

# 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\***

**Mapper\***   S3 location of the map function or the name of the Hadoop streaming command to run.


**Reducer\***   S3 location of the reduce function or the name of the Hadoop streaming command to run.

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

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

**Arguments**

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

**Action on failure**   What to do if the step fails.


Cancel Save


**Step 2** will combine those 4 reduced files to 1


**Add step** ✕


**Step type** Streaming program

**Name\***

**Mapper\***   S3 location of the map function or the name of the Hadoop streaming command to run.


**Reducer\***   S3 location of the reduce function or the name of the Hadoop streaming command to run.

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

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

**Arguments**

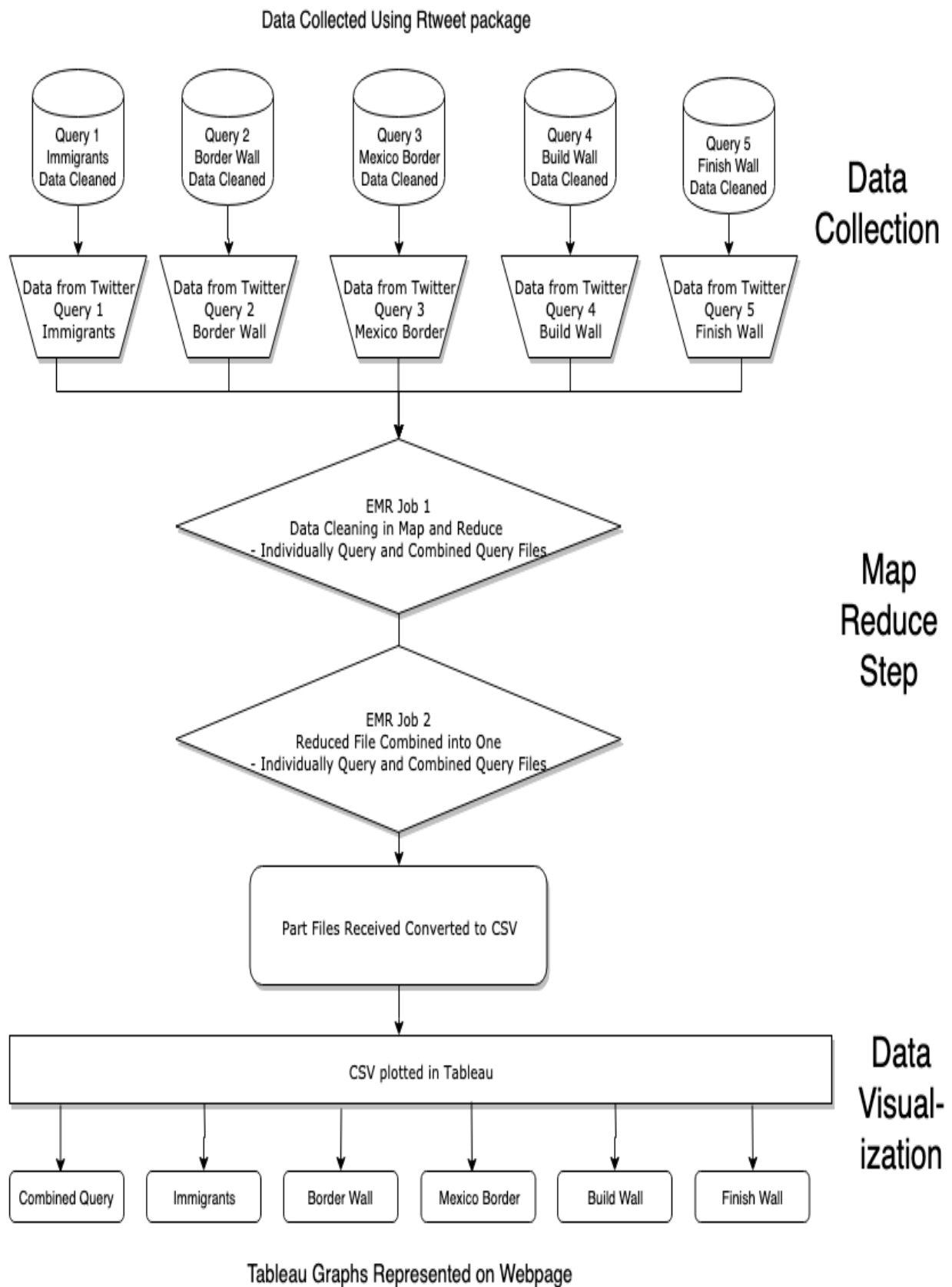
```
-jobconf mapred.reduce.tasks=1
```

**Action on failure**   What to do if the step fails.

Cancel Save

3. Part-00000 file received

## Job Flow for Twitter Data



## Job Flow for Common Crawl Data

Data Collected Using Common Crawler WET files

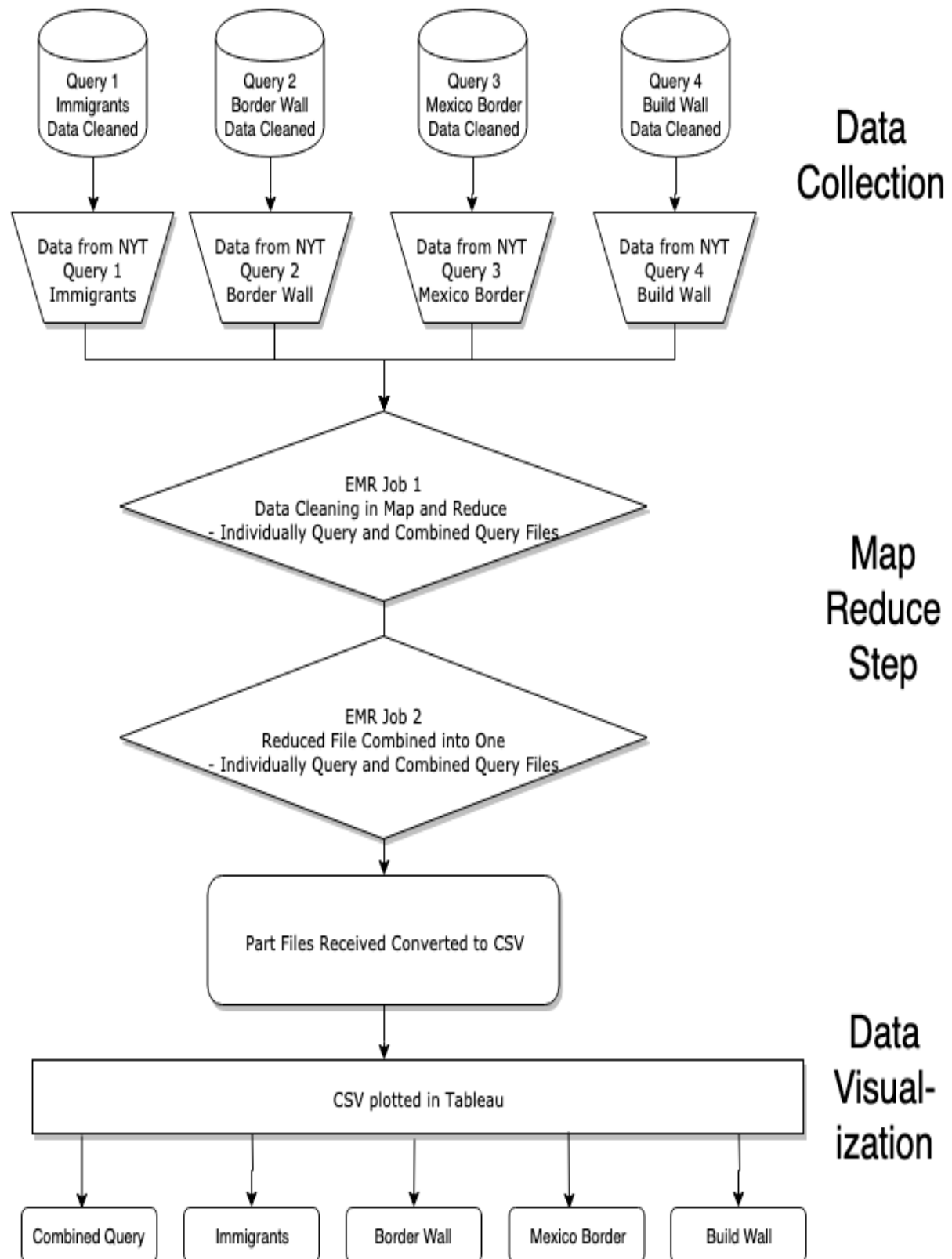
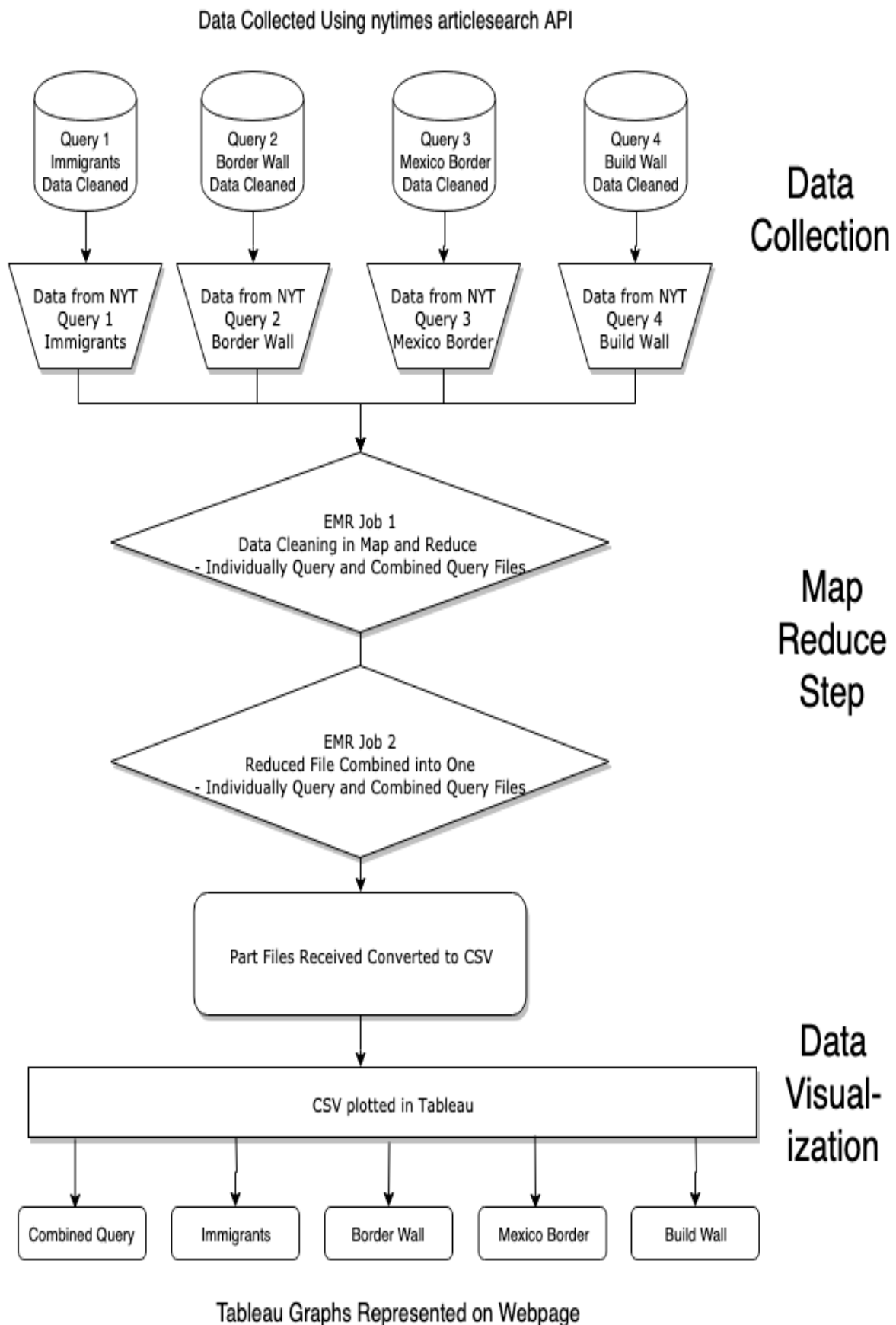


