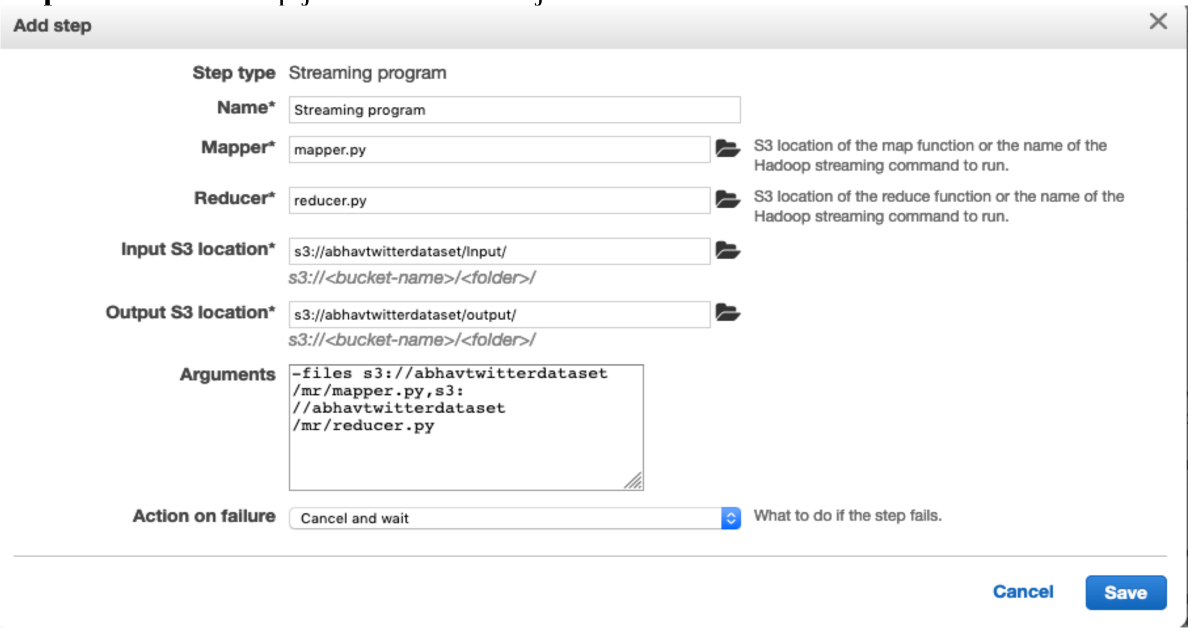
Lab 2 Data Intensive Computing

Name - Abhav Luthra
Person No – 50288904

Methodology for Aws Setup

1. Create Bucket on Aws and Input files
2. Schedule EMR job with steps as bellow

Step 1 will Run 8 Map jobs and 4 reduce jobs



Add step

Step type: Streaming program

Name*: Streaming program

Mapper*: mapper.py — S3 location of the map function or the name of the Hadoop streaming command to run.

Reducer*: reducer.py — S3 location of the reduce function or the name of the Hadoop streaming command to run.

Input S3 location*: s3://abhavtwitterdataset/Input/
s3://<bucket-name>/<folder>/

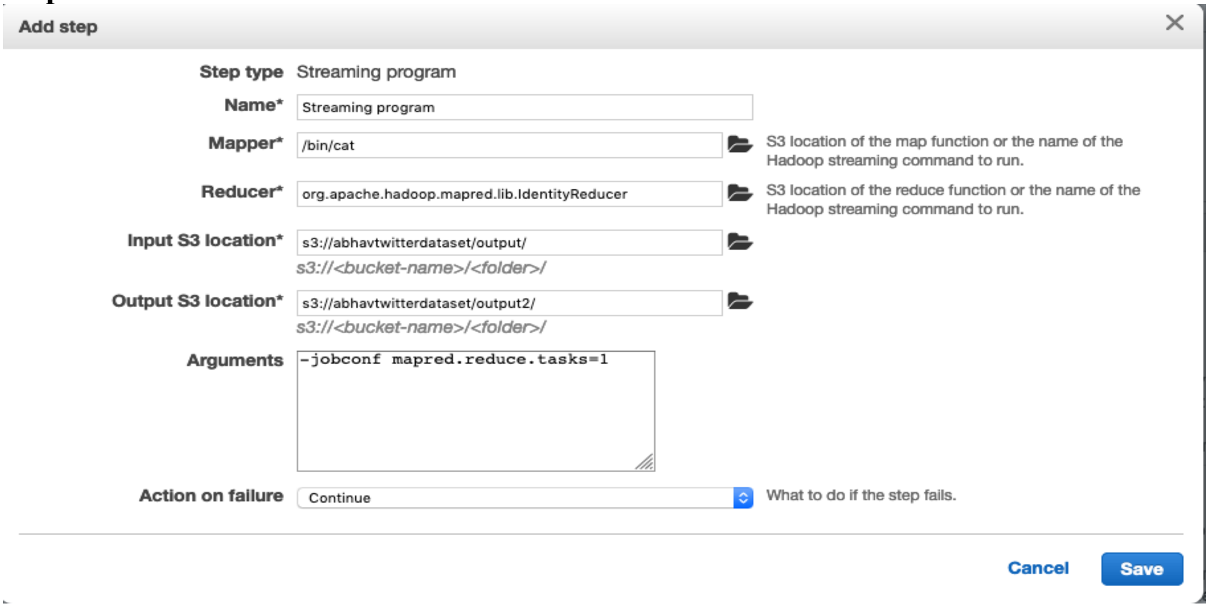
Output S3 location*: s3://abhavtwitterdataset/output/
s3://<bucket-name>/<folder>/

Arguments:

```
-files s3://abhavtwitterdataset
/mr/mapper.py,s3:
//abhavtwitterdataset
/mr/reducer.py
```

Action on failure: Cancel and wait — What to do if the step fails.

Step 2 will combine those 4 reduced files to 1



Add step

Step type: Streaming program

Name*: Streaming program

Mapper*: /bin/cat — S3 location of the map function or the name of the Hadoop streaming command to run.

Reducer*: org.apache.hadoop.mapred.lib.IdentityReducer — S3 location of the reduce function or the name of the Hadoop streaming command to run.

Input S3 location*: s3://abhavtwitterdataset/output/
s3://<bucket-name>/<folder>/

Output S3 location*: s3://abhavtwitterdataset/output2/
s3://<bucket-name>/<folder>/

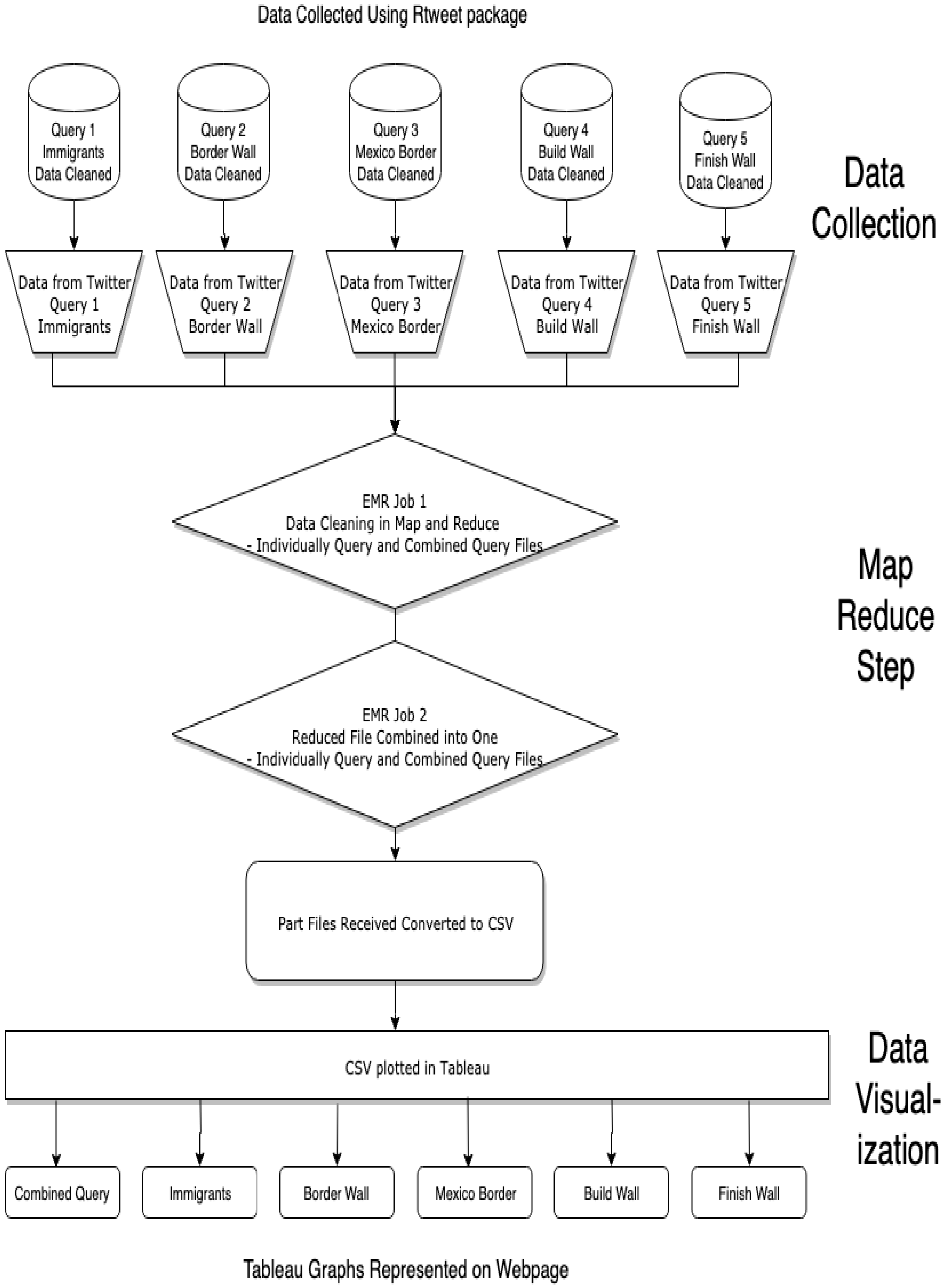
Arguments:

```
-jobconf mapred.reduce.tasks=1
```

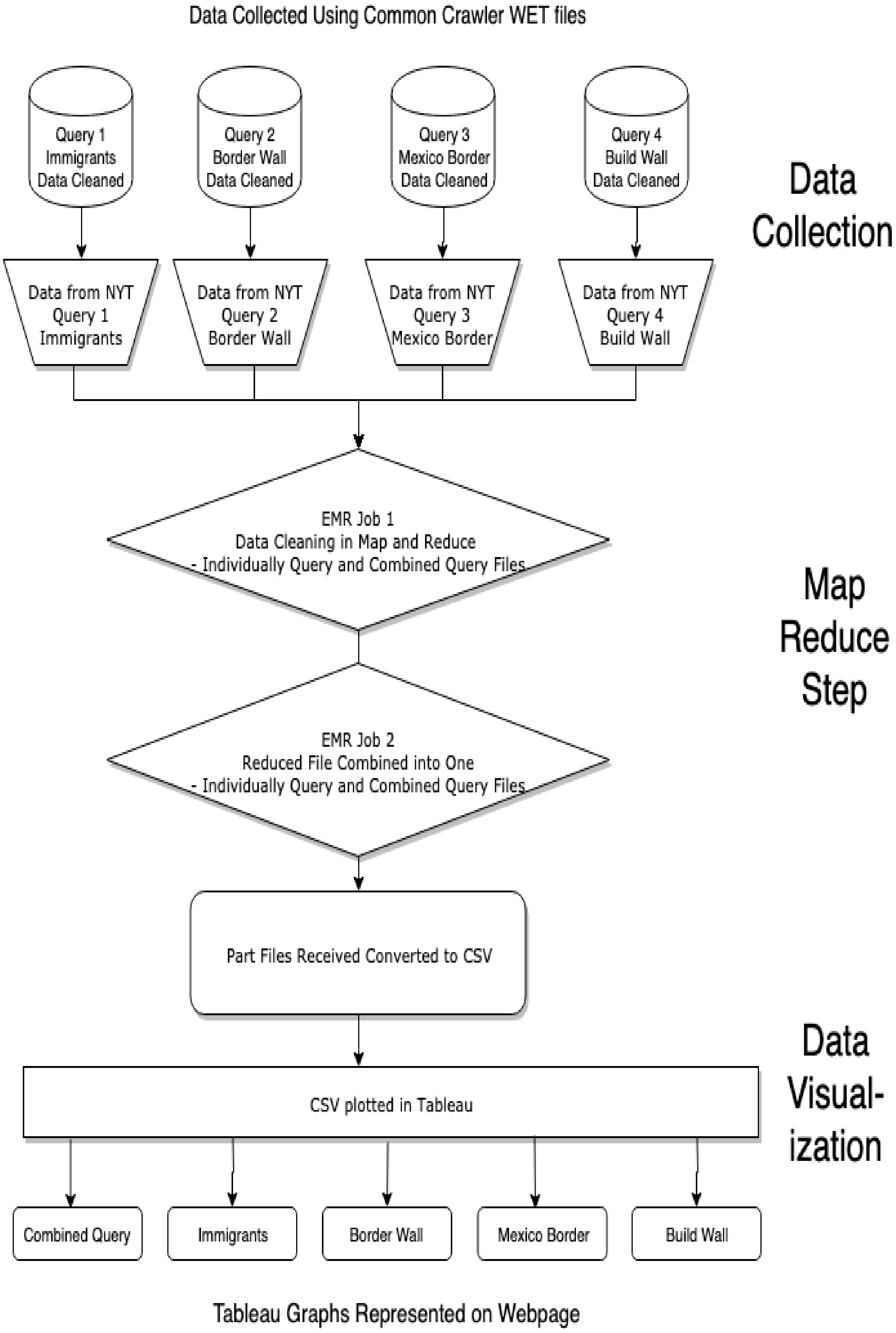
Action on failure: Continue — What to do if the step fails.

3. Part-00000 file received

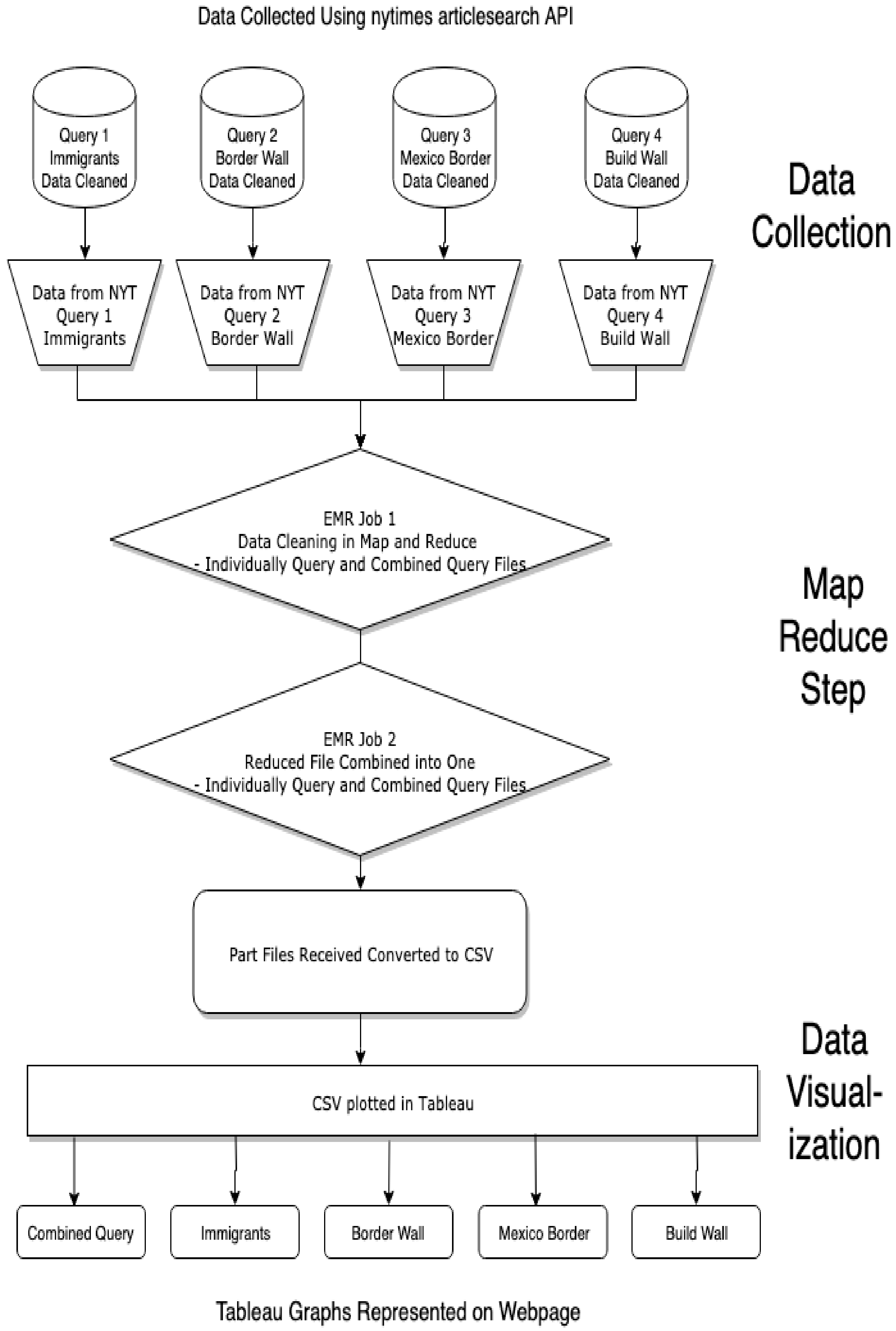
Job Flow for Twitter Data


Data Collected Using Rtweet package
Query 1 Immigrants Data Cleaned
Query 2 Border Wall Data Cleaned
Query 3 Mexico Border Data Cleaned
Query 4 Build Wall Data Cleaned
Query 5 Finish Wall Data Cleaned
Data Collection
Data from Twitter Query 1 Immigrants
Data from Twitter Query 2 Border Wall
Data from Twitter Query 3 Mexico Border
Data from Twitter Query 4 Build Wall
Data from Twitter Query 5 Finish Wall
EMR Job 1 Data Cleaning in Map and Reduce - Individually Query and Combined Query Files
Map Reduce Step
EMR Job 2 Reduced File Combined into One - Individually Query and Combined Query Files
Part Files Received Converted to CSV
CSV plotted in Tableau
Data Visual-ization
Combined Query
Immigrants
Border Wall
Mexico Border
Build Wall
Finish Wall
Tableau Graphs Represented on Webpage

Job Flow for Common Crawl Data


Data Collected Using Common Crawler WET files
Query 1 Immigrants Data Cleaned
Query 2 Border Wall Data Cleaned
Query 3 Mexico Border Data Cleaned
Query 4 Build Wall Data Cleaned
Data Collection
Data from NYT Query 1 Immigrants
Data from NYT Query 2 Border Wall
Data from NYT Query 3 Mexico Border
Data from NYT Query 4 Build Wall
EMR Job 1
Data Cleaning in Map and Reduce
- Individually Query and Combined Query Files
Map Reduce Step
EMR Job 2
Reduced File Combined into One
- Individually Query and Combined Query Files
Part Files Received Converted to CSV
Data Visual-ization
CSV plotted in Tableau
Combined Query
Immigrants
Border Wall
Mexico Border
Build Wall
Tableau Graphs Represented on Webpage

Job Flow for NYT Data


Data Collected Using nytimes articlesearch API
Query 1 Immigrants Data Cleaned
Query 2 Border Wall Data Cleaned
Query 3 Mexico Border Data Cleaned
Query 4 Build Wall Data Cleaned
Data Collection
Data from NYT Query 1 Immigrants
Data from NYT Query 2 Border Wall
Data from NYT Query 3 Mexico Border
Data from NYT Query 4 Build Wall
EMR Job 1 Data Cleaning in Map and Reduce - Individually Query and Combined Query Files
Map Reduce Step
EMR Job 2 Reduced File Combined into One - Individually Query and Combined Query Files
Part Files Received Converted to CSV
Data Visual-ization
CSV plotted in Tableau
Combined Query
Immigrants
Border Wall
Mexico Border
Build Wall
Tableau Graphs Represented on Webpage

Query list used – **Border Wall, Build Wall, Immigrant, Finish Wall (only Twitter as no data found), Mexico Border**

Data Collection

1. Twitter

Data collected on 29 March and 9 April 2019. Count of tweets-
Border Wall - 27449
Build Wall – 32198 + 26556
Finish Wall – 664 + 3583
Mexico Border - 28937
Immigrant - 33345
Total - 152732

2. NYT

Data collected on 29 March 2019 using ArticleSearch Api and Beautiful Soup. Article Count-
Border Wall - 577
Build Wall - 170
Mexico Border - 207
Immigrant – 943
Total – around 1800

3. Common Crawl

Data collected on 15 April 2019. Article Count-
Border Wall - 2514
Build Wall - 220
Mexico Border – 84
Immigrants – 728314

Data Processing

1. Twitter

Removed Retweets. Count of tweets-
Border Wall – 5632
Build Wall – 5452 + 7480 = 12932
Finish Wall – 325 + 1360 = 1685
Mexico Border - 1721
Immigrant – 7246
Total = 29216

2. NYT

Sentence Tokenized for Co-occurrence using nltk

```
nltk.sent_tokenize(article)
```

3. Common Crawl

Sentence Tokenized for Co-occurrence using nltk

```
nltk.sent_tokenize(article)
```

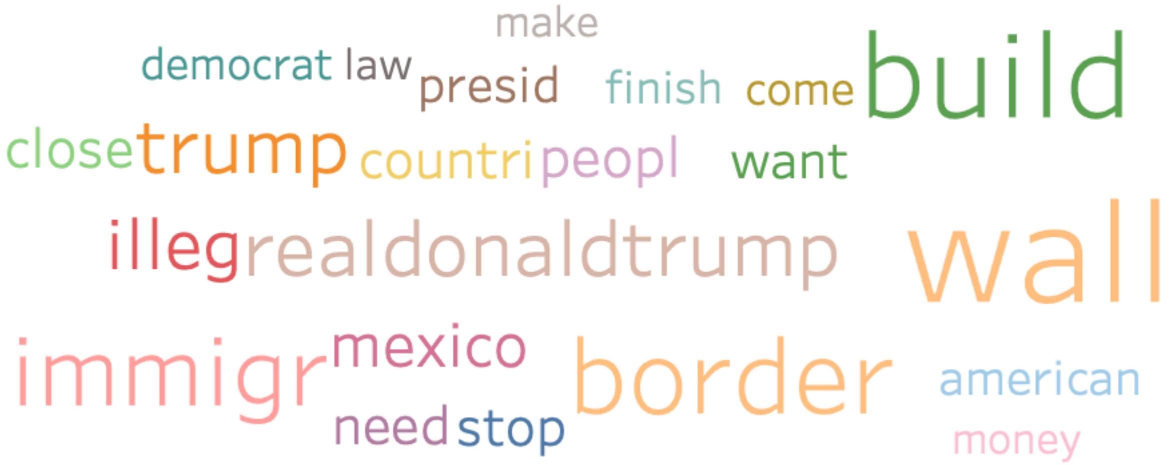
Mapper and Reducer
Mapper.py – Word Count
Mapper2.py – Word Co- occurrence – NYT and Common Crawl
Mapper3.py – Word Co- occurrence – Twitter
MapperCC.py py – Word Count – Common Crawl

Minor changes done to mapper to improve pre-processing.

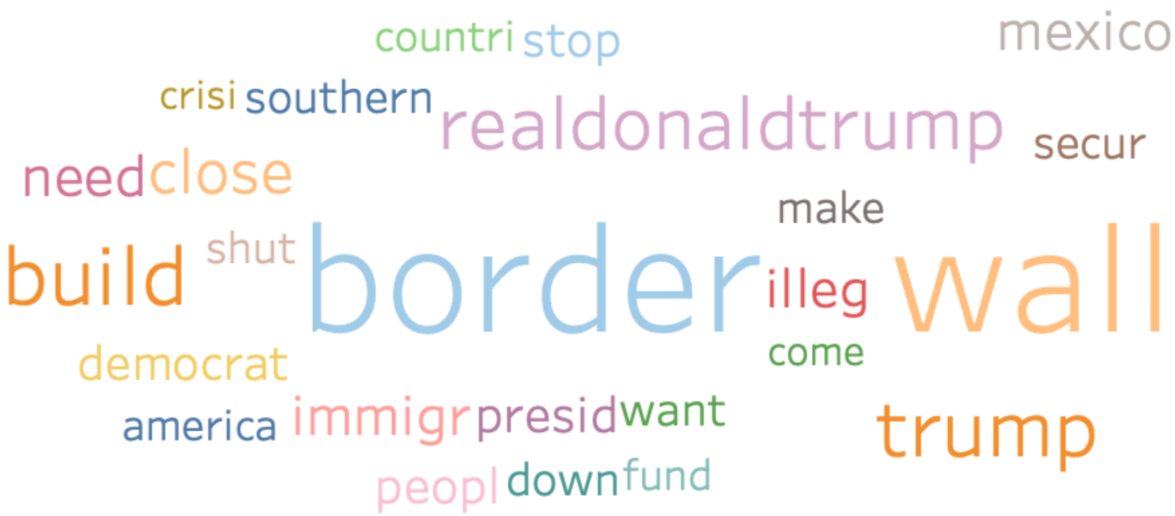
Data Visualization
In Common Crawl, only Border Wall and Immigrant data is used for co-occurrence and full word count as rest return irrelevant words

Twitter

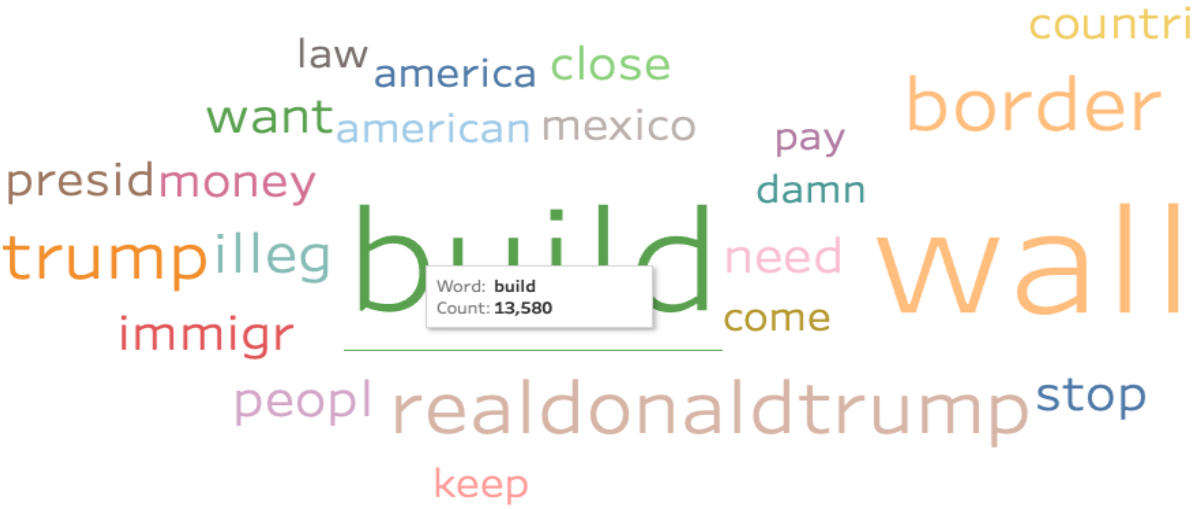
Word Count
Word Cloud on Full Data



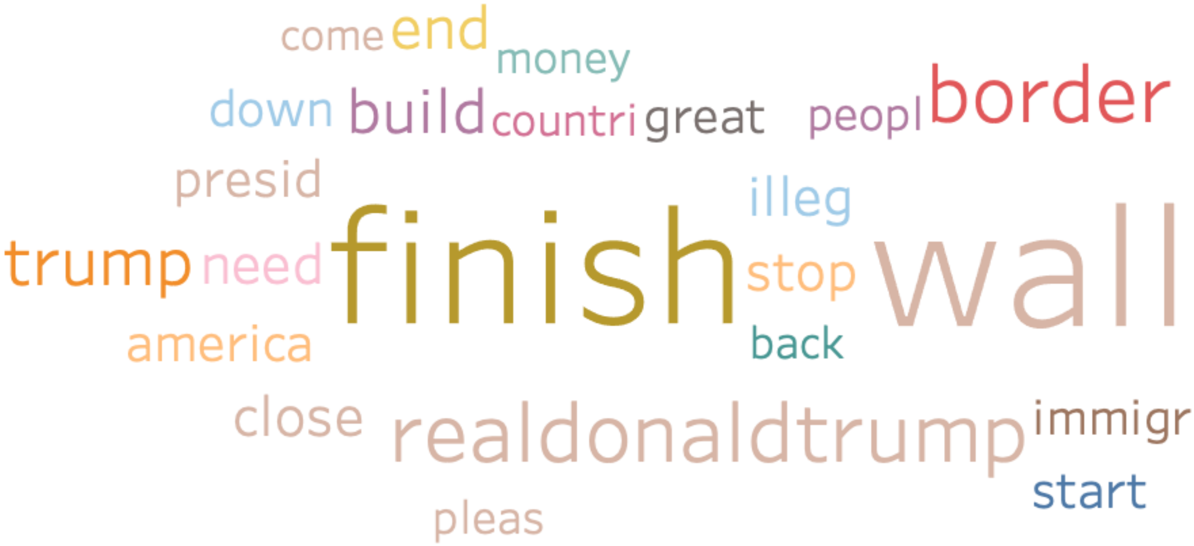
Word Cloud on Border Wall



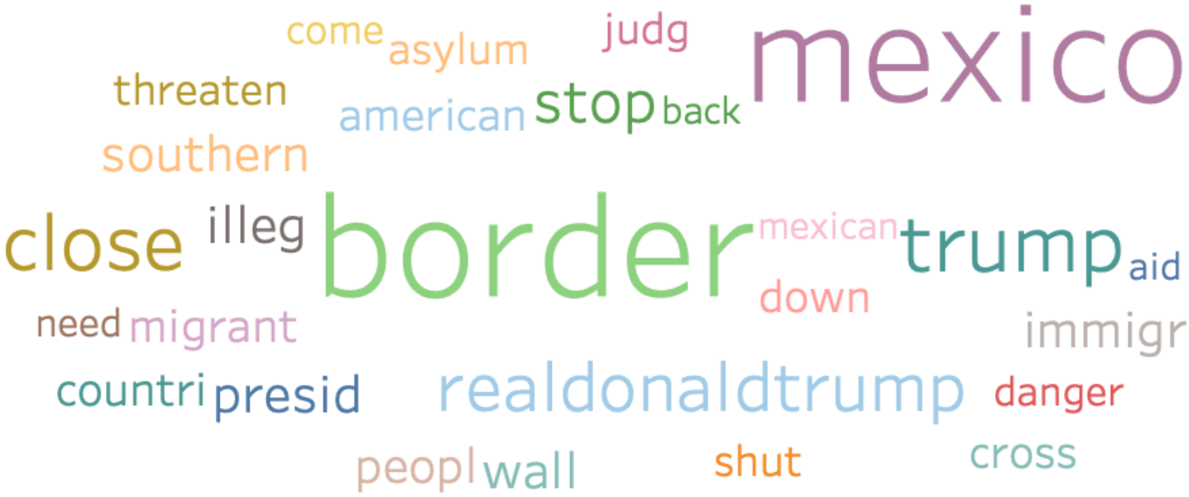
Word Cloud on Build Wall



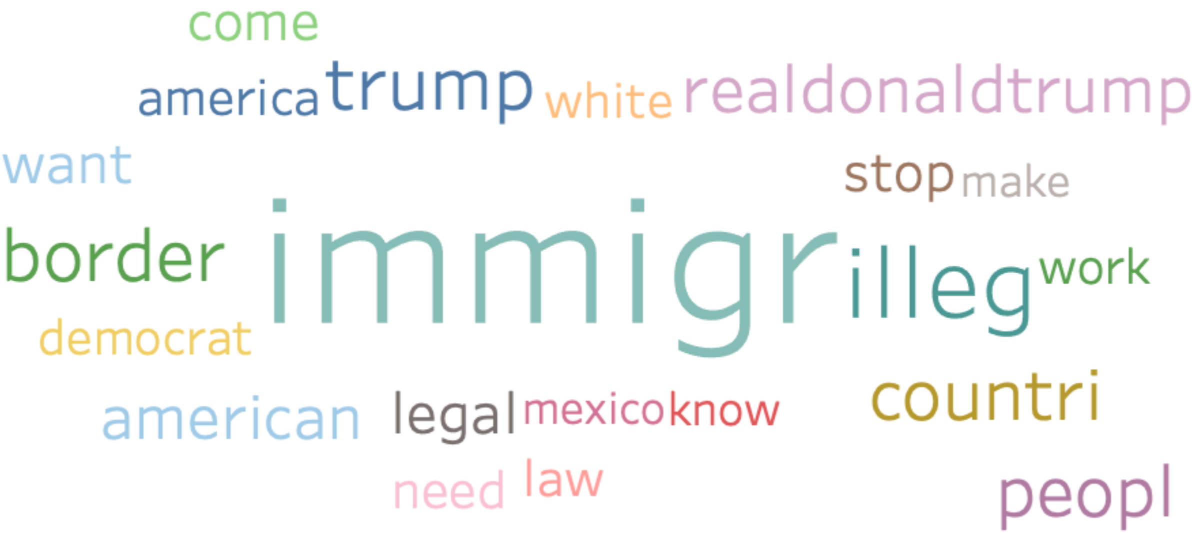
Word Cloud on Finish Wall



Word Cloud on Mexico Border

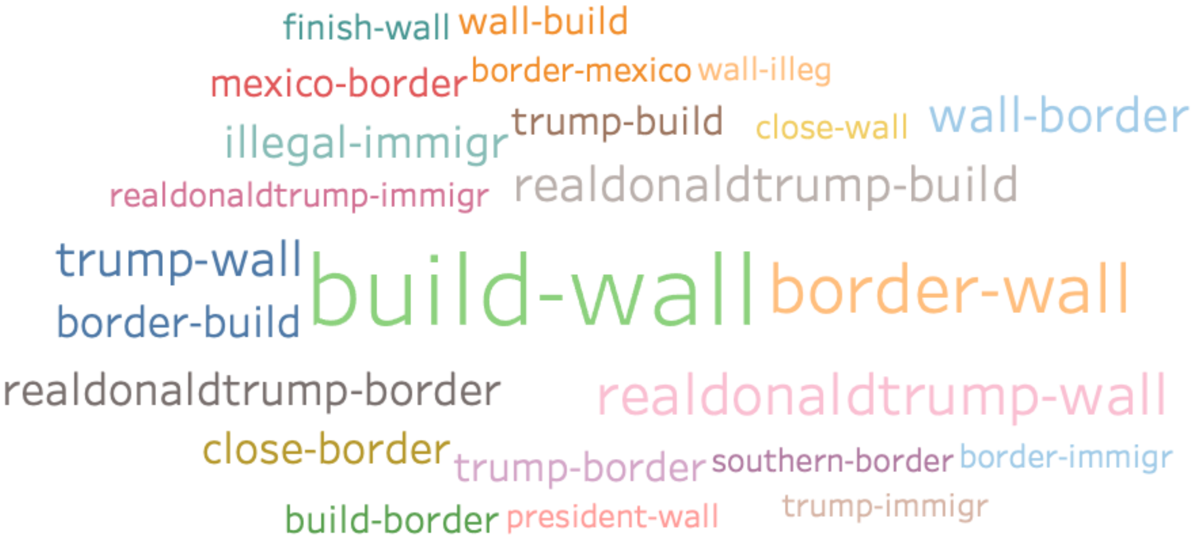


Word Cloud on Immigrant

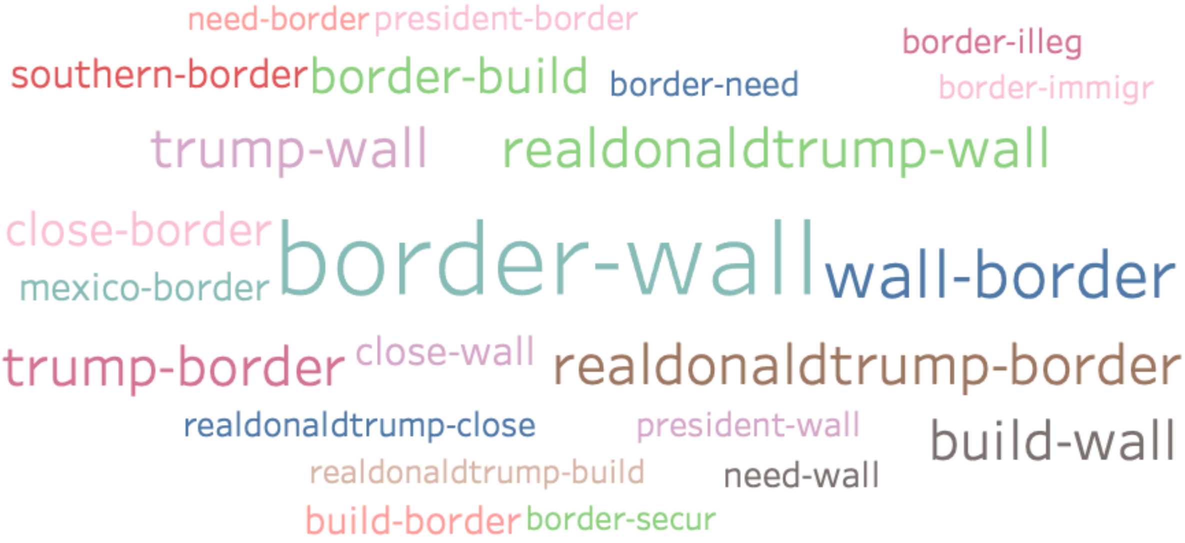


Word Co-occurrence

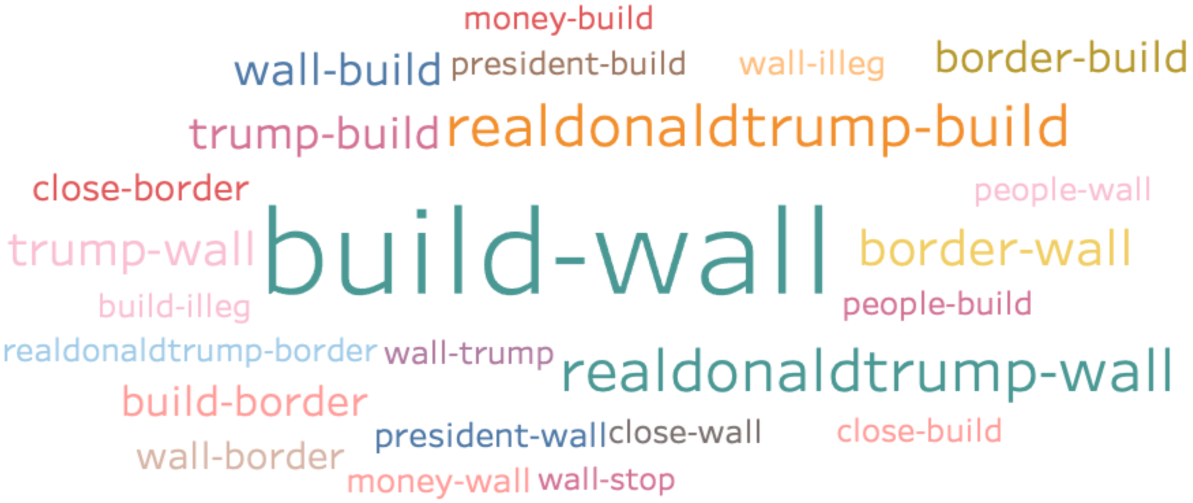
Word Cloud on Full Data



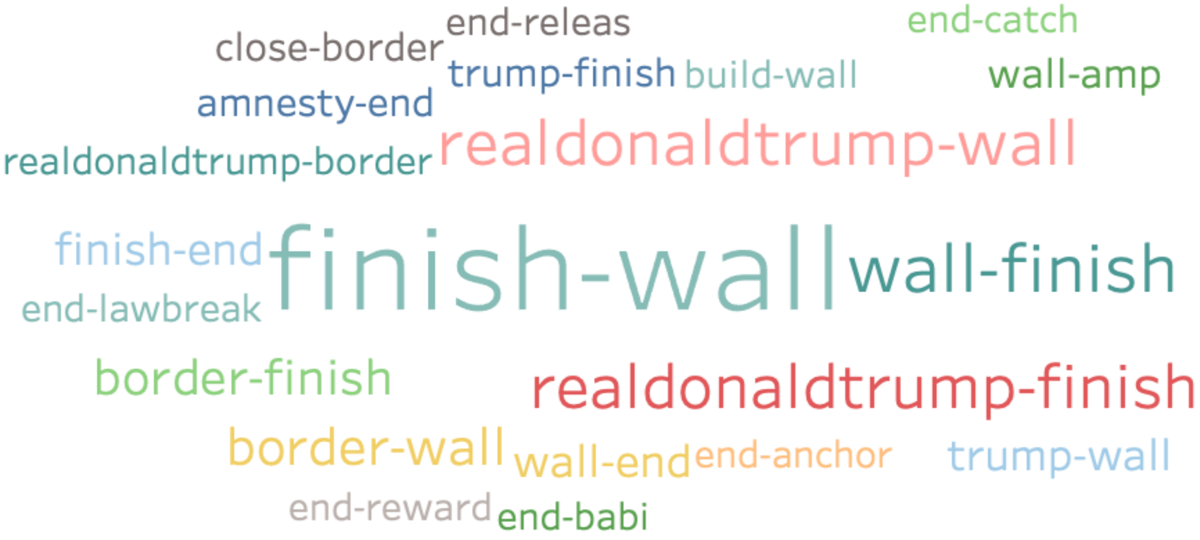
Word Cloud on Border Wall



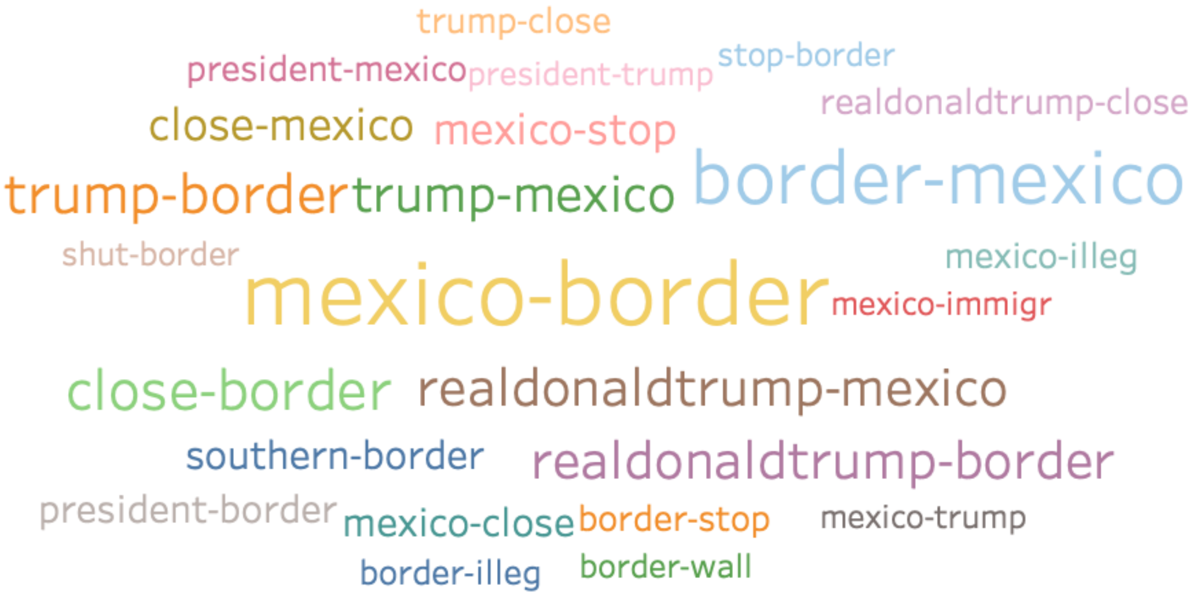
Word Cloud on Build Wall



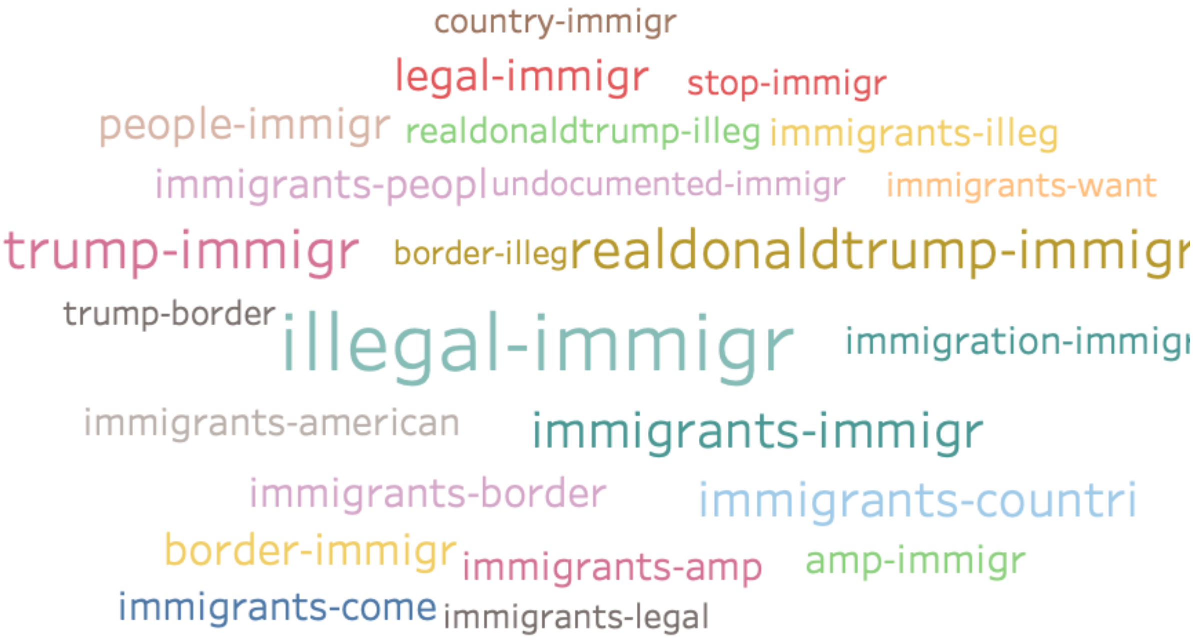
Word Cloud on Finish Wall



Word Cloud on Mexico Border



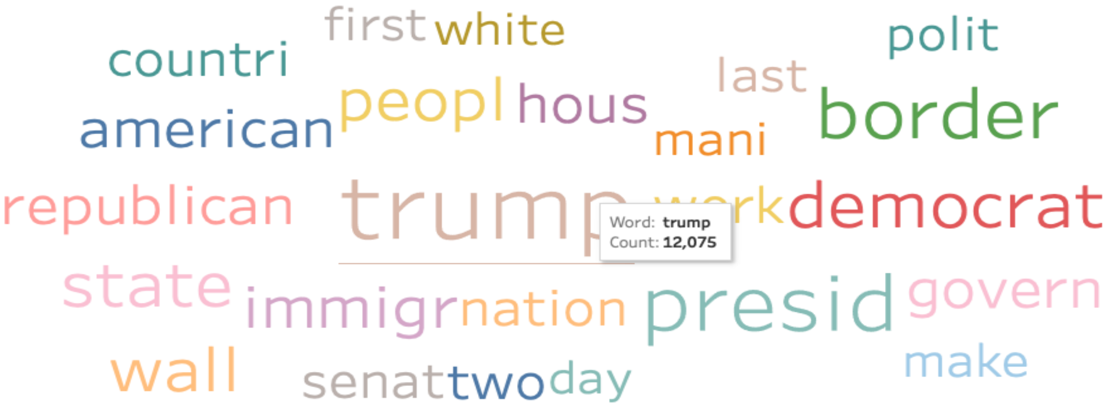
Word Cloud on Immigrant



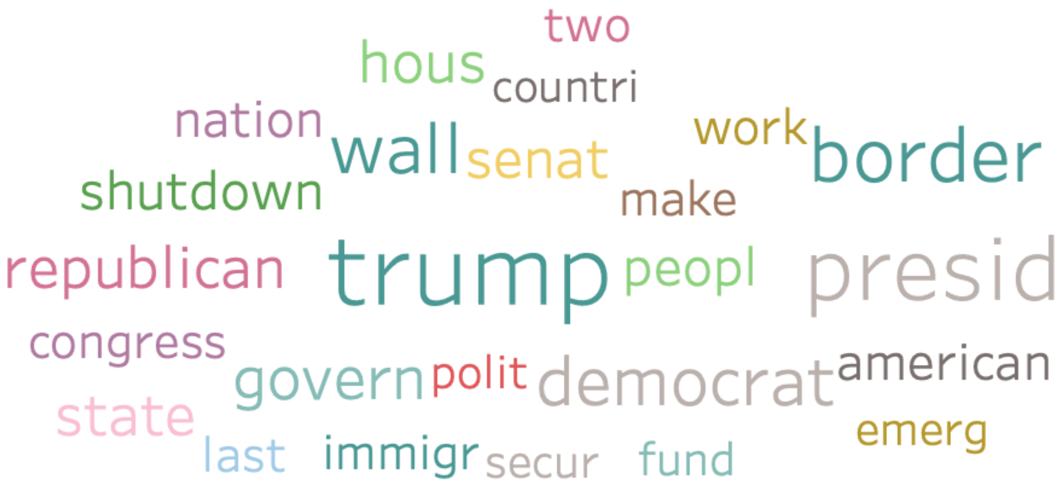
New York Times

Word Count

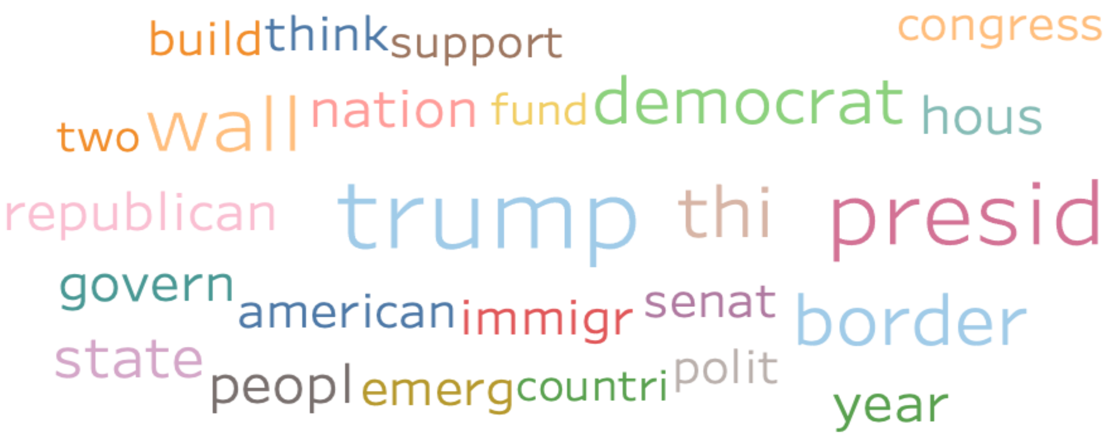
Word Cloud on Full Data



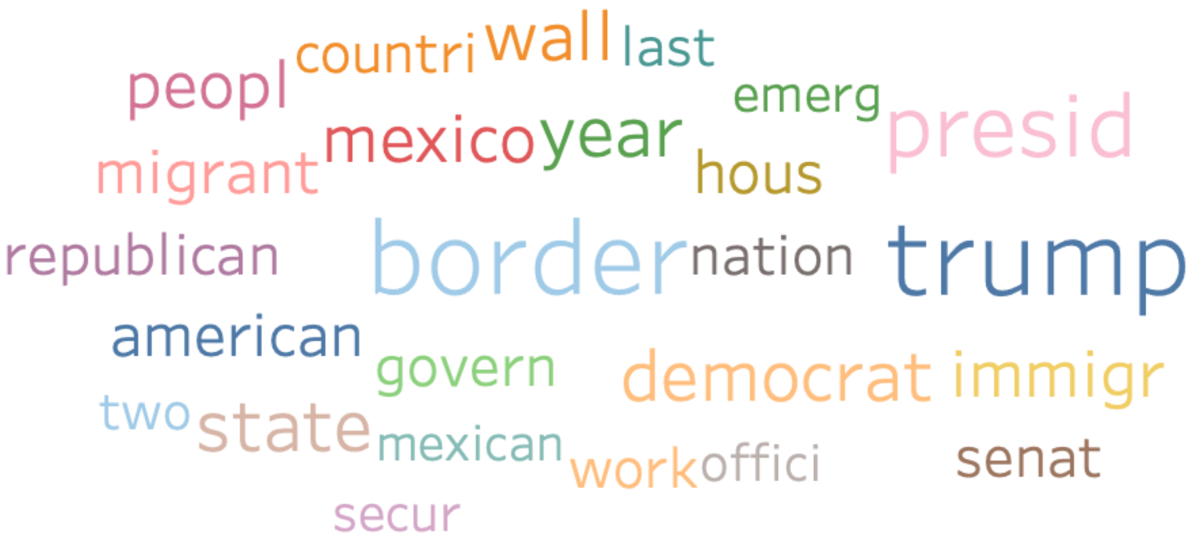
Word Cloud on Border Wall



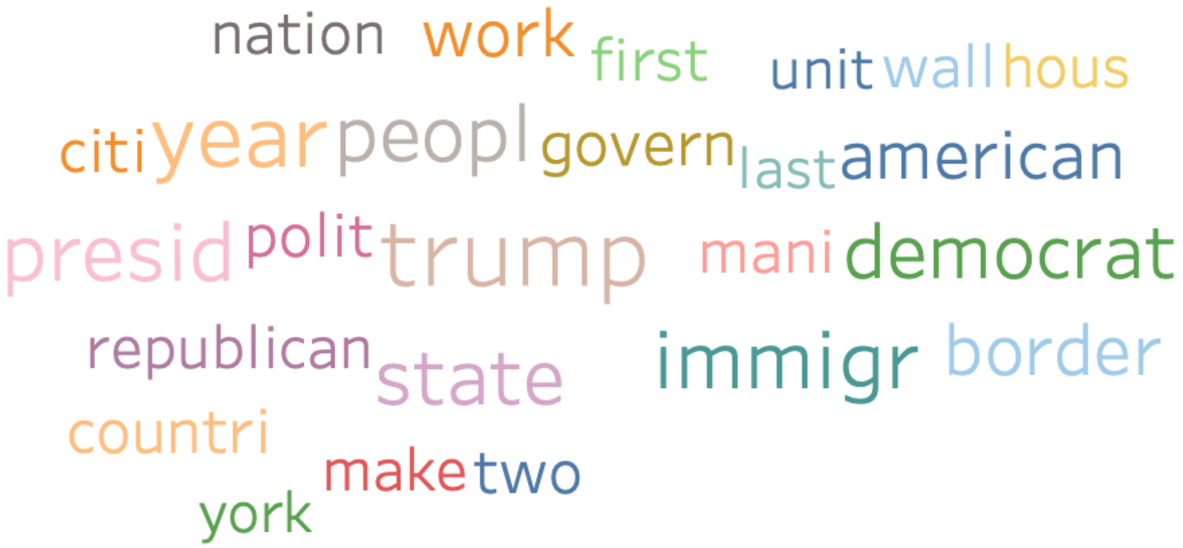
Word Cloud on Build Wall



Word Cloud on Mexico Border

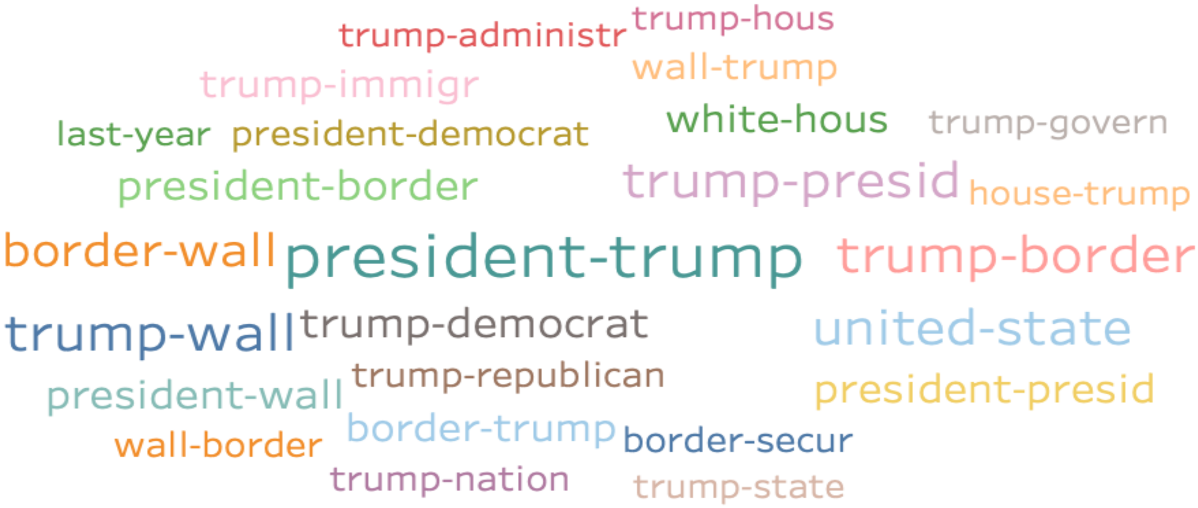


Word Cloud on Immigrant

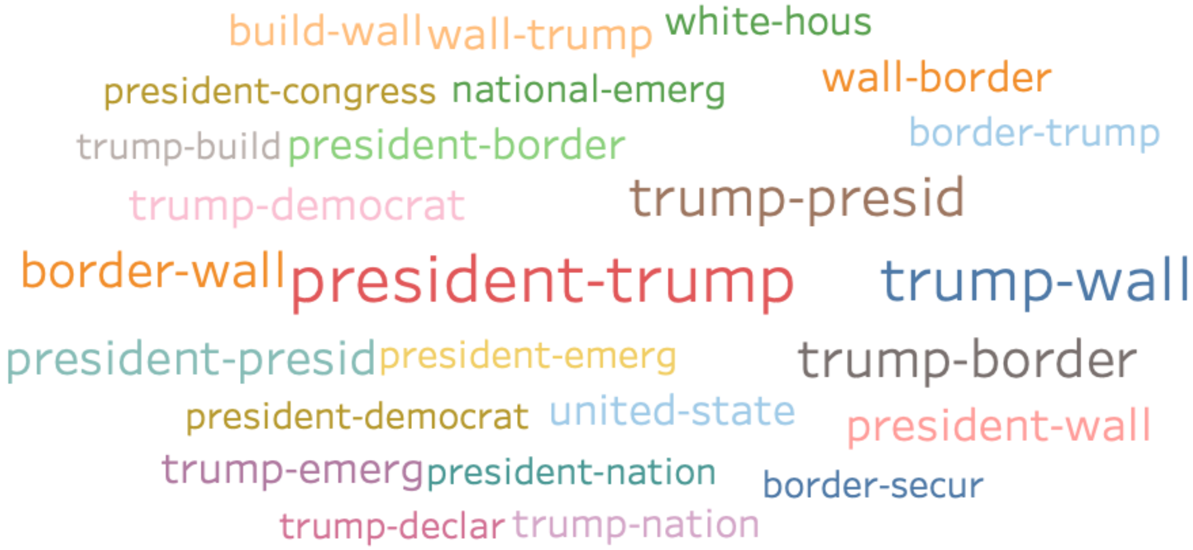


Word Co-occurrence

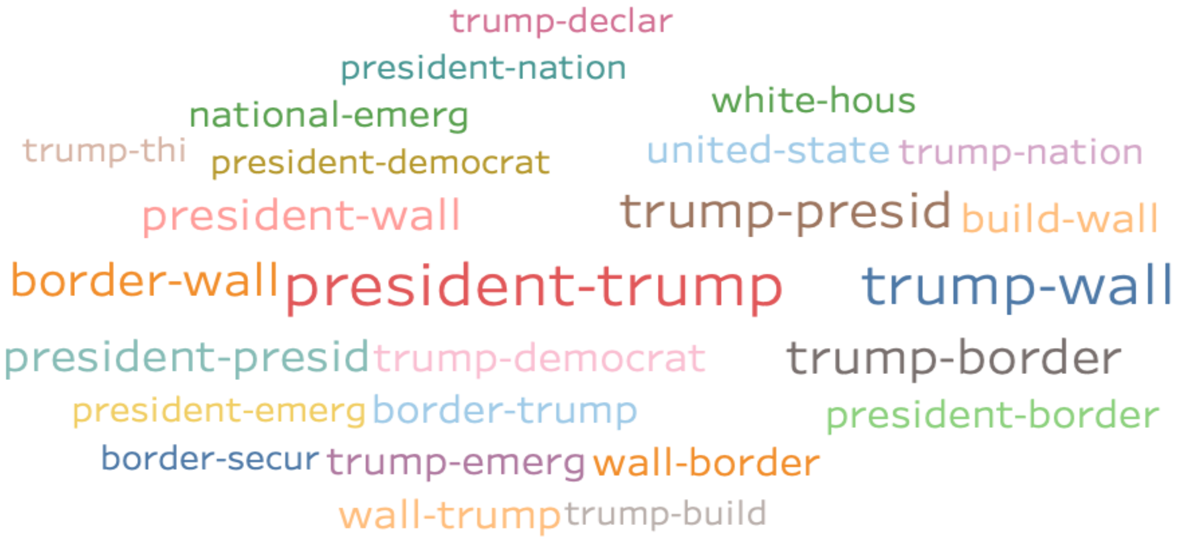
Word Cloud on Full Data



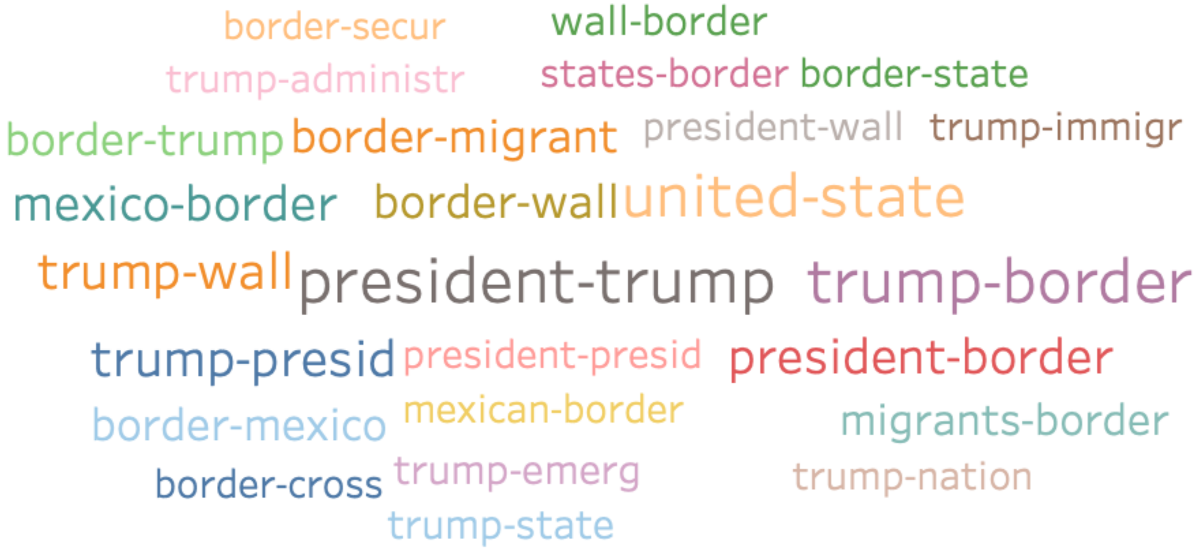
Word Cloud on Border Wall



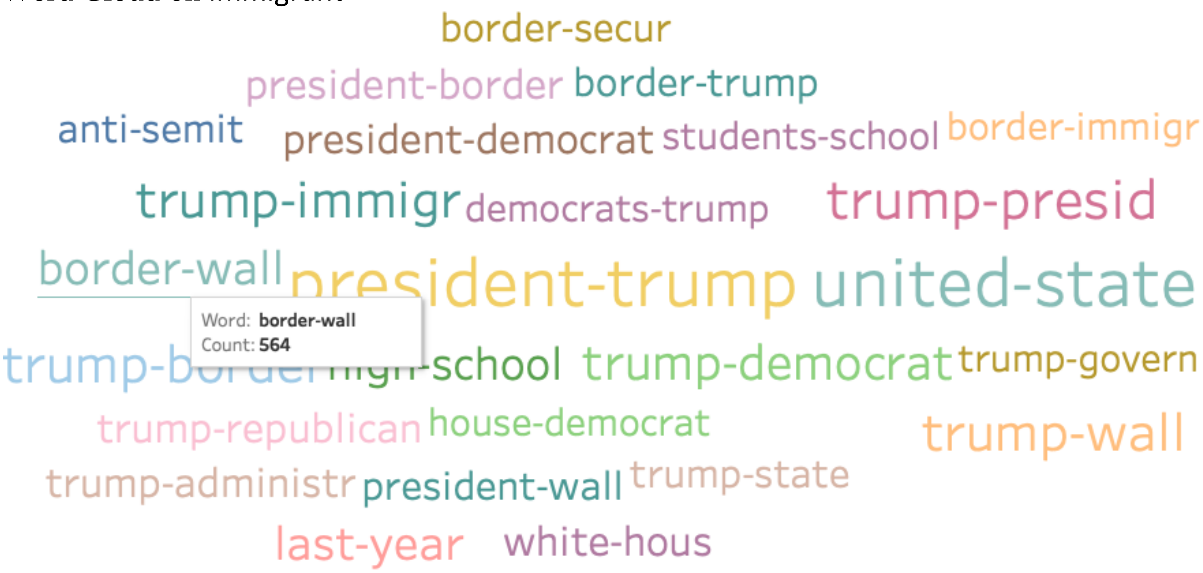
Word Cloud on Build Wall



Word Cloud on Mexico Border



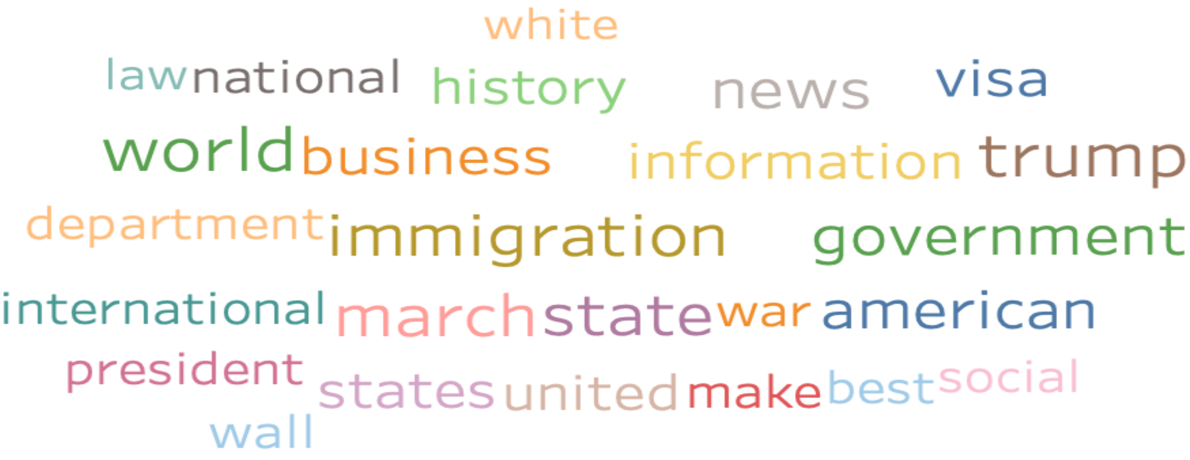
Word Cloud on Immigrant



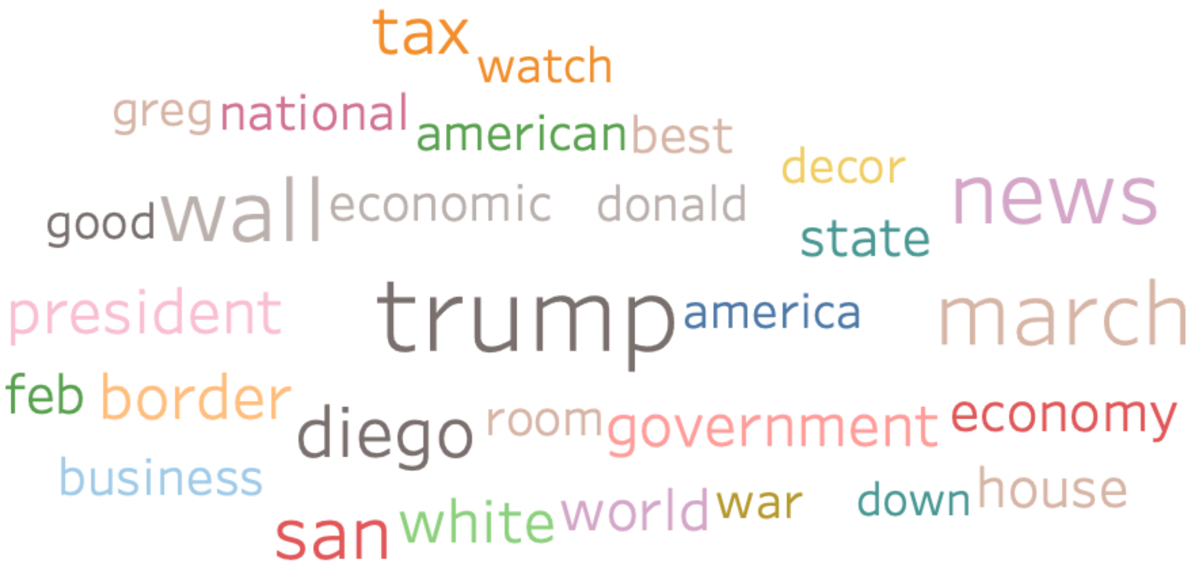
Common Crawl

Word Count

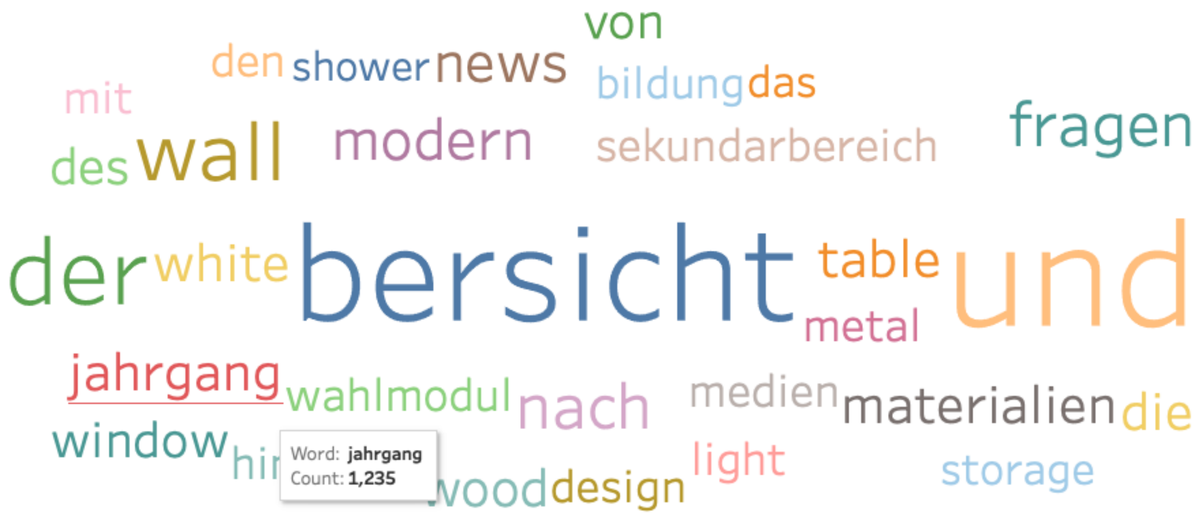
Word Cloud on Full Data



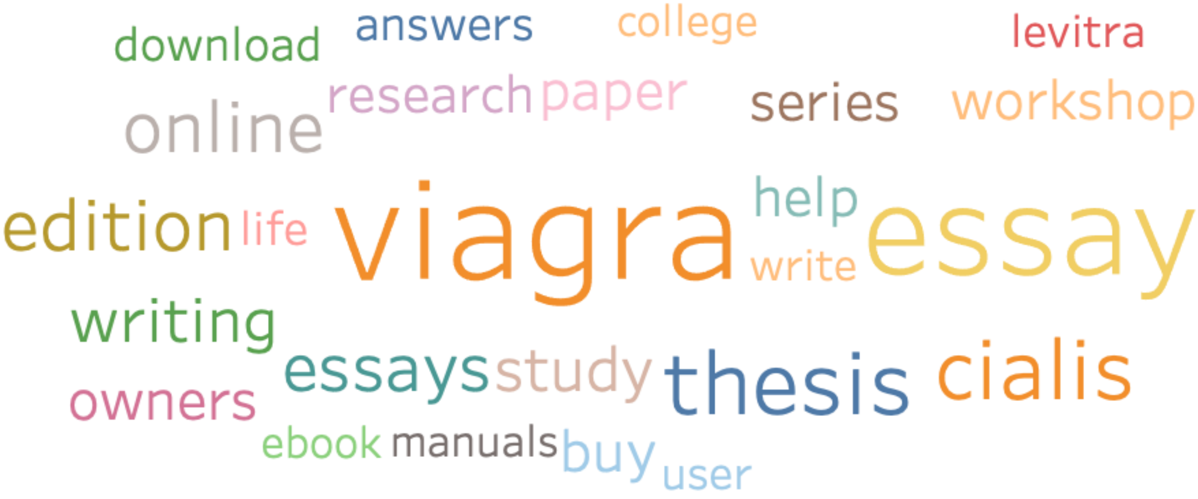
Word Cloud on Border Wall



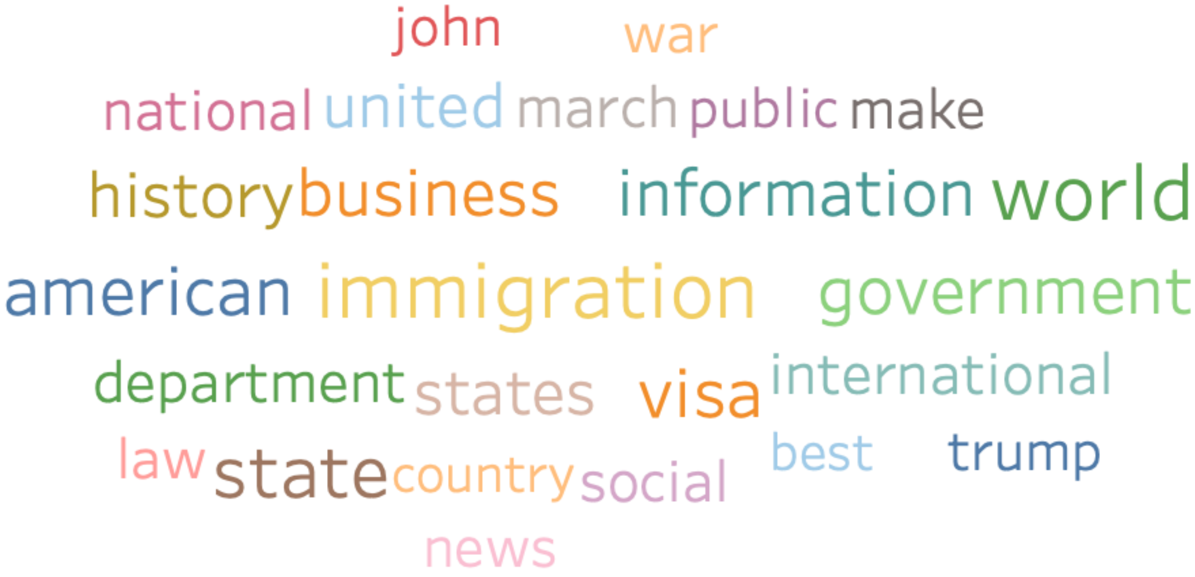
Word Cloud on Build Wall



Word Cloud on Mexico Border



Word Cloud on Immigrant



Word Co-occurrence
Unable to run even with 20 instances on AWS.