Data Intensive Computing (CSE 587) – Lab 2

PROBLEM STATEMENT:

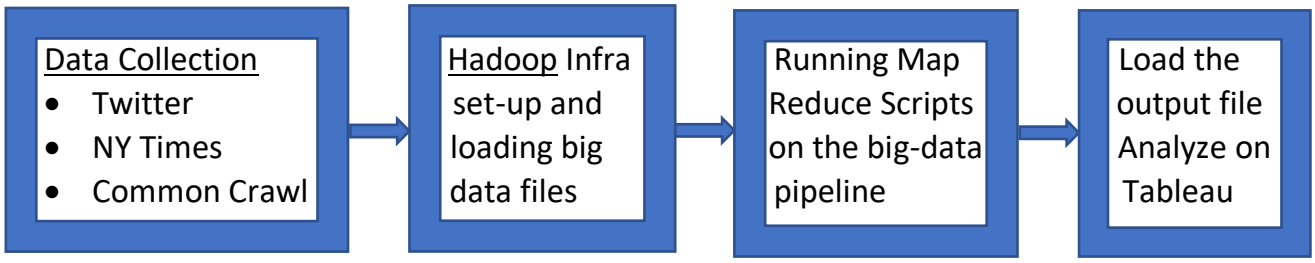
Collect data about from at least three sources, one opinion-based social media in twitter, research data in New York Times, and the third is the common crawl data for the same topic or key phrase, and similar time periods. Process the three data sets collected individually using classical big data methods. Compare the outcomes using popular visualization methods.

SOLUTION:

The topic of interest chosen for data collection and visualization is ***"IMMIGRATION".***

Topic Related Key-words: US Immigration, U.S. Immigrant, US Civil wars, U.S. Citizenship and Immigration Services, USCIS, immigrant visa, Migrator, Migration, Refugees, Pilgrim, Green Card, Emigrate, Emigration, Illegal immigration, Legal immigration, border wall, etc.

The entire Data Flow is as below:
Data is collected and scrapped from various sources (Twitter, New York Times and Common Crawl). The data is then loaded onto the Hadoop infrastructure and Map-Reduce scripts are run on it. The word count and word co-occurrence files are then loaded into the Tableau for visualizing the most frequent words as '*word-cloud*'.



→ *Prototype data collection*

Social Media – Tweets from Twitter

1. TwitteR library is loaded and respective keys are set-up and authorized.
2. Tweets are collected using "searchTwitter" command.
3. Tweets are collected (**> 20,000 tweets**) for each day and stored separately and collectively.

Research Data - New York Times

1. articleAPI key is set-up using NYTimes developer's account.

2. Articles are selected for various sub-topics using "api.search" command for the entire year 2019.
3. URLs are collected for each topic (Total URLs – 257) and the data is scrapped from each URL using "BeautifulSoup". More than **500 articles** have been collected for NY Times.

Common Crawl

1. Code Reference: - https://www.bellingcat.com/resources/2015/08/13/using-python-to-mine-common-crawl/
2. Libraries are loaded and an **"index-list"** from common crawl site is chosen. The domain passed on the index list is **"americanimmigrationcouncil.org".**
3. All the URLs are collected (3671 hits with 1393 links discovered).
4. Data is then scrapped from each URL using "BeautifulSoup". More than **500 articles** have been collected for Common Crawl.

→ *Set up big data infrastructure*

Hadoop Infrastructure is set-up for storing and analyzing the big data pipeline. Install a Hadoop infrastructure in the one of the following ways:
(i) cloudera docker image from the document provided, or
(ii) virtual machine (VM) image for data storage in HDFS and Hadoop infrastructure, or
(iii) amazon aws/ any other.

For **NY Times** data, the VM is installed for data storage for loading data and running map reduce files. README.txt file contains the instructions on setting the VM.
For **Twitter** and **Common Crawl**, Amazon AWS is utilized.

→ *Data Pre and Post Processing*

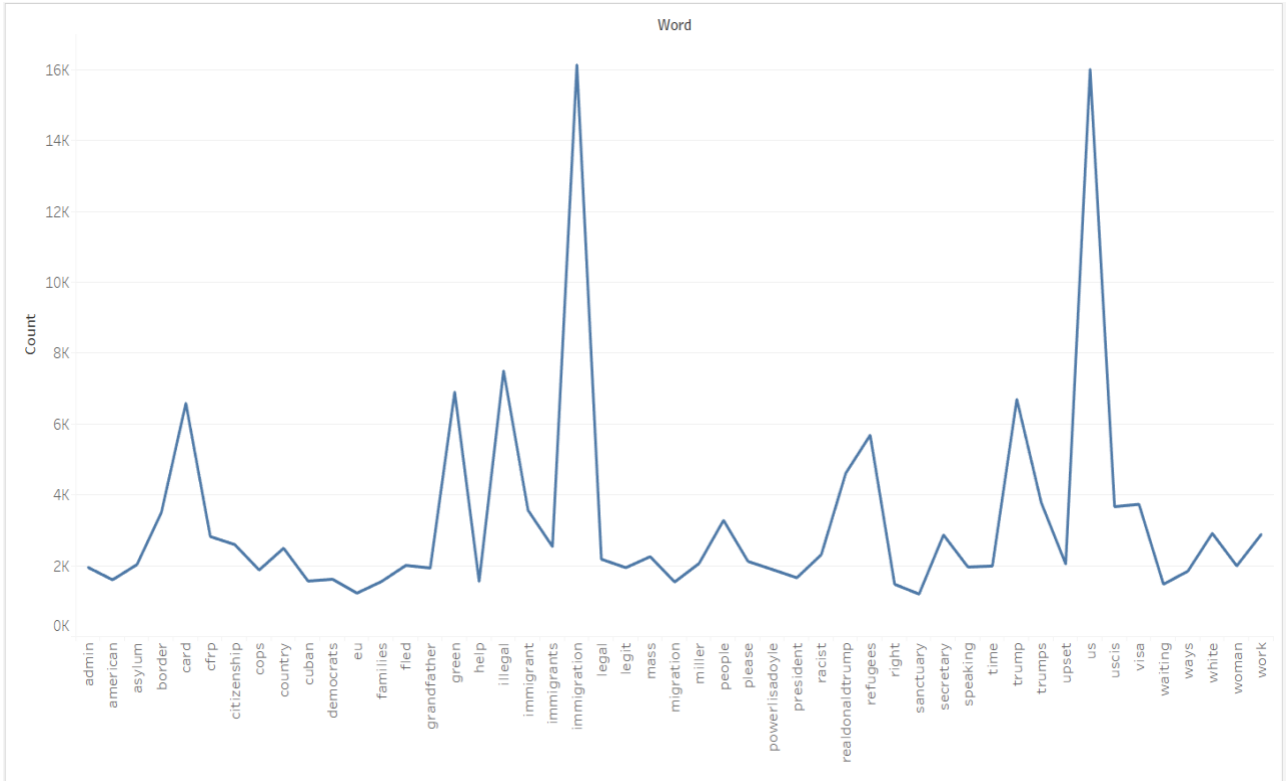
Pre-processing is done on each of the data files generated from the data collection. Using python, Mapper and Reducer scripts have been to written to remove all the punctuations (?, ! etc.), unwanted symbols/characters (@,# etc.), **stop words**(said, would etc.). The output of mapper is then fed to the Reducer script where the data is grouped using key, value pair where keys are the words and the value is the count of each word. Reducer gives the collective **count of the words** as output. Using this file as input a python code is used to calculate the top 50 words and prepended it in the respective files.
In case of co-occurrence the top 50 words are read from the input file and each of these words are searched in the input file articles. These words and their **co-occurring words** are then treated as keys and fed to the reducer.
The output obtained for both "word-count" and "word-cooccurrence" is then converted into a csv file and is further used for analyzing and visualization.

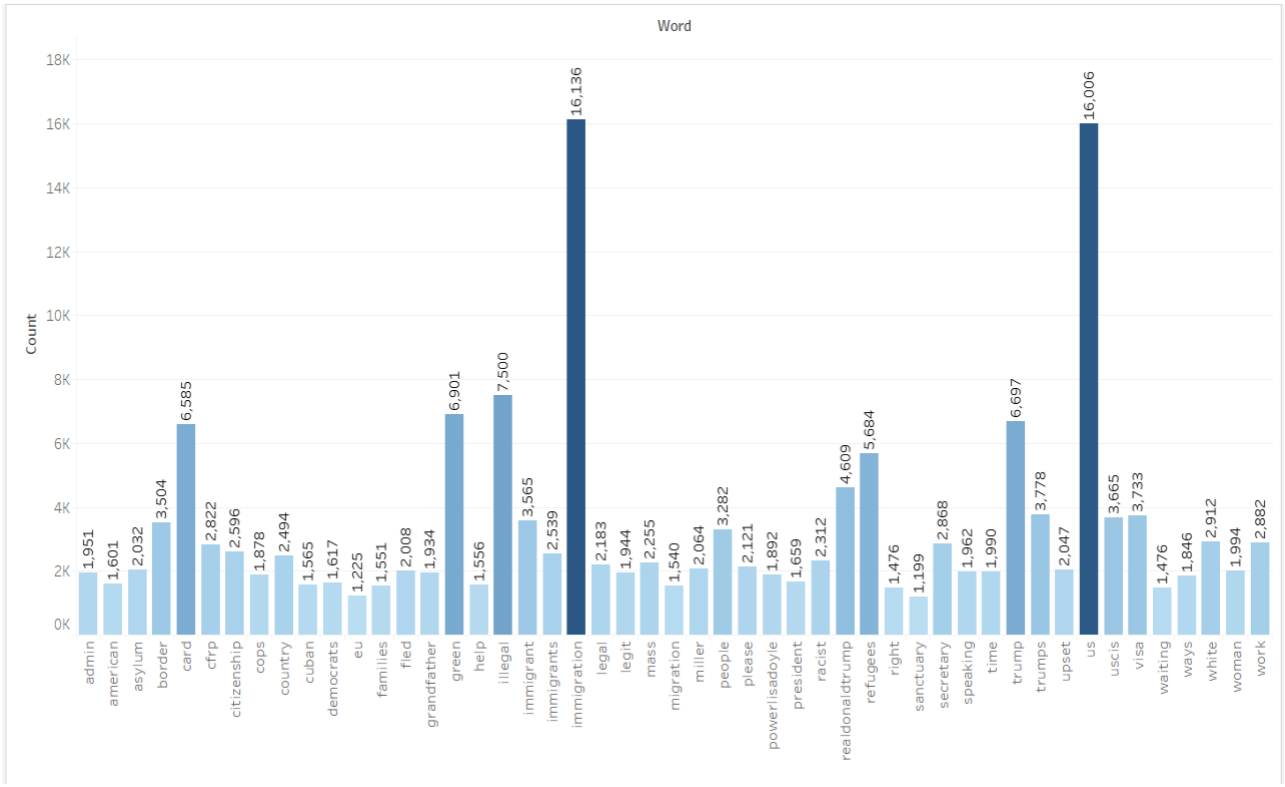
→ *Analyze and visualize*

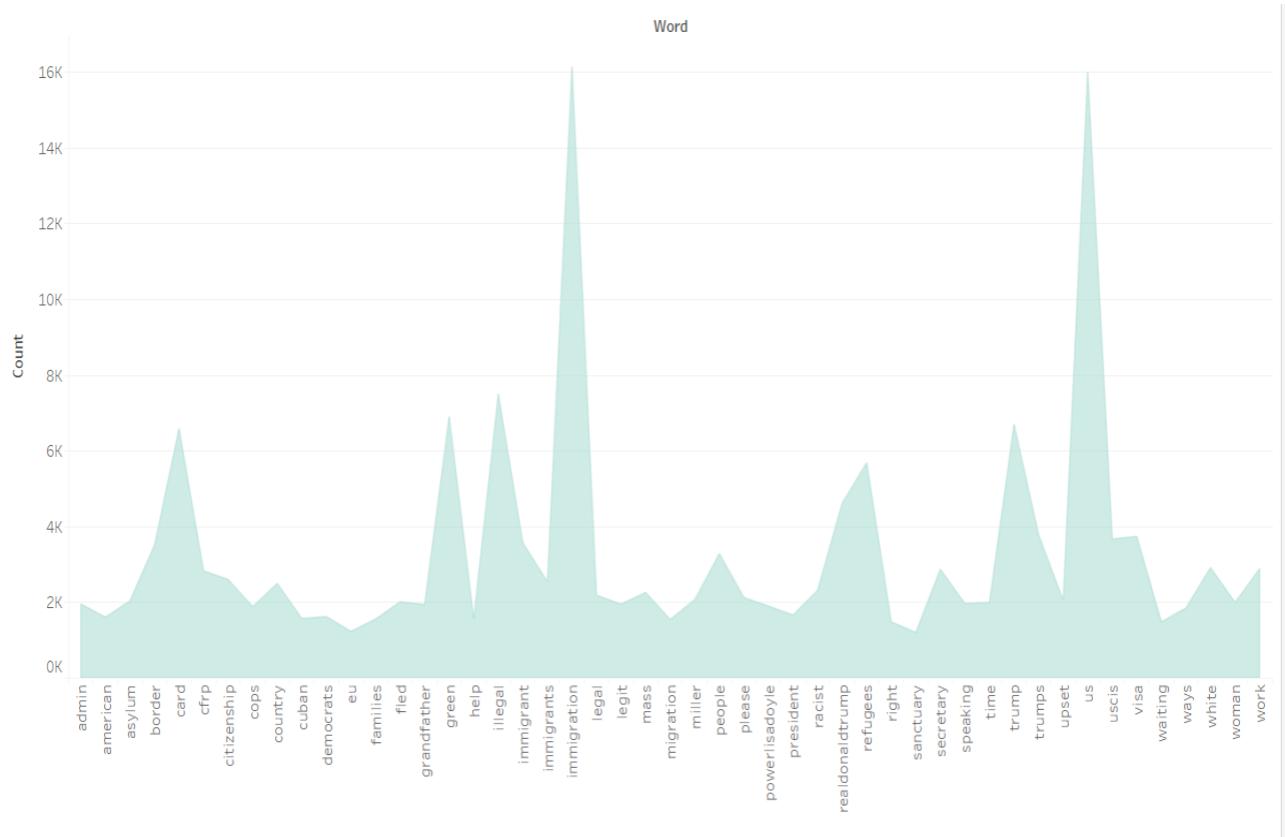
Twitter Top 50 Word Count

Word Cloud

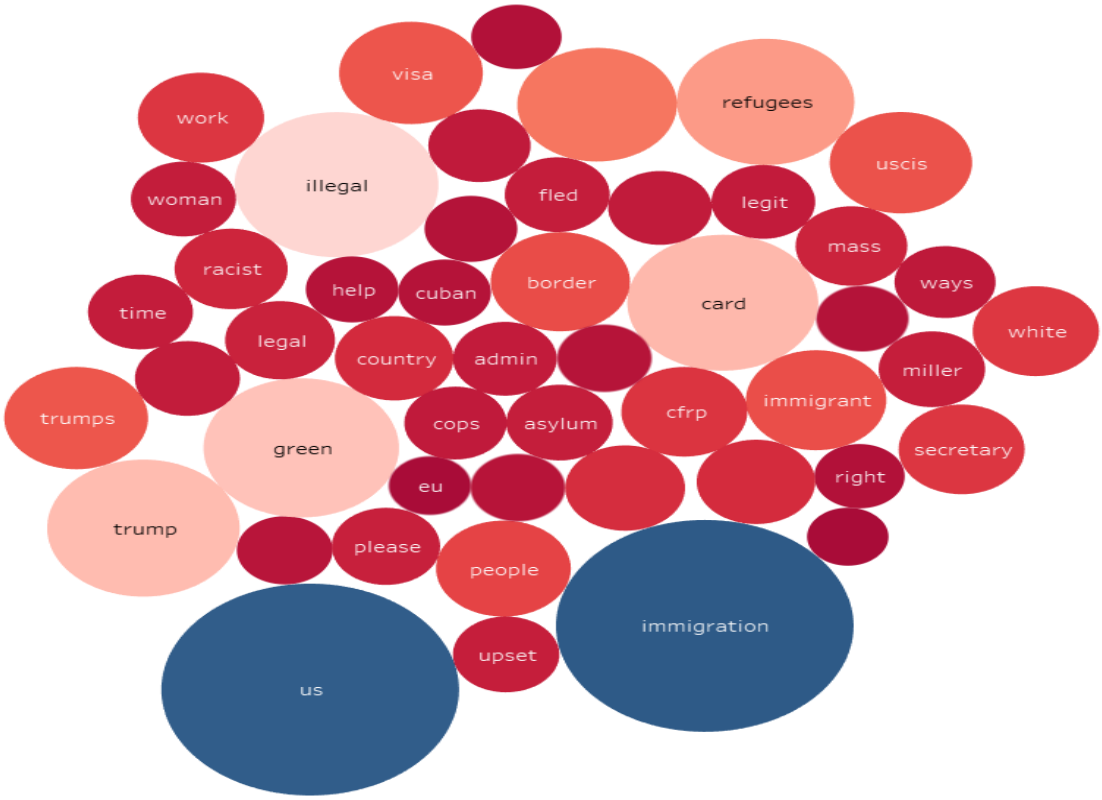
Line, Bar and Area Plot




Word
Count
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
admin 1,951
american 1,601
asylum 2,032
border 3,504
card 6,585
cfrp 2,822
citizenship 2,596
cops 1,878
country 2,494
cuban 1,565
democrats 1,617
eu 1,225
families 1,551
fled 2,008
grandfather 1,934
green 6,901
help 1,556
illegal 7,500
immigrant 3,565
immigrants 2,539
immigration 16,136
legal 2,183
legit 1,944
mass 2,255
migration 1,540
miller 2,064
people 3,282
please 2,121
powerlisadoyle 1,892
president 1,659
racist 2,312
realdonaldtrump 4,609
refugees 5,684
right 1,476
sanctuary 1,199
secretary 2,868
speaking 1,962
time 1,990
trump 6,697
trumps 3,778
upset 2,047
us 16,006
uscis 3,665
visa 3,733
waiting 1,476
ways 1,846
white 2,912
woman 1,994
work 2,882

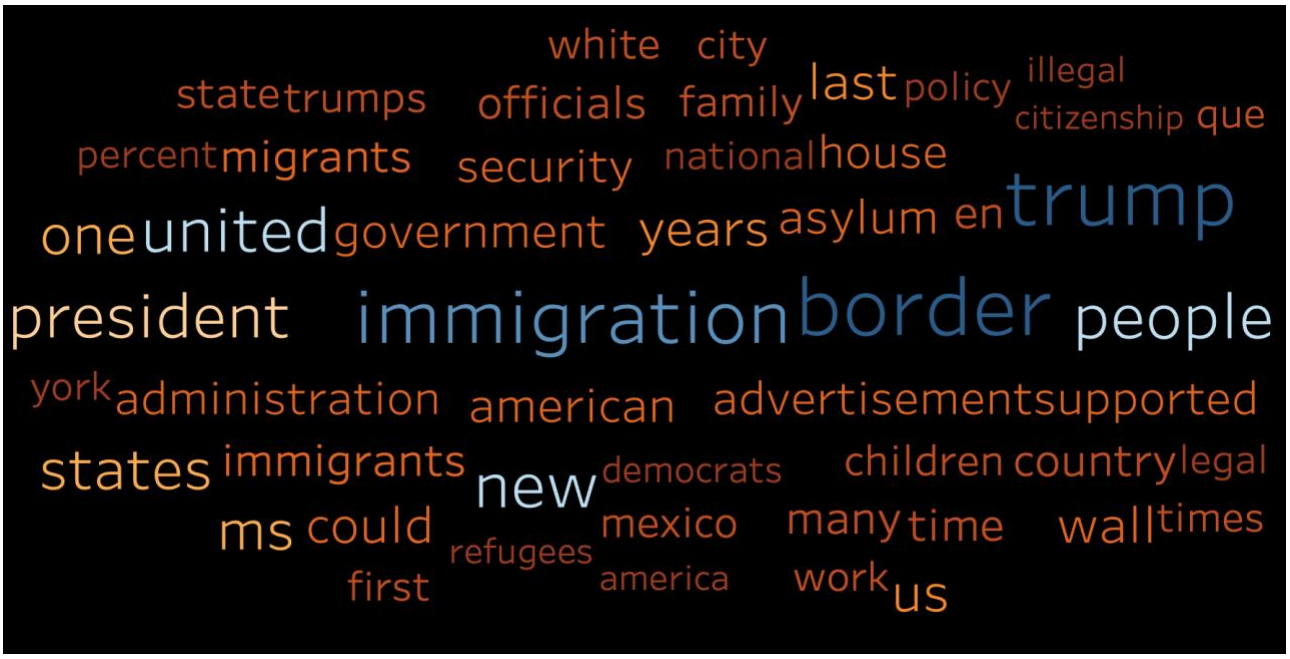

Word
Count
16K
14K
12K
10K
8K
6K
4K
2K
0K
admin
american
asylum
border
card
cfrp
citizenship
cops
country
cuban
democrats
eu
families
fled
grandfather
green
help
illegal
immigrant
immigrants
immigration
legal
legit
mass
migration
miller
people
please
powerlisadoyle
president
racist
realdonaldtrump
refugees
right
sanctuary
secretary
speaking
time
trump
trumps
upset
us
uscis
visa
waiting
ways
white
woman
work

Round Plot

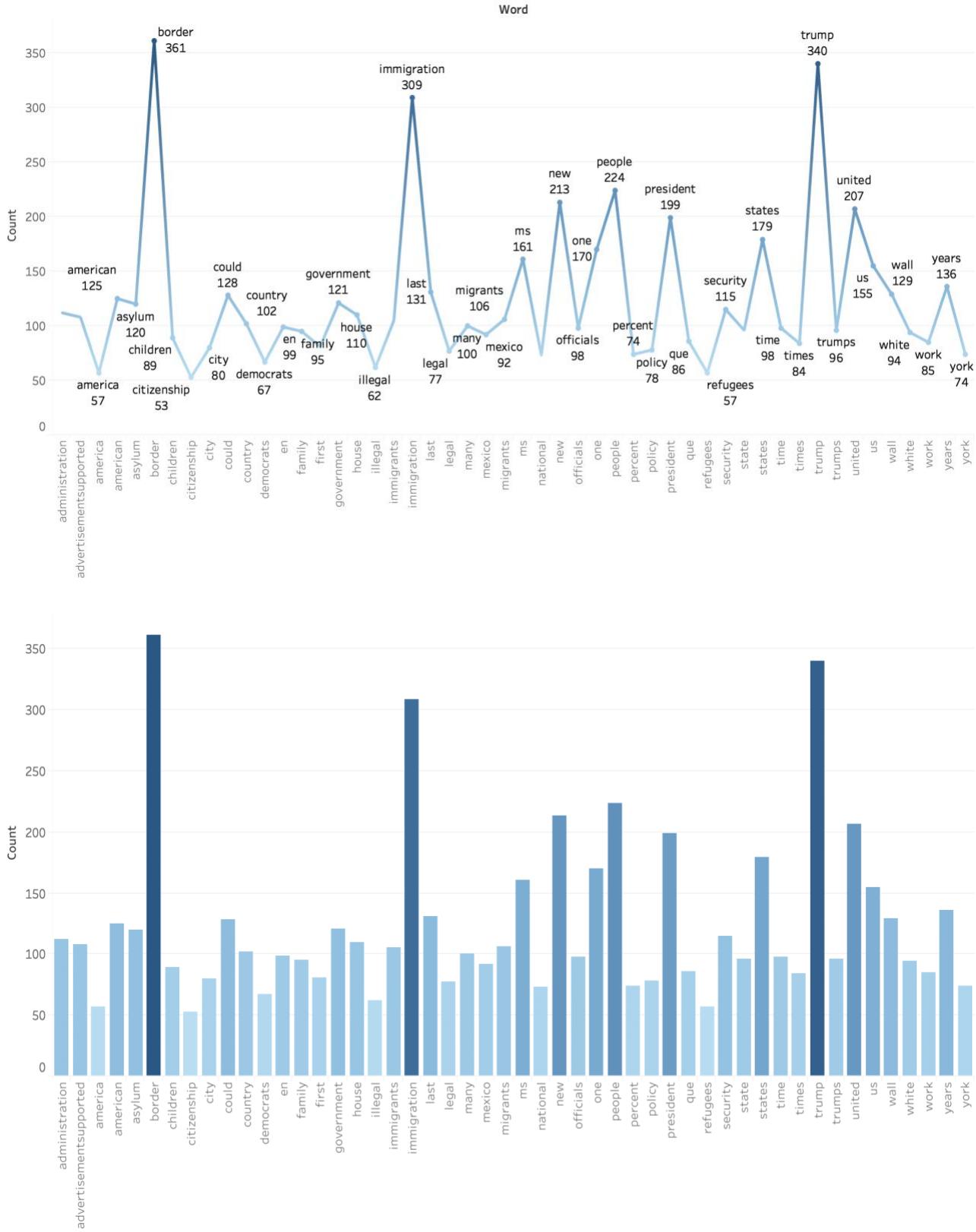


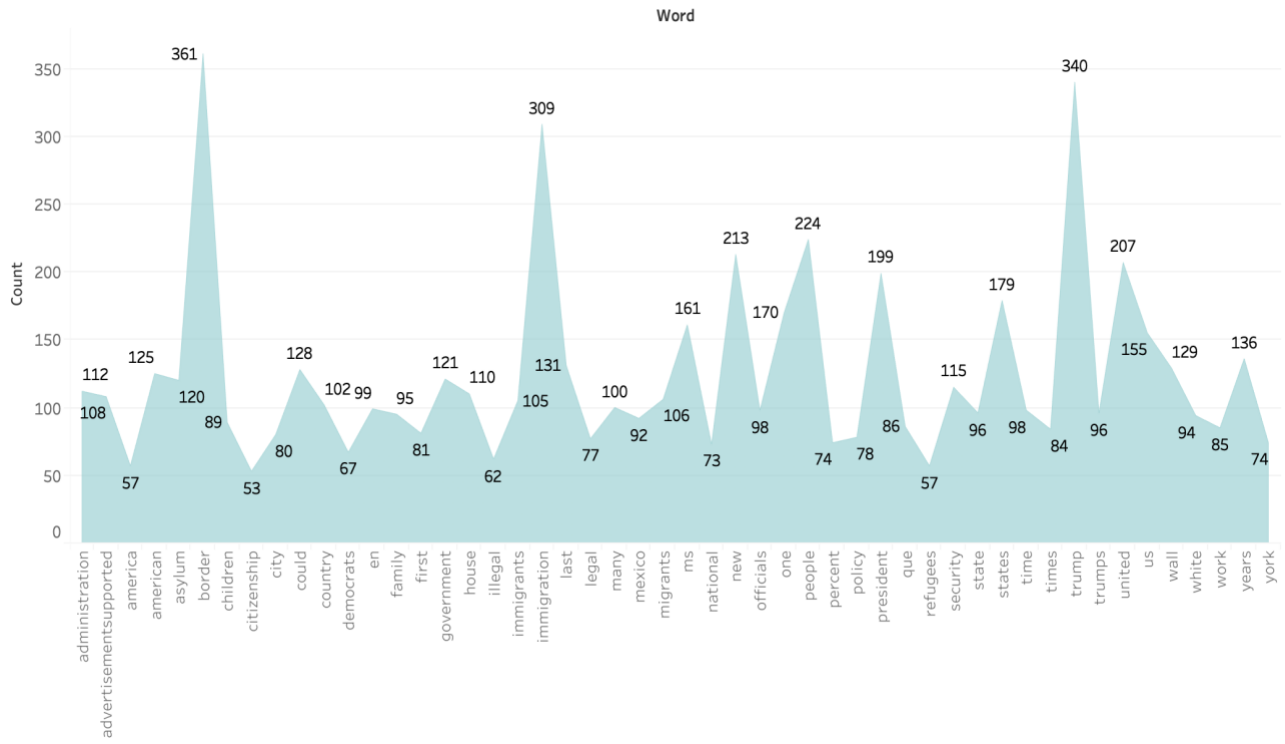
New York Times Top 50 Word Count

Word Cloud

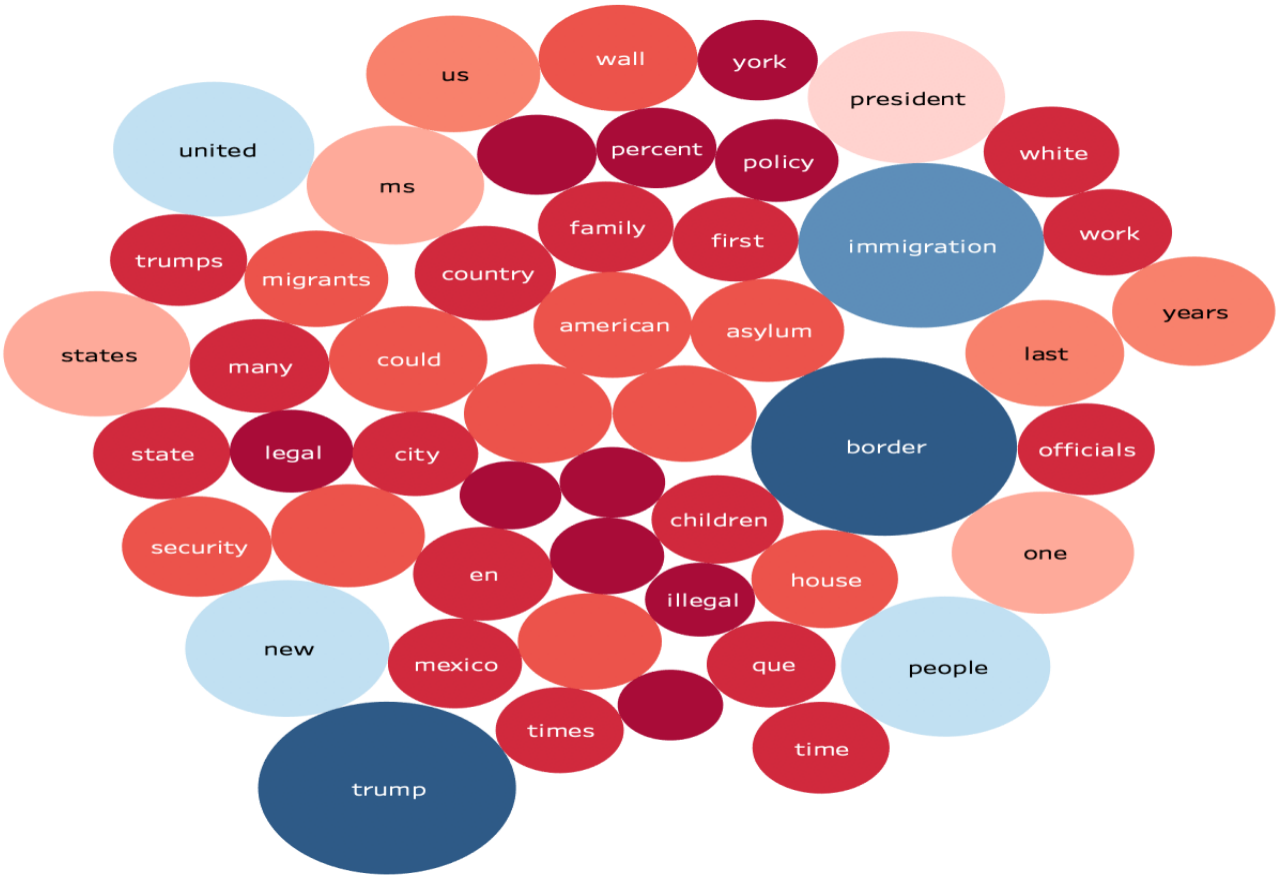


Line, Bar and Area Plot



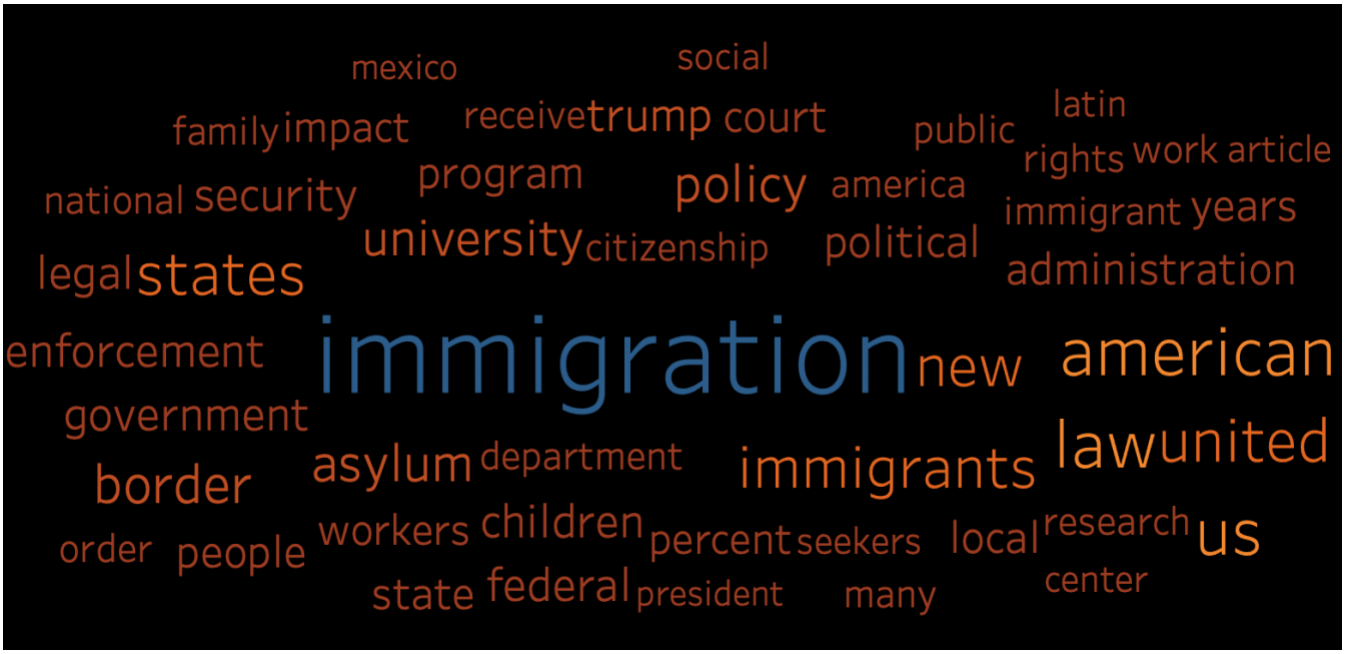

Word
Count
0
50
100
150
200
250
300
350
112
108
57
125
120
361
89
53
80
128
102
99
67
95
81
121
110
62
105
309
131
77
100
92
106
161
73
213
98
170
74
224
78
199
86
57
115
96
179
98
84
340
96
207
155
129
94
136
85
74
administration
advertisementsupported
america
american
asylum
border
children
citizenship
city
could
country
democrats
en
family
first
government
house
illegal
immigrants
immigration
last
legal
many
mexico
migrants
ms
national
new
officials
one
people
percent
policy
president
que
refugees
security
state
states
time
times
trump
trumps
united
us
wall
white
work
years
york

Round Plot

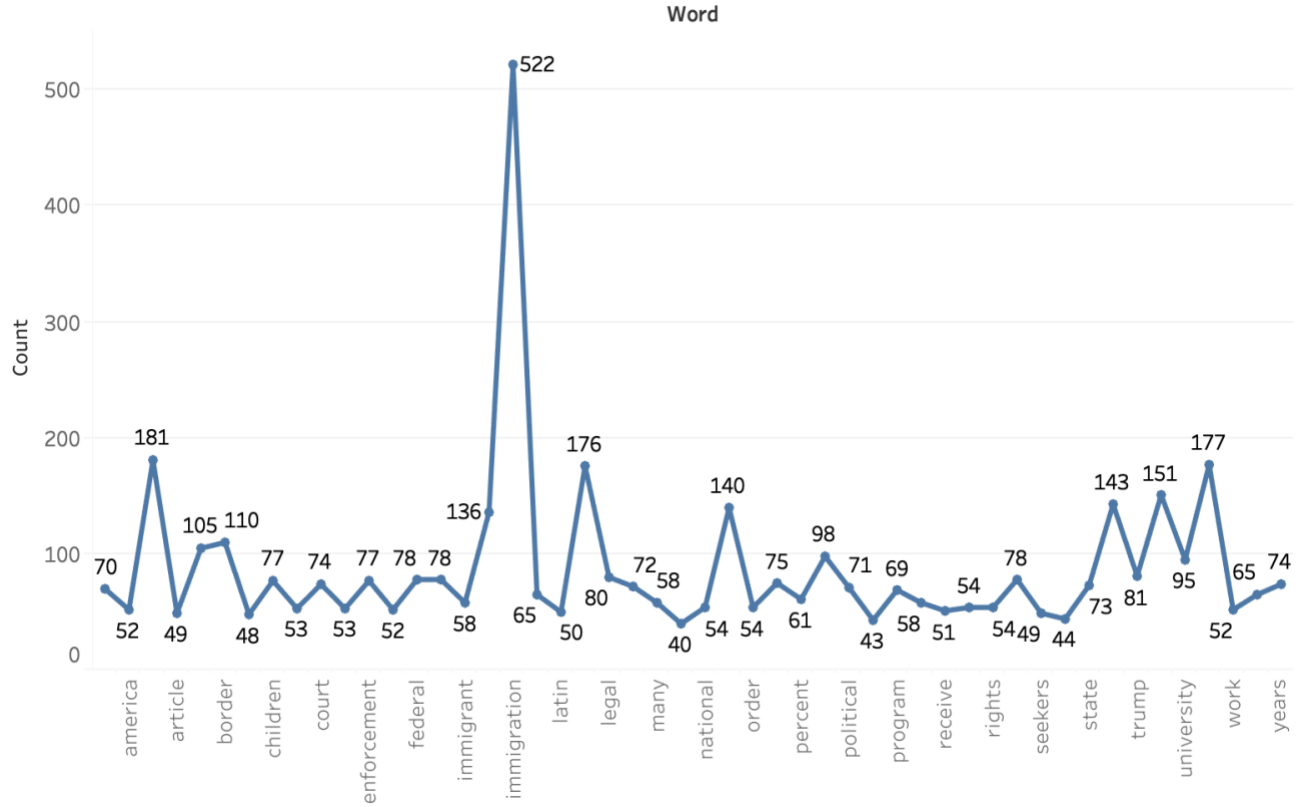

us
wall
york
president
united
percent
policy
white
ms
family
first
immigration
work
trumps
migrants
country
years
american
asylum
states
many
could
last
state
legal
city
border
officials
children
security
one
en
house
illegal
new
mexico
que
people
times
time
trump

Common Crawl Top 50 Word Count

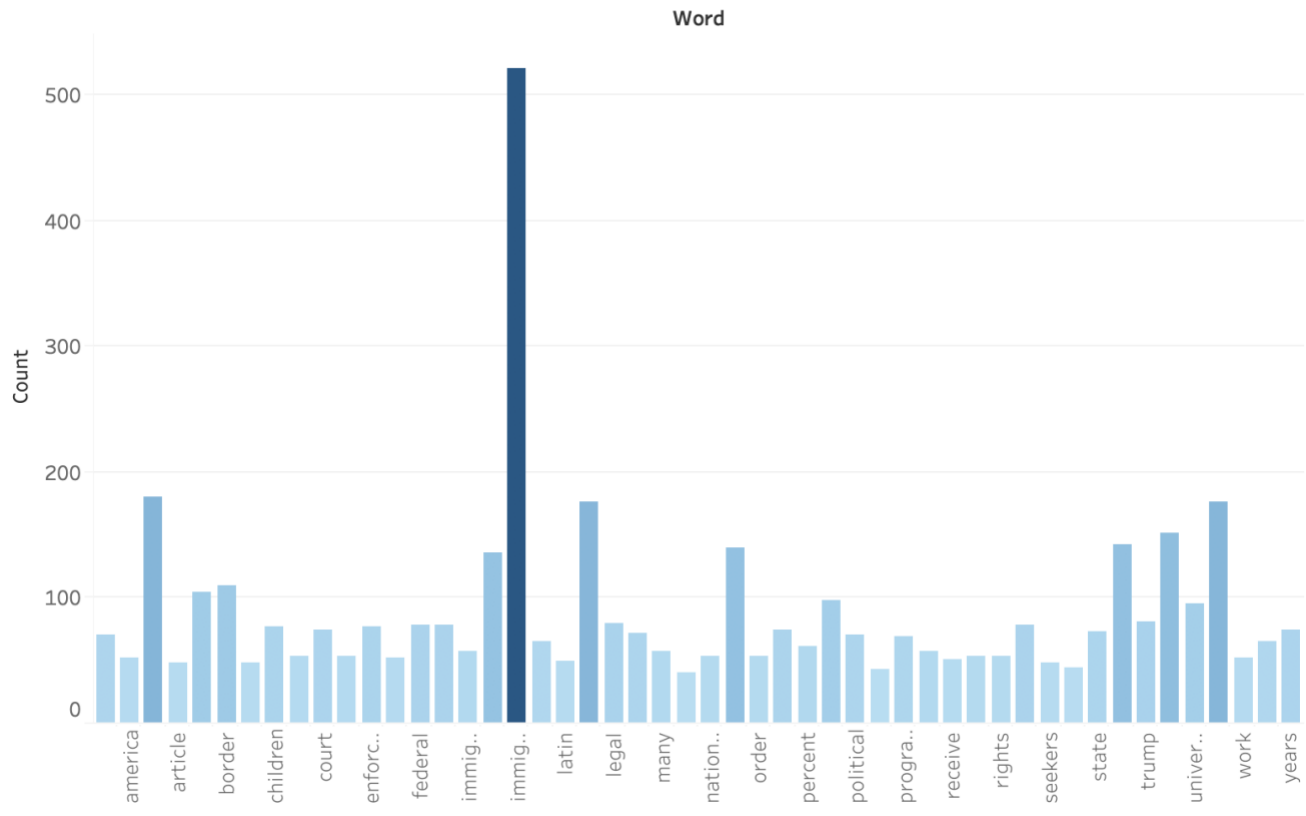
Word Cloud

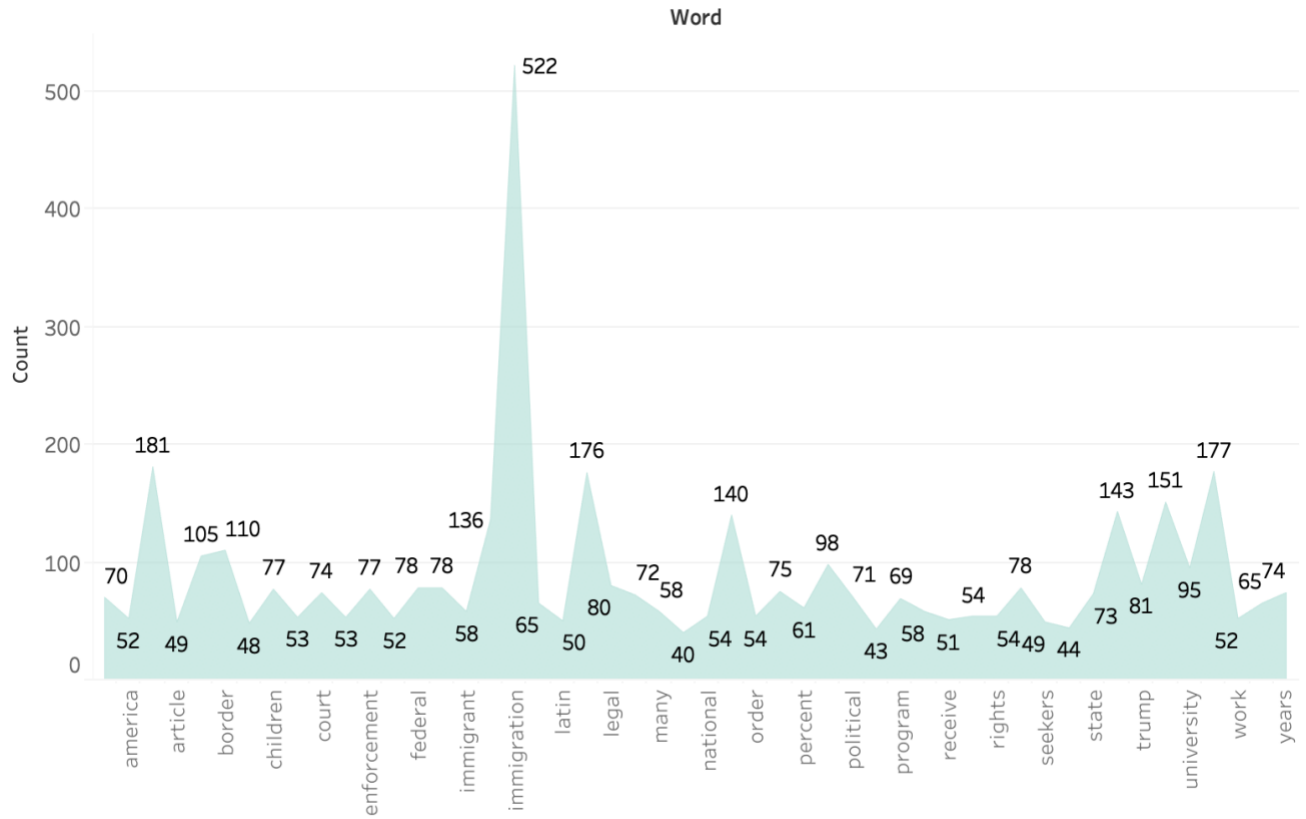


Line, Bar and Area Plot

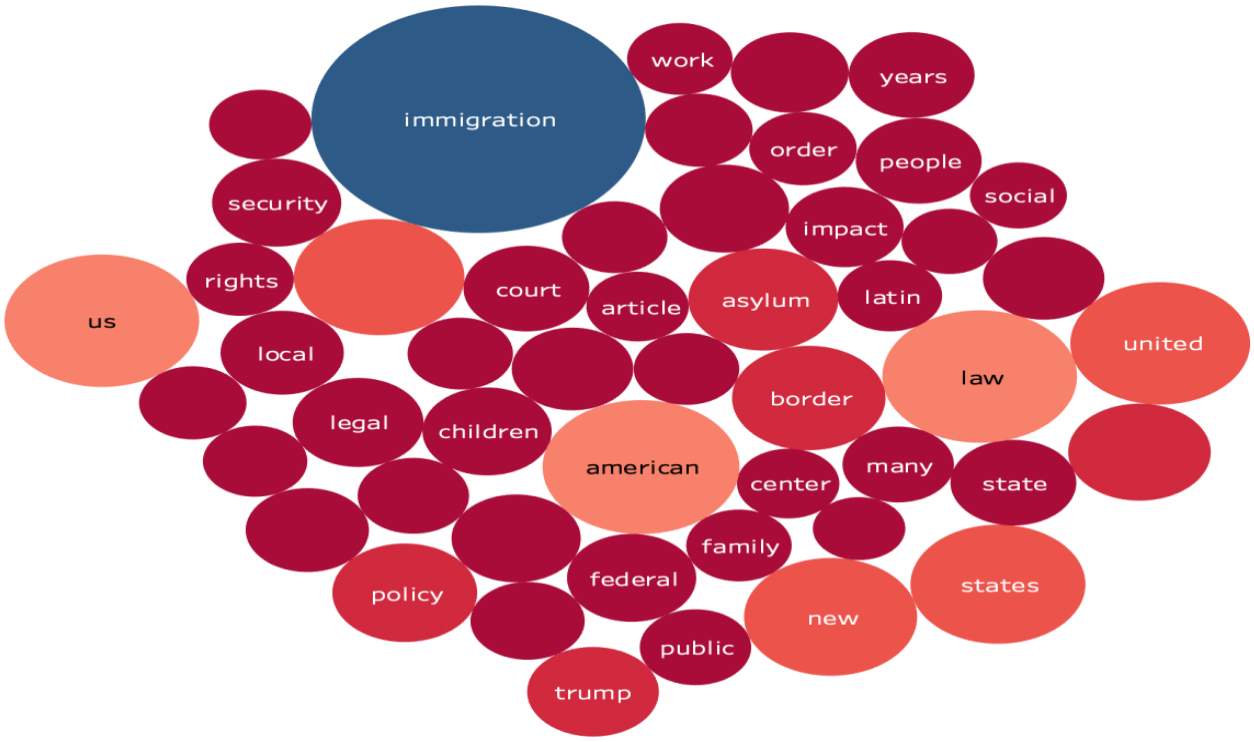


TEAM MEMBERS: - PRADEEP KUMAR JOSHI, ABHISHEK GOSWAMI



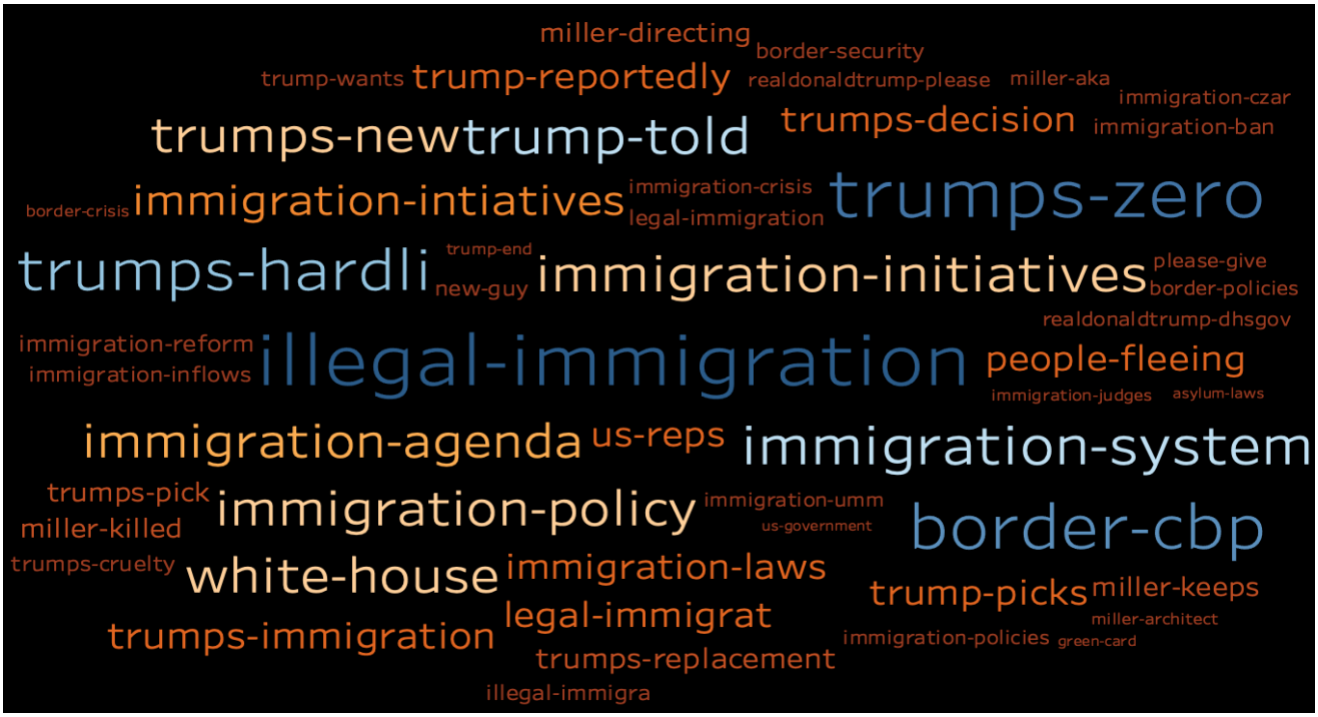


Round Plot

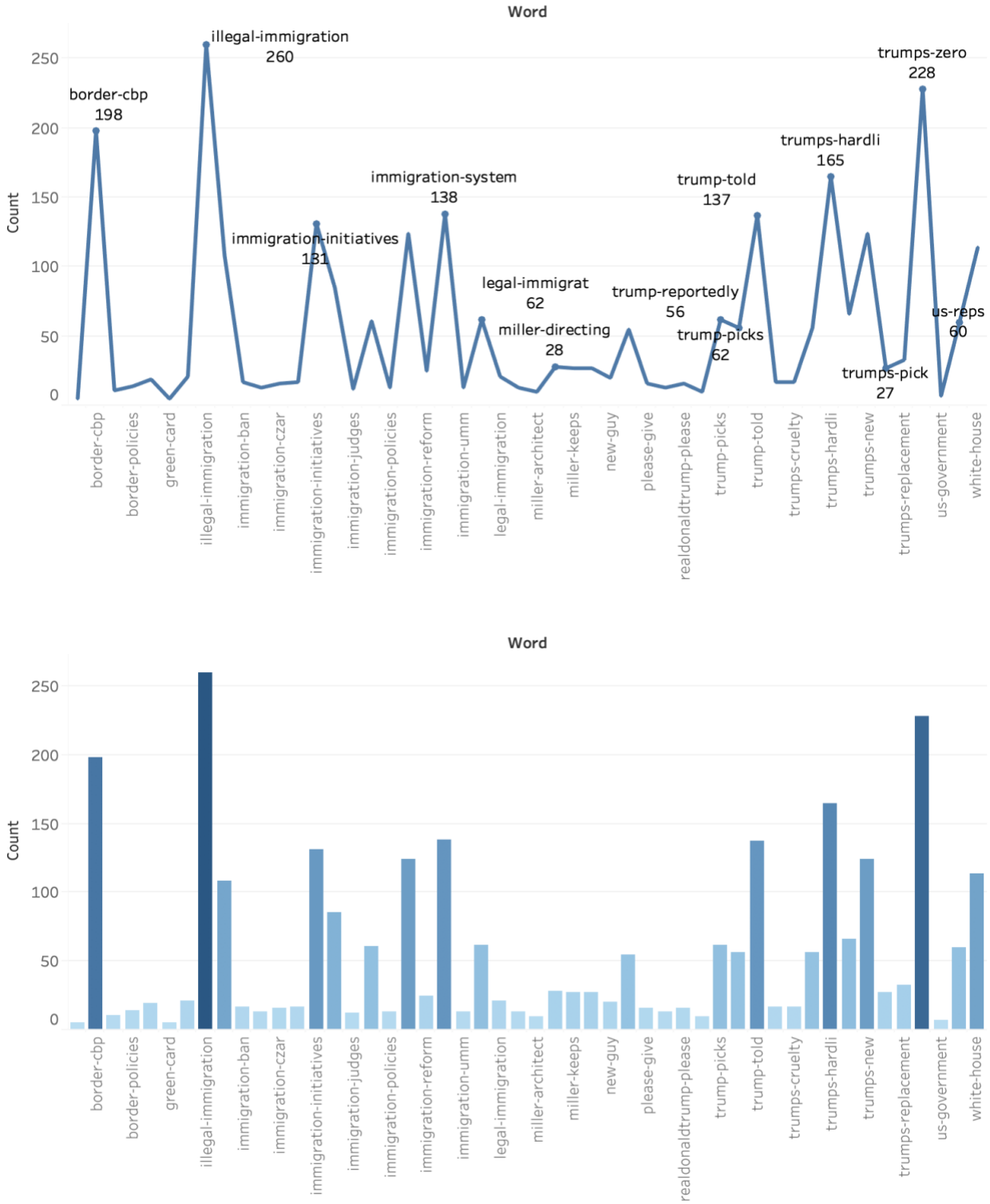


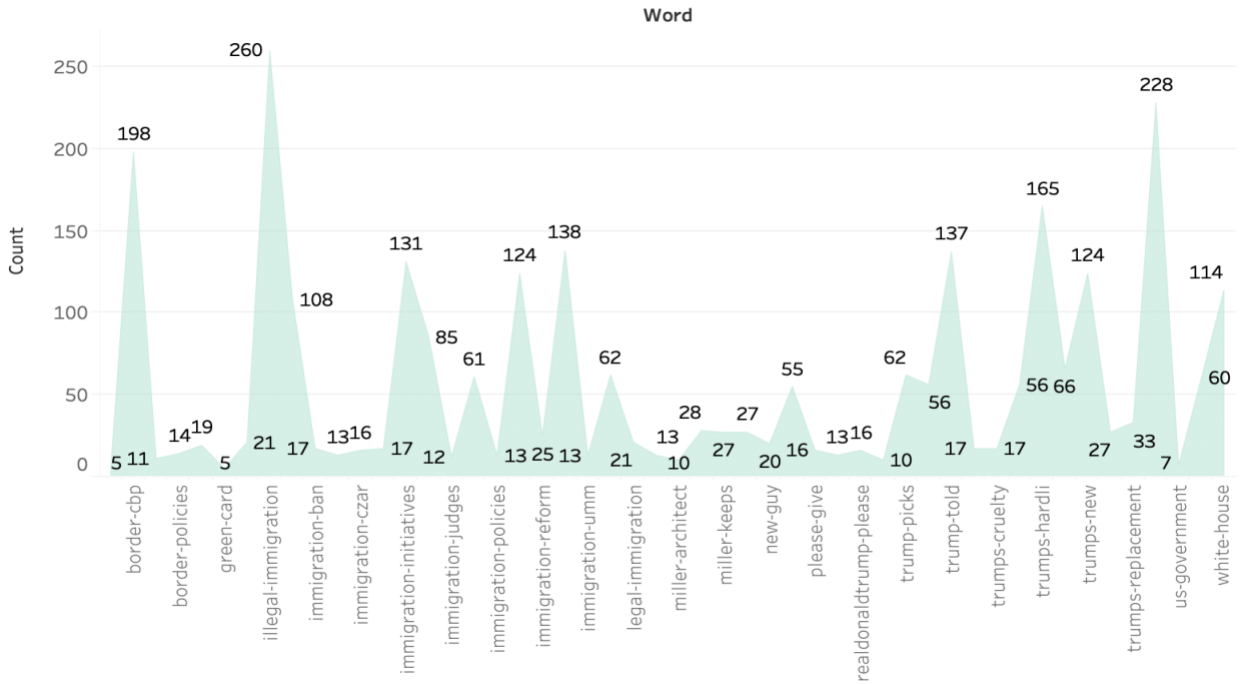
Twitter Top 50 Word Co-occurrence

Word Cloud



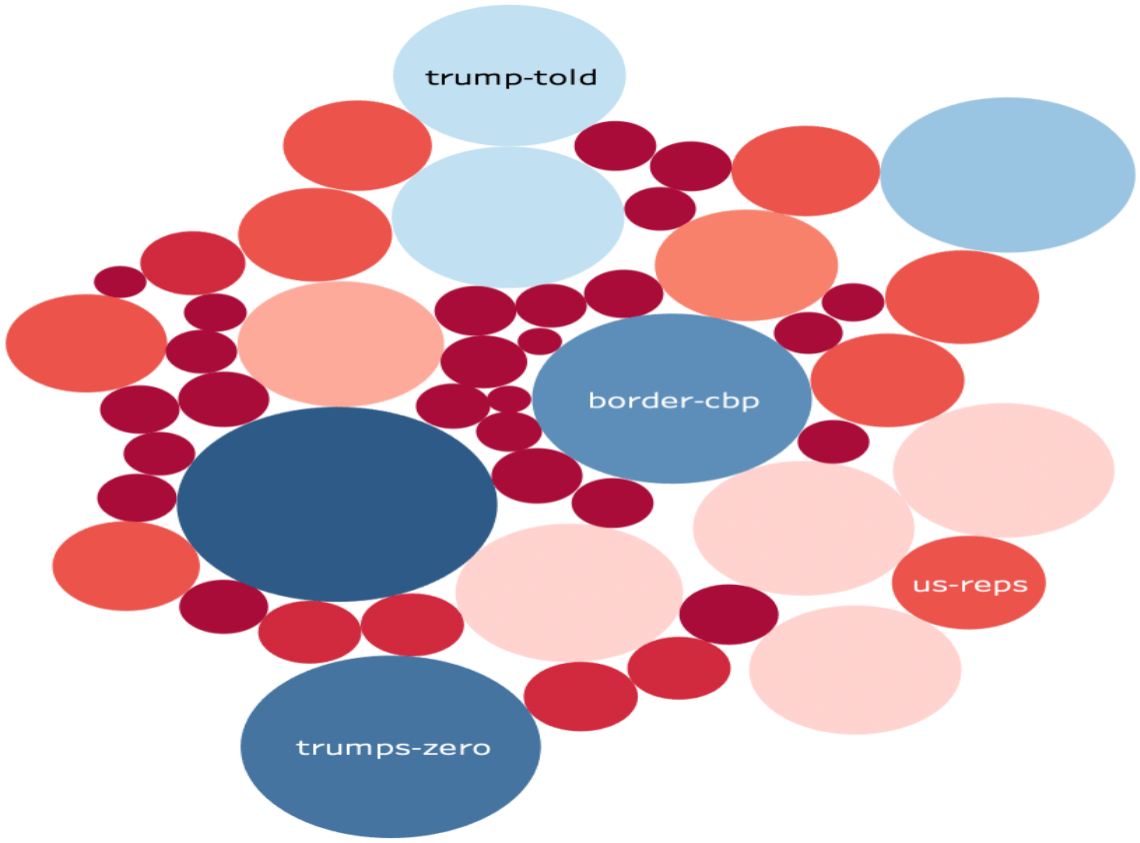
Line, Bar and Area Plot





Round Plot

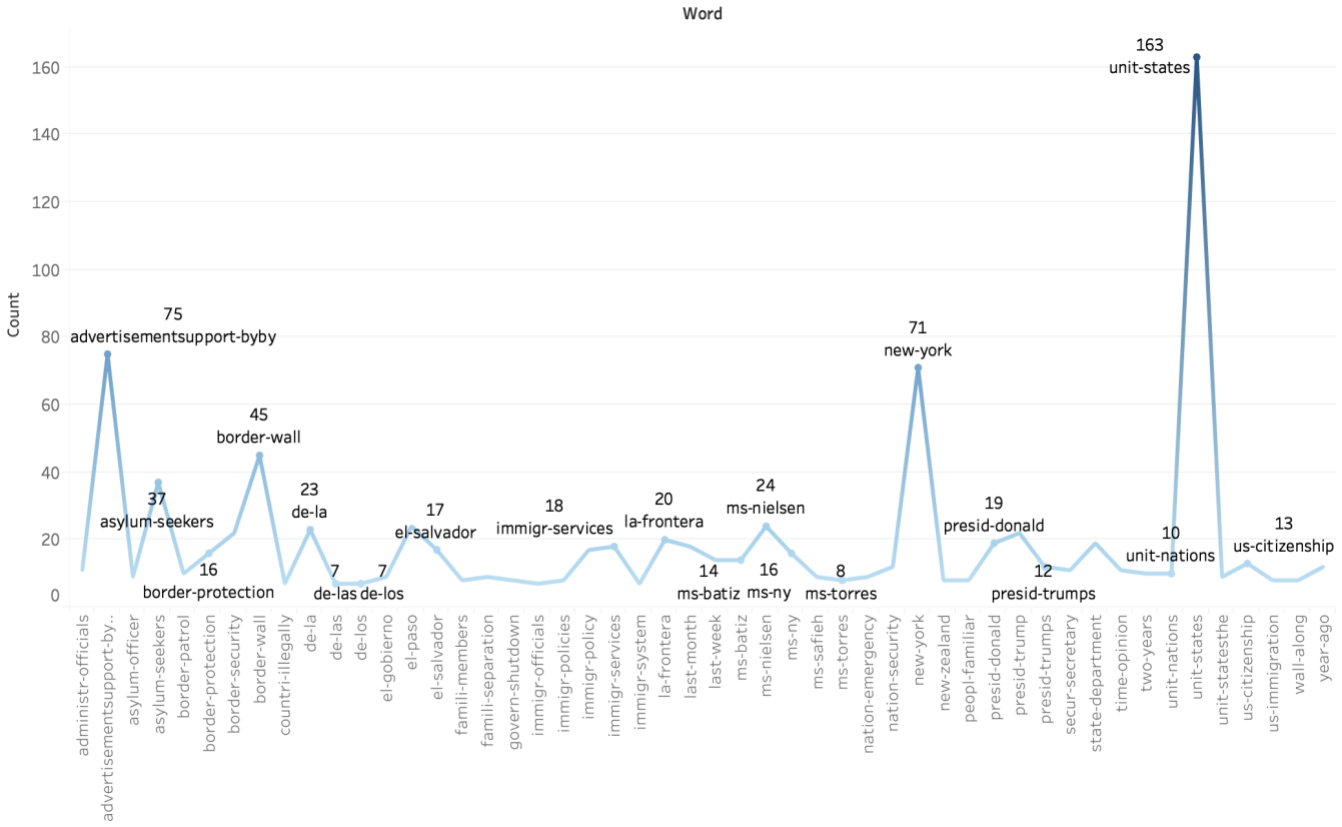
Round Plot

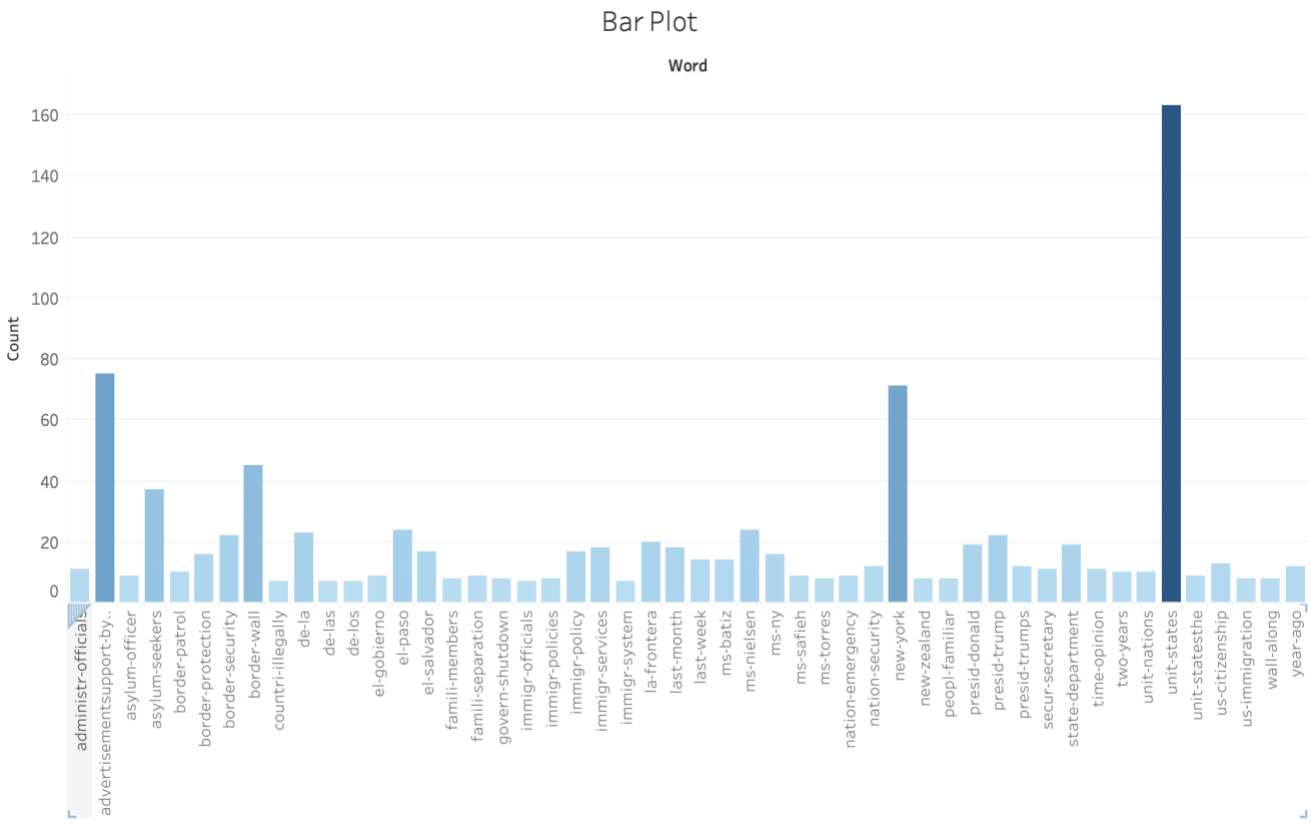


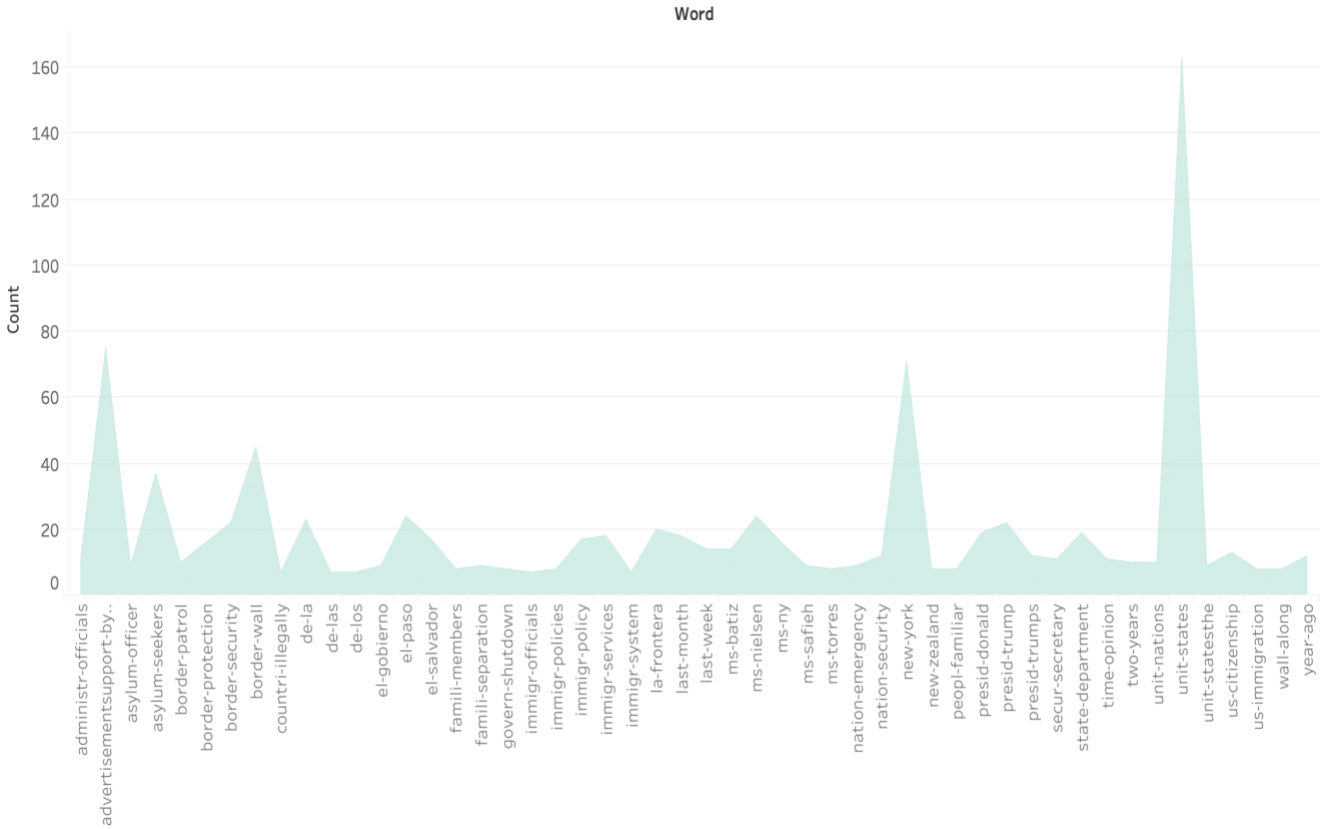
New York Times Top 50 Word Co-occurrence

Word Cloud

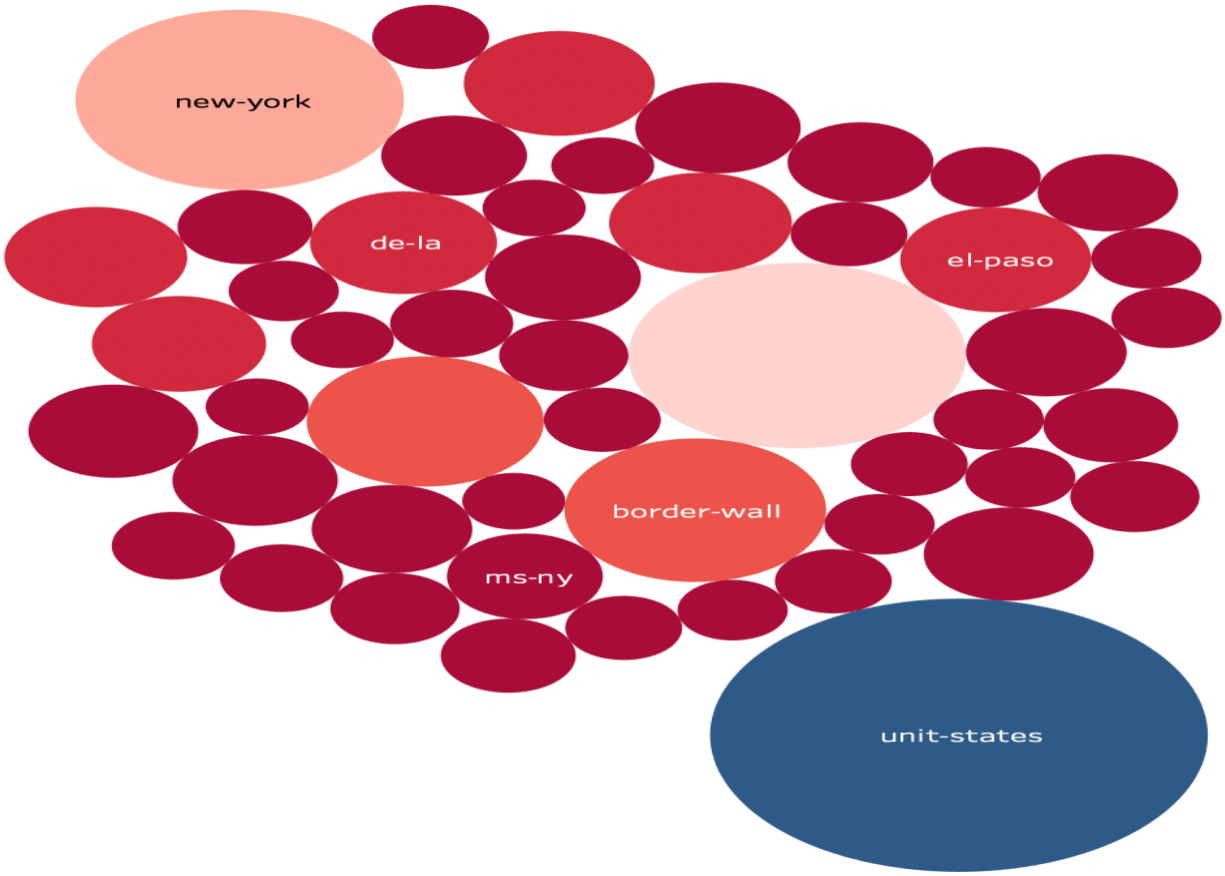
Line, Bar and Area Plot




Bar Plot
Word
Count
0
20
40
60
80
100
120
140
160
administr-officials
advertisementsupport-by..
asylum-officer
asylum-seekers
border-patrol
border-protection
border-security
border-wall
countri-illegally
de-la
de-las
de-los
el-gobierno
el-paso
el-salvador
famili-members
famili-separation
govern-shutdown
immigr-officials
immigr-policies
immigr-policy
immigr-services
immigr-system
la-frontera
last-month
last-week
ms-batiz
ms-nielsen
ms-ny
ms-safieh
ms-torres
nation-emergency
nation-security
new-york
new-zealand
peopl-familiar
presid-donald
presid-trump
presid-trumps
secur-secretary
state-department
time-opinion
two-years
unit-nations
unit-states
unit-statesthe
us-citizenship
us-immigration
wall-along
year-ago

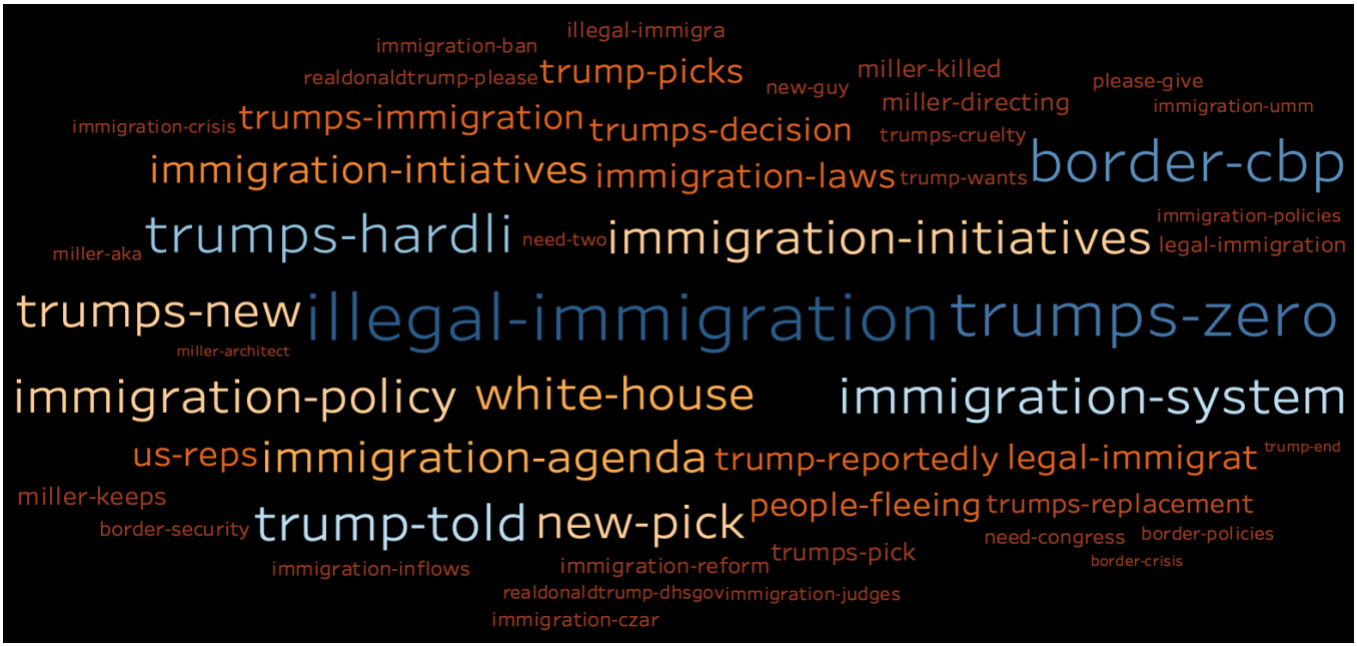

Word
Count
0
20
40
60
80
100
120
140
160
administr-officials
advertisementsupport-by..
asylum-officer
asylum-seekers
border-patrol
border-protection
border-security
border-wall
countri-illegally
de-la
de-las
de-los
el-gobierno
el-paso
el-salvador
famili-members
famili-separation
govern-shutdown
immigr-officials
immigr-policies
immigr-policy
immigr-services
immigr-system
la-frontera
last-month
last-week
ms-batiz
ms-nielsen
ms-ny
ms-safieh
ms-torres
nation-emergency
nation-security
new-york
new-zealand
peopl-familiar
presid-donald
presid-trump
presid-trumps
secur-secretary
state-department
time-opinion
two-years
unit-nations
unit-states
unit-statesthe
us-citizenship
us-immigration
wall-along
year-ago

Round Plot

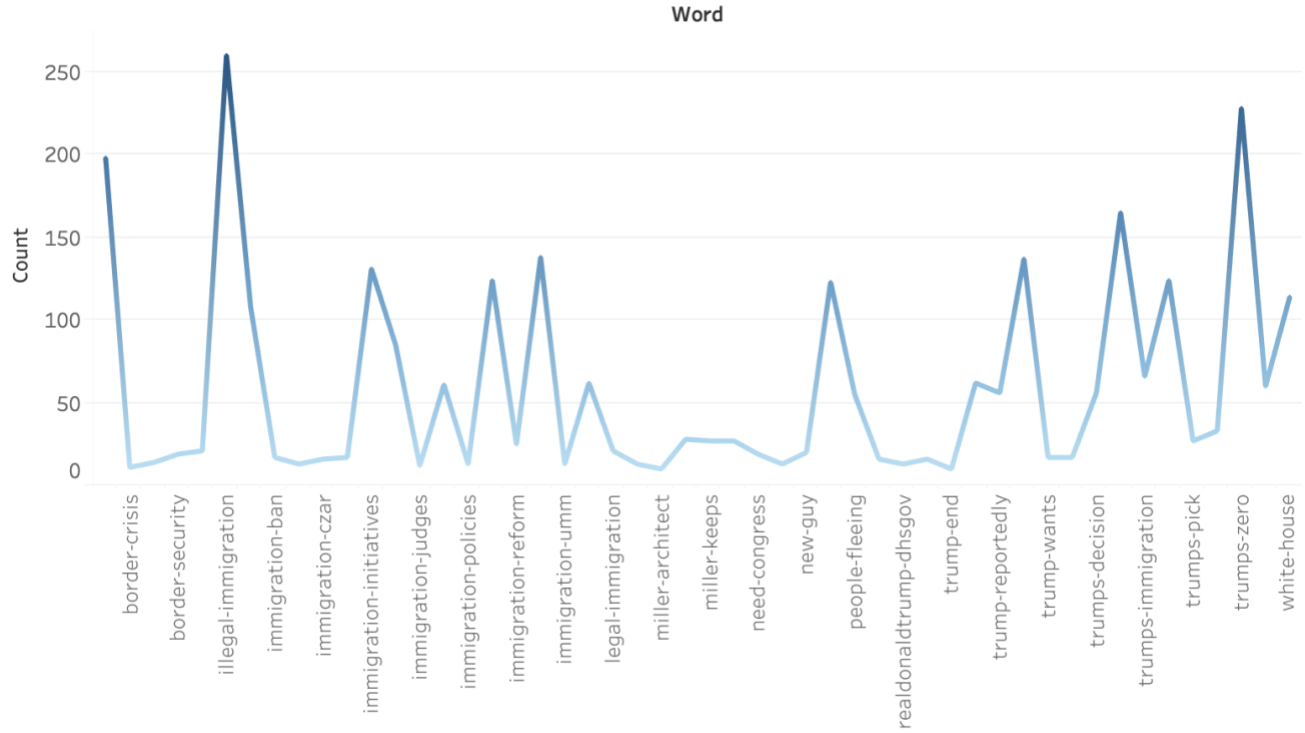


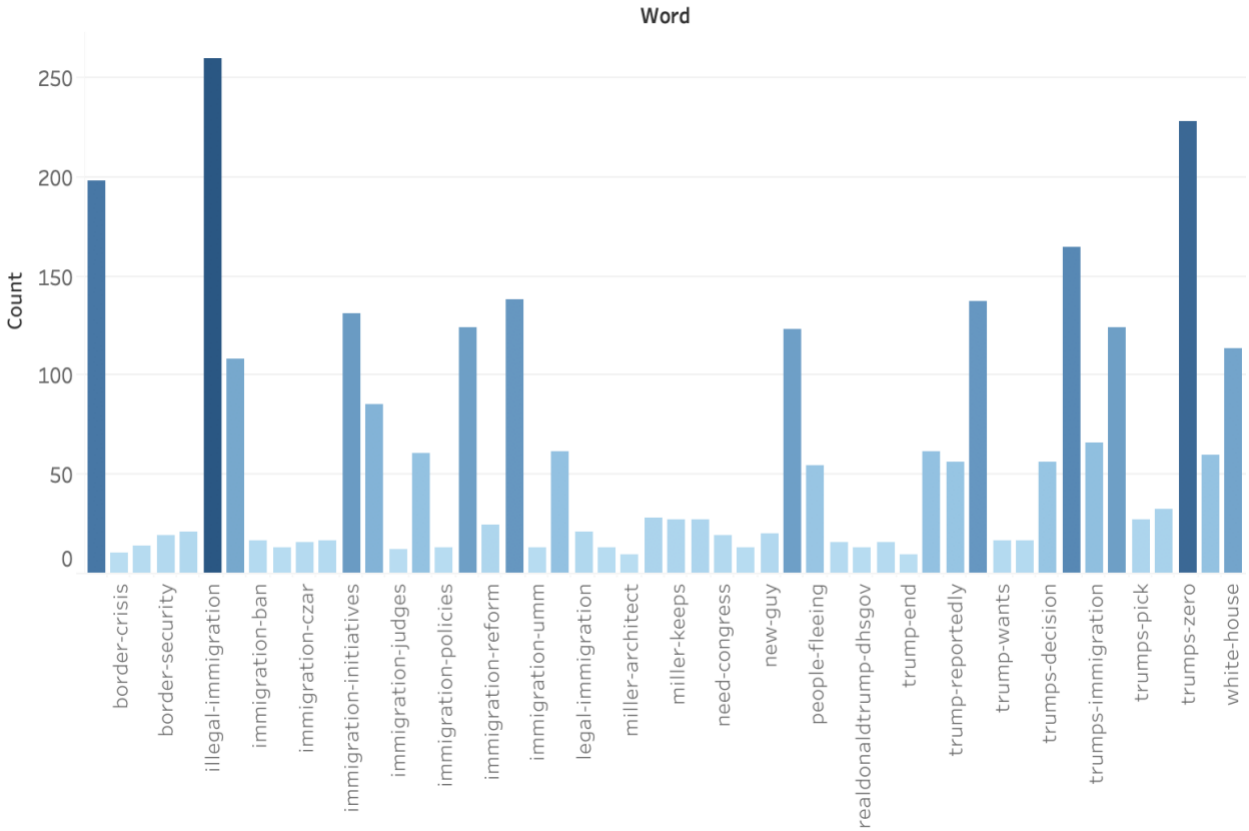
Common Crawl Top 50 Word Co-occurrence

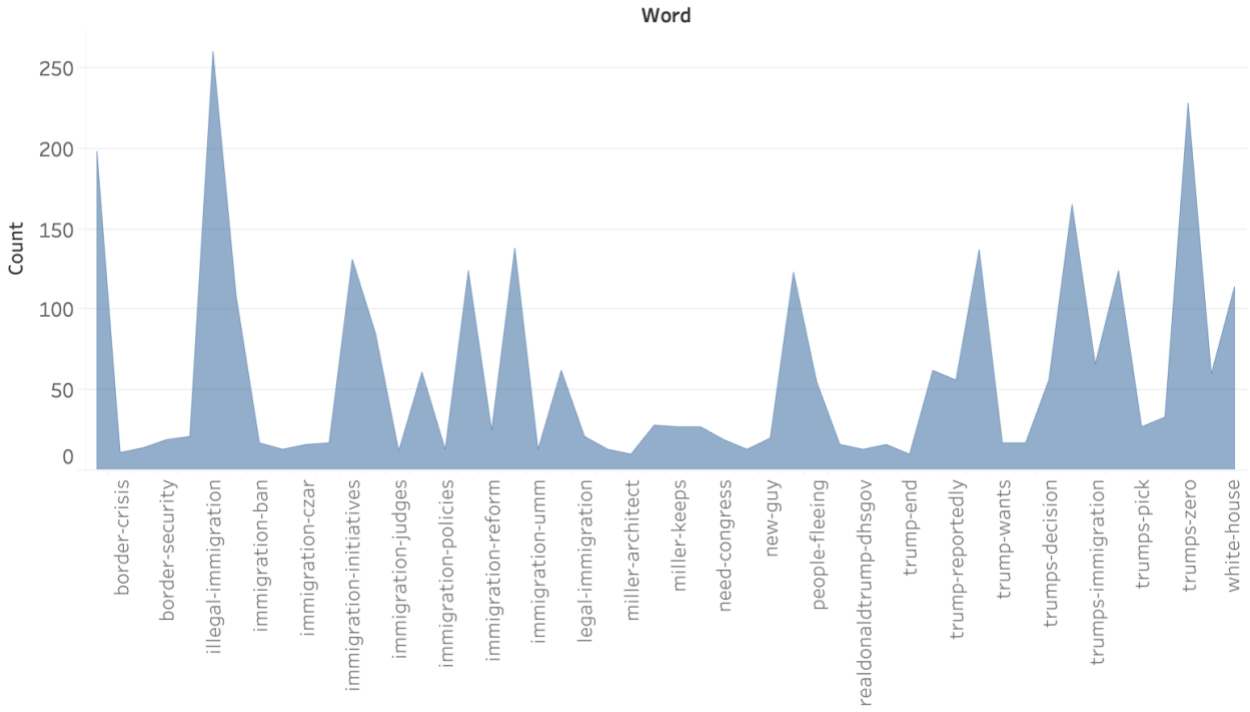
Word Cloud



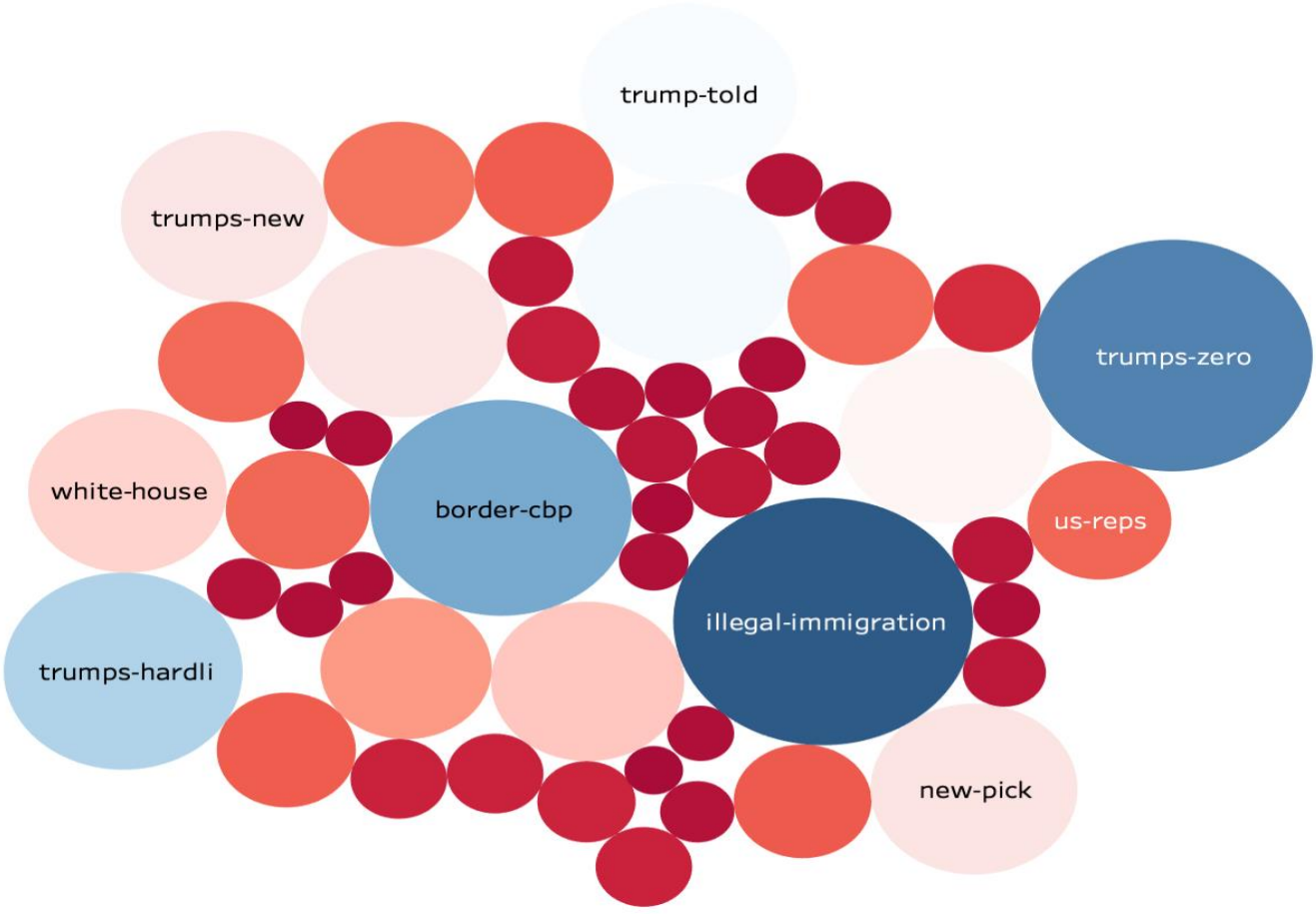
Line, Bar and Area Plot






Word
Count
250
200
150
100
50
0
border-crisis
border-security
illegal-immigration
immigration-ban
immigration-czar
immigration-initiatives
immigration-judges
immigration-policies
immigration-reform
immigration-umm
legal-immigration
miller-architect
miller-keeps
need-congress
new-guy
people-fleeing
realdonaldtrump-dhsgov
trump-end
trump-reportedly
trump-wants
trumps-decision
trumps-immigration
trumps-pick
trumps-zero
white-house

Round Plot


trump-told
trumps-new
trumps-zero
white-house
border-cbp
us-reps
illegal-immigration
trumps-hardli
new-pick

→ *Directory Structure*

goswami5Lab2.zip

- **Report.pdf**
- **Video.mp4**
- **README.txt**
- **part1** (folder)
  - **Code** (folder) -> all python and R codes for fetching Twitter, NY Times and Common Crawl Data.
  - **Data** (folder)
    - Twitter (folder) -> all tweets are inside *"tweets.txt"* file
    - NYT (folder) -> all NYT articles are inside *"nytimes.txt"*. Subtopic files are also submitted.
    - Commoncrawl (folder) -> all Common Crawl articles are inside *"common_crawl.txt"*
- **part2** (folder)
  - pg345.txt, mapper.py and reducer.py as provided by the professor.
  - Output files (_SUCCESS and part-00000) and screenshot (Demo_Output) obtained after running the demo.
- **part3** (folder)
  - **Twitter** (folder)
    - Code (folder) -> Contains all the mapper and reducer scripts for word count and word co-occurrence.
    - Images (folder) -> Contains all the *.twbx, image files* for word count and word co-occurrence to create Twitter cloud.
    - Output (folder) - > Contains folders and output files for word count and word co-occurrence for Twitter data.
  - **NYT** (folder)
    - Code (folder) -> Contains all the mapper and reducer scripts for word count and word co-occurrence.
    - Images (folder) -> Contains all the *.twbx, image files* for word count and word co-occurrence to create NYT cloud.
    - Output (folder) - > Contains folders and output files for word count and word co-occurrence for New York Times data.
  - **Commoncrawl** (folder)
    - Code (folder) -> Contains all the mapper and reducer scripts for word count and word co-occurrence.
    - Images (folder) -> Contains all the *.twbx, image files* for word count and word co-occurrence to create Common Crawl cloud.

  - Output (folder) - > Contains folders and output files for word count and word co-occurrence for Common Crawl data.
- **Webpage** (folder)
  - Interactive webpage file showing the results based on data collected from various sources. https://public.tableau.com/profile/abhishek.goswami#!/vizhome/Wordcountandcooccurencevisualization/DICLAB2Visualization

CONCLUSION:

The analysis and visualization of word count and word co-occurrence for Twitter, New York Times and Common Crawl is shown in the above plots. Hadoop infrastructure and AWS console was used for running the word-count and word co-occurrence using the appropriate mapper and reducer scripts. The files obtained from the reducer output were then used further in Tableau for visualization.