Tableau Graphs Represented on Webpage

## Job Flow for NYT Data



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.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.

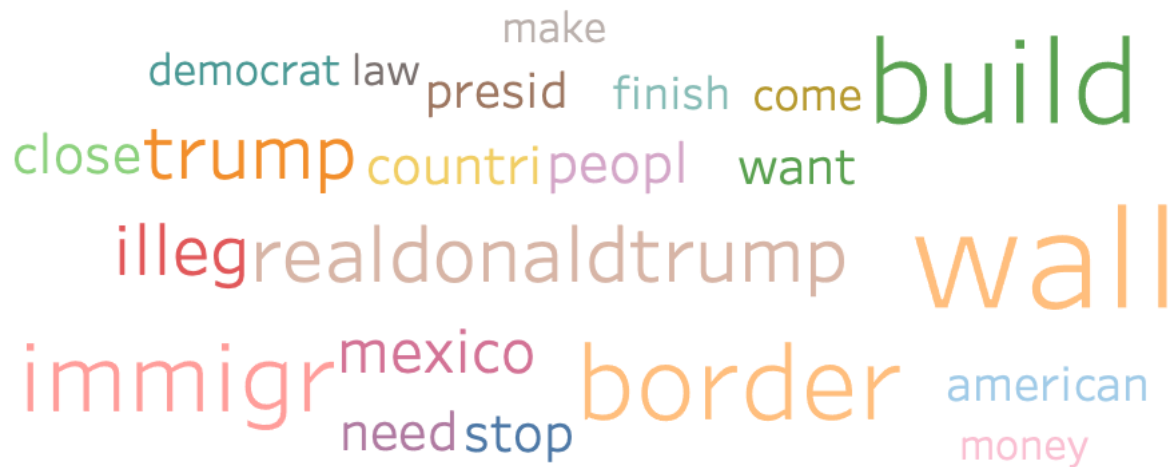
## Data Visualization

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



## A word cloud visualization of terms related to the 'Build the Wall' campaign. The words are arranged in a circular pattern, with their size and color corresponding to their frequency. The word 'build' is the largest and most central, colored green. Other prominent words include 'wall' (orange), 'trump' (orange), 'border' (orange), 'america' (blue), 'mexico' (grey), 'money' (pink), 'immigr' (red), 'people' (purple), 'real' (brown), 'dona' (brown), 'ld' (brown), 'trump' (brown), 'stop' (blue), 'keep' (red), 'pay' (purple), 'damn' (blue), 'need' (pink), 'come' (yellow), 'law' (grey), 'want' (green), 'presid' (grey), 'countri' (yellow), 'close' (green), 'american' (blue), 'illeg' (teal), and 'country' (yellow). A small white box with a black border is overlaid on the word 'build', containing the text 'Word: build' and 'Count: 13,580'.

## A word cloud featuring various terms associated with the US-Mexico border wall. The most prominent words are 'finish' and 'wall', both in large, bold, brownish-gold fonts. Other significant words include 'border' in red, 'build' in purple, 'great' in green, 'people' in blue, 'stop' in orange, 'back' in teal, 'immigr' in brown, 'start' in blue, 'close' in brown, 'pleas' in brown, 'america' in orange, 'trump' in orange, 'need' in pink, 'presid' in brown, