dataset for k = 10

Evaluation and Final Results

- It is seen from the above analysis that the mismatch rate is on the lower side (<25%) and in some cluster even lower than that. Also, all of the clusters have majority of the articles with positive sentiment. Thus, there seems to be some level of association between a topics and sentiment but it is hard to draw that inference as we do not have any clusters with negative majority sentiment. It could have been easier to test association between topics and sentiment if we could see lower mismatch rate for a cluster or cluster/s with negative sentiment label. This demands for further investigation based on the entire dataset referring back to using 5% of randomly selected data from original articles dataset.

- Increasing the number of clusters to K = 10 do not change the nature of majority labels for the clusters for both countries and politicians OR mismatch rate much, thus it further strengthens the conclusion above.

- There is future scope for using the entire article dataset (with over 250k articles) and a further extended lists of topics as mentioned above to be able to derive further number of clusters.