Data Mining from the News and building a Network Analysis

May 15, 2020

1 Project Proposal

Team Members:

`Name: Amar Kumar Reddy ,akm352@drexel.edu`

`Name: Astha jain , aj887@drexel.edu`

`Name: Nupur Roy Chowdhury, nr572@drexel.edu`

2 Abstract:

We live in an age of media and information, where the importance of understanding the intricacies of the News cannot be overstated. Building graph representations (i.e. social networks) of data opens up a whole host of possibilities for data science applications, such as finding the most influential individuals on the news or identifying clusters (cliques of people) based on connections. How The sprawling web of politicians, companies look like, using Graph theory and analysing the network representing this vast web of information through news channels.

The fundamental premise behind building our network will be two-fold and quite simple:

1. If two people are mentioned in the same article, they have co-mention relation.
2. The more articles mention the same two people, the closer they have co-mention relation.

Named Entity Recognition:

We are using Named Entity Recognition is a Natural Language Processing task for extracting information from text which recognises entities. This is achieved using statistical models trained on our large dataset, where we make the model learn to recognise and categorise entities based on the context in which they appear in words. We will then be using one of these models for the words tagged as persons, organisations for each article.

Example: If an article mentions Person D and Person B , and two other, separate articles mention Person D and Person A, we'll say that Person D is friends with Person B, and Person D is also friends with Person A, only twice as much.

We not only get a pictorial representation of the friendship group, we can also start seeing hidden relationships:

although Person A wasn't mentioned in the same article as Person B, we can guess with some certainty that the two are related (and that they are related via their mutual friend, Person D). We can also tell that Person D is the alpha male in the group, having influence over both person A and Person B.

Business Problem from the project:

- We are analysing the below following for our project:
- We aim to build a network analysis where we could answer the questions like
- Who are the most influential individuals on the news?
- Does Everyone Know Each Other in the Network?
- But How Do All These People Know Each Other?
- How Many People and Connections are in our Network?
- What does the sprawling web of politicians, companies and celebrities really look like?
- How is one person in news related to another person in another news article?

3 Data collection:

The source I used was from the GDELT Project, which is a free, open platform of all world events, monitored in real time from across the globe. The GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages.

We have fetched their raw data, which they also make available completely free of charge. They publish daily CSVs with thousands of events which occurred during that day, but more importantly, they include the URL of the news source which reported on the event.

I used the file containing the events that ocurred on the 23rd of April, and extracted the top 100 articles around the globe. I limited my search in such a way to somewhat narrow down the volume of entities.

The type of raw data we got contained the source urls. I used Goose to extract the content from each web page. The aim of the GOOSE is to take any news article or article-type web page and not only extract what is the main body of the article but also all meta data and most probable image candidate.

Goose will try to extract the following information:

Main text of an article Main image of article Any YouTube/Vimeo movies embedded in article Meta Description Meta tags

Importing Libraries

[2]:

```
import os
import numpy
import pandas as pd
import seaborn as sns
import re
from goose3 import Goose
# load Flair and NLTK
```

```
import torch
from flair.data import Sentence
from flair.models import SequenceTagger
from nltk import tokenize
import networkx as nx
from itertools import combinations
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\nupur_nsxs2zt\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

[2]:
```
True
```

[11]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

Reading the data from the file

[4]:
```
FILE_ONE = '20200423.gkg.csv'
df_file_one = pd.read_csv(FILE_ONE, sep = '\t')
df_file_one.head()
```

[4]:
```
       DATE  NUMARTS                                             COUNTS  \
0  20200423        2                                                NaN   
1  20200423        1                                                NaN   
2  20200423        1                                                NaN   
3  20200423        1                                                NaN   
4  20200423        1  AFFECT#19##1#Spain#SP#SP#40#-4#SP;CRISISLEX_T0...   

                                              THEMES  \
0                                                NaN   
1  TAX_ETHNICITY;TAX_ETHNICITY_ITALIAN;TAX_WORLDL...   
2  SOC_POINTSOFINTEREST;SOC_POINTSOFINTEREST_AIRP...   
3  LEADER;TAX_FNCACT;TAX_FNCACT_GOVERNOR;CRISISLE...   
4  TAX_FNCACT;TAX_FNCACT_SENATOR;USPEC_POLITICS_G...   

                                           LOCATIONS  \
0  2#Alabama, United States#US#USAL#32.799#-86.80...   
1   2#New York, United States#US#USNY#42.1497#-74...   
2  4#Kuala Lumpur, Kuala Lumpur, Malaysia#MY#MY14...   
3  4#Salisbury, Mashonaland East, Zimbabwe#ZI#ZI0...   
4  4#Manila, Manila, Philippines#RP#RPD9#14.6042#...   

                                             PERSONS  \
0  javon kinlaw;clemson a j terrell;desmond trufa...   
1                      william fisher;kristin fisher   
```

```
2           wisma putra;mohd shukrie;queen elizabeth
3                                      scott gensler
4                     panfilo lacson;rodrigo duterte

                                ORGANIZATIONS  \
0                                  young;espn
1  instagram;texas medical center;league city
2                    kl international airport
3                                         NaN
4                                         NaN

                                                TONE  \
0  -1.7293997965412,1.93285859613428,3.6622583926...
1  -2.05128205128205,2.30769230769231,4.358974358...
2  1,4.33333333333333,3.33333333333333,7.66666666...
3  1.01694915254237,2.03389830508475,1.0169491525...
4  -1.14155251141553,2.28310502283105,3.424657534...

                                       CAMEOEVENTIDS             SOURCES  \
0                                                NaN   espn.com;espn.com
1  919758687,919758695,919758696,919759218,919759...    houstoniamag.com
2                                          919783853  theedgemarkets.com
3                                                NaN            wifr.com
4                                                NaN        inquirer.net

                                          SOURCEURLS
0  https://www.espn.com/blog/atlanta-falcons/post...
1  https://www.houstoniamag.com/coronavirus/2020/...
2  https://www.theedgemarkets.com/article/mahb-do...
3  https://www.wifr.com/content/news/Illinois-gar...
4  https://newsinfo.inquirer.net/1263337/lacson-o...
```

Creating empty dataframe for generating the dataset

[5]:

```
df_cleaned = pd.
 ↪DataFrame(columns=['url','title','meta_description','domain','date','content'])
df_cleaned
```

[5]:

```
Empty DataFrame
Columns: [url, title, meta_description, domain, date, content]
Index: []
```

[6]:

```
    count = 0
    with Goose() as g:
        for url in df_file_one['SOURCEURLS']:
            count += 1
            if count < 100:
                try:
                    article = g.extract(url=url)
                    df_cleaned=df_cleaned.append(
                        {
                            'url' : url ,
                            'title' : article.title,
                            'meta_description':article.meta_description,
                            'domain':article.domain,
                            'date':'04-23-2020',
                            'content':article.cleaned_text
                        }, ignore_index=True)
                except:
                    continue
            else:
                break
```

```
C:\Users\nupur_nsxs2zt\anaconda3\lib\site-
packages\dateutil\parser\_parser.py:1218: UnknownTimezoneWarning: tzname PDT
identified but not understood.  Pass `tzinfos` argument in order to correctly
return a timezone-aware datetime.  In a future version, this will raise an
exception.
  category=UnknownTimezoneWarning)
```

Extracting the Title, Meta description, domain and the content of each article allowed me to organise my data into a nice pandas DataFrame:

[7]:
```
df_cleaned.head()
```

[7]:
```
                                                 url  \
0  https://www.espn.com/blog/atlanta-falcons/post...
1  https://www.houstoniamag.com/coronavirus/2020/...
2  https://www.theedgemarkets.com/article/mahb-do...
3  https://www.wifr.com/content/news/Illinois-gar...
4  https://newsinfo.inquirer.net/1263337/lacson-o...

                                               title  \
0  Trade or not, Falcons can't afford first-round...
1  A Local E.R. Doctor and Former Astronaut Talks...
2  MAHB donates comfort kits to front liners at H...
3           Illinois garden centers to reopen in May
4  Lacson on P10M reward for COVID-19 vaccine: ...

                                    meta_description                  domain  \
0  Moving into the top 5 would be difficult for t...            www.espn.com
1  Dr. William Fisher, a former astronaut and fat...    www.houstoniamag.com
2  KUALA LUMPUR (April 23): Malaysia Airports Bhd...  www.theedgemarkets.com
3  Garden centers across the state get the green ...            www.wifr.com
4  Why not invest more in the countrys research ...   newsinfo.inquirer.net
```

```
         date                                            content
0  04-23-2020  Kevin Negandhi dives into the careers of the t...
1  04-23-2020  A few weeks ago, we spoke to an Italian doctor...
2  04-23-2020  KUALA LUMPUR (April 23): Malaysia Airports Bhd...
3  04-23-2020  Garden centers in the region celebrate as Gove...
4  04-23-2020  MANILA, Philippines  Why not invest more in t...
```

Stored the dataframe with top 100 Articles around the globe in the given csv file

[8]:
```
#df_cleaned.to_csv('100Articles.csv')
```

Displaying the top 10 Data from the Dataset 100Articles

[44]:
```
data = pd.read_csv("100Articles.csv")
data.head(10)
```

[44]:
```
   Unnamed: 0                                                url  \
0           0  https://www.espn.com/blog/atlanta-falcons/post...
1           1  https://www.houstoniamag.com/coronavirus/2020/...
2           2  https://www.theedgemarkets.com/article/mahb-do...
3           3  https://www.wifr.com/content/news/Illinois-gar...
4           4  https://newsinfo.inquirer.net/1263337/lacson-o...
5           5  https://www.wyomingnews.com/news/local_news/cr...
6           6  https://packerswire.usatoday.com/2020/04/23/pa...
7           7  https://www.jdsupra.com/legalnews/bitblog-bi-w...
8           8  https://theriver973.iheart.com/content/2020-04...
9           9  https://daytimeconfidential.com/2020/04/23/per...

                                               title  \
0  Trade or not, Falcons can't afford first-round...
1  A Local E.R. Doctor and Former Astronaut Talks...
2  MAHB donates comfort kits to front liners at H...
3           Illinois garden centers to reopen in May
4  Lacson on P10M reward for COVID-19 vaccine: ...
5  CRAFT asks city, county for $260,000 in 2021 f...
6  Packers take LB Zack Baun in final Matt Miller...
7              BitBlog Bi-Weekly Update - April 2020
8                     Harrisburg's Real Rock Variety
9  Perkie's Observations: Jordan Worries About TJ...

                                    meta_description  \
0  Moving into the top 5 would be difficult for t...
1  Dr. William Fisher, a former astronaut and fat...
2  KUALA LUMPUR (April 23): Malaysia Airports Bhd...
3  Garden centers across the state get the green ...
4  Why not invest more in the countrys research ...
5  CHEYENNE  At the beginning of April, flights ...
6  Matt Miller's final mock draft at Bleacher Rep...
```

```
7  While the world at large shelters in place due...
8                                  Real Rock Variety
9  (There were 567 flashbacks today, including tw...

                     domain        date  \
0              www.espn.com  04-23-2020
1      www.houstoniamag.com  04-23-2020
2    www.theedgemarkets.com  04-23-2020
3              www.wifr.com  04-23-2020
4     newsinfo.inquirer.net  04-23-2020
5       www.wyomingnews.com  04-23-2020
6  packerswire.usatoday.com  04-23-2020
7           www.jdsupra.com  04-23-2020
8    theriver973.iheart.com  04-23-2020
9   daytimeconfidential.com  04-23-2020

                                             content
0  Kevin Negandhi dives into the careers of the t...
1  A few weeks ago, we spoke to an Italian doctor...
2  KUALA LUMPUR (April 23): Malaysia Airports Bhd...
3  Garden centers in the region celebrate as Gove...
4  MANILA, Philippines  Why not invest more in t...
5  Your notification has been saved.\n\nThere was...
6  A review and breakdown of predictions for the ...
7  While the world at large shelters in place due...
8                                                NaN
9  Carly's not amused when Cyrus stops by the Met...
```

Taking the top 20 data and getting us the count of occurences of each of the unique values in this column.

[76]:
```
data['domain'].value_counts()[:20]
```

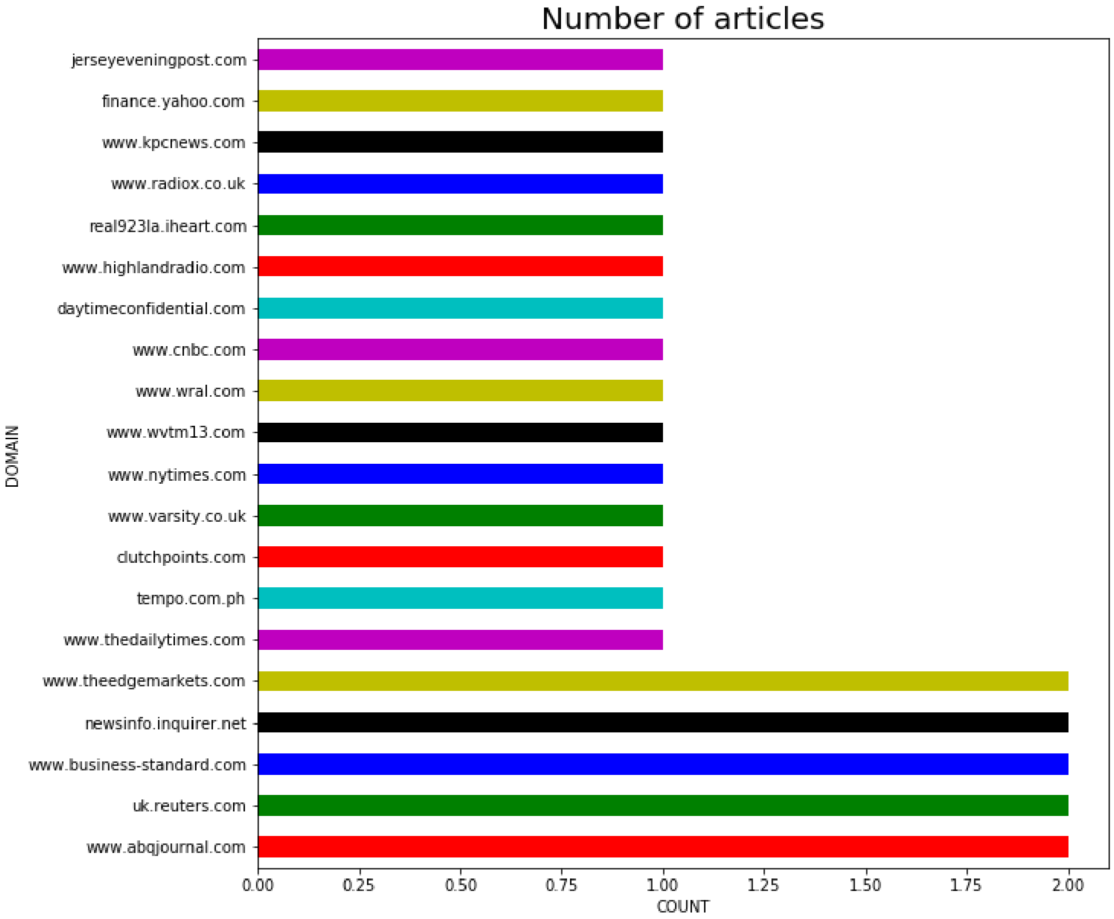
[76]:
```
www.abqjournal.com           2
uk.reuters.com               2
www.business-standard.com    2
newsinfo.inquirer.net        2
www.theedgemarkets.com       2
www.thedailytimes.com        1
tempo.com.ph                 1
clutchpoints.com             1
www.varsity.co.uk            1
www.nytimes.com              1
www.wvtm13.com               1
www.wral.com                 1
www.cnbc.com                 1
daytimeconfidential.com      1
www.highlandradio.com        1
```

```
real923la.iheart.com        1
www.radiox.co.uk            1
www.kpcnews.com             1
finance.yahoo.com           1
jerseyeveningpost.com       1
Name: domain, dtype: int64
```

Plotting the sample Bar plot

```
[116]: data['domain'].value_counts()[:20].plot(kind='barh',figsize=(10, 10),color =␣
         ↪list('rgbkymc'))
       plt.xlabel("COUNT", fontsize = 10)
       plt.ylabel("DOMAIN", fontsize = 10)
       #plt.grid()
       plt.title("Number of articles", fontsize=20)
```

```
[116]: Text(0.5, 1.0, 'Number of articles')
```



3.0.1 Intended methodology

Recognising The Entities:

Flair library: It provides state-of-the-art NLP solutions in a few lines of code. Flair uses a neural language model (Akbik et al., 2018) to assign tags to text data, beating most previous models' accuracy in the process.

Example: We'll tell Flair's sentence tagger to predict entities within the sentence:
"Boris went to Seattle with Donald to meet Microsoft"

[9]:

```
tagger = SequenceTagger.load('ner')
```

```
2020-04-25 15:48:39,903 loading file /Users/asthajain/.flair/models/en-ner-
conll03-v0.4.pt
```

[10]:

```
sentence = Sentence('Boris went to Seattle with Donald to meet Microsoft')
tagger.predict(sentence)
for entity in sentence.get_spans('ner'):
    print(entity)
```

```
../torch/csrc/utils/tensor_numpy.cpp:141: UserWarning: The given NumPy array is
not writeable, and PyTorch does not support non-writeable tensors. This means
you can write to the underlying (supposedly non-writeable) NumPy array using the
tensor. You may want to copy the array to protect its data or make it writeable
before converting it to a tensor. This type of warning will be suppressed for
the rest of this program.
```

```
PER-span [1]: "Boris"
LOC-span [4]: "Seattle"
PER-span [6]: "Donald"
ORG-span [9]: "Microsoft"
```

It successfully identified all four entities, and was able to recognise the two people (PER), one location (LOC) and one organisation (ORG).

Building The Graph:

NetworkX package: To build the actual social network, we'll use NetworkX package.

We'll use the combinations functionality from itertools to, well, find all possible combinations given a list of items. First, we'll sort our entities alphabetically — this is to ensure that in every pair we find , the alphabetically superior entity appears on the left hand side, and we don't duplicate pairs of A-B and B-A, for instance:

[11]:

```
sentence.to_dict(tag_type='ner')
sent_dict = sentence.to_dict(tag_type='ner')
df_ner = pd.DataFrame(data={
  'entity': [entity['text'] for entity in sent_dict['entities']],
  'type': [entity['type'] for entity in sent_dict['entities']]
 }
)
df_ner=df_ner[df_ner['type'].isin(['PER', 'ORG'])]
df_ner = df_ner.sort_values('entity')
combs = list(combinations(df_ner['entity'], 2))
df_links = pd.DataFrame(data=combs, columns=['from', 'to'])
df_links
```

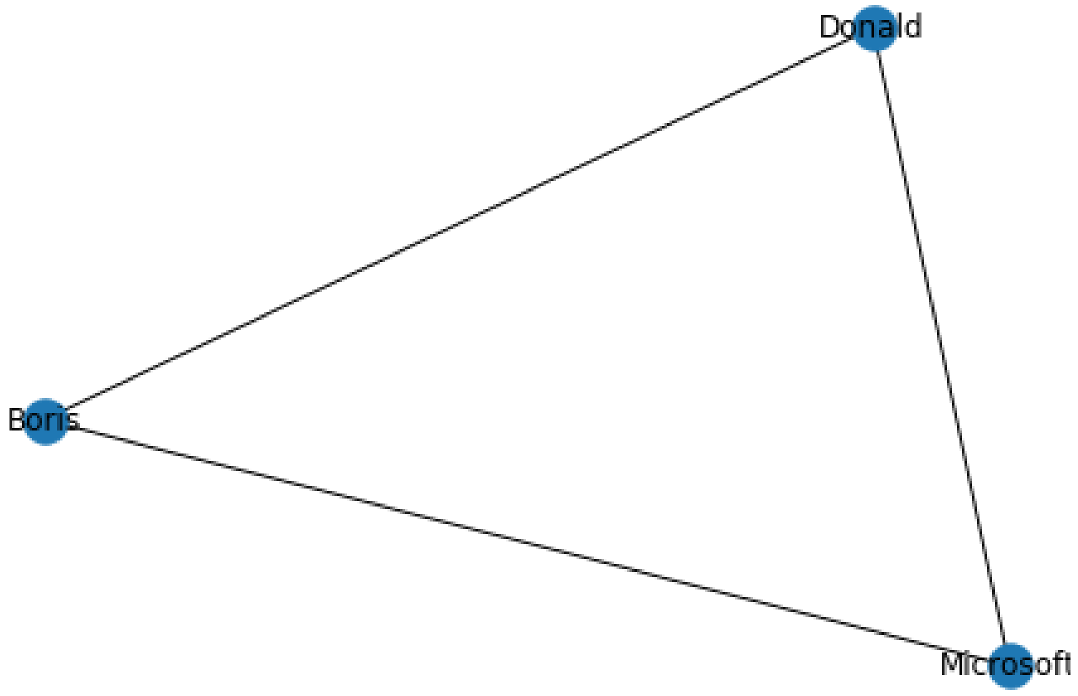
[11]:

```
     from         to
0   Boris     Donald
1   Boris  Microsoft
2  Donald  Microsoft
```

We are ready to create our graph and plot it!

[14]:

```
G = nx.Graph()
for link in df_links.index:
  G.add_edge(df_links.iloc[link]['from'],
             df_links.iloc[link]['to'])
nx.draw(G, with_labels=True)
```



We have demonstrated the sample by using just one dummy sentence and successfully able to form a csv file with top 100 articles. We look forward to work the same way with all the 100 articles in the project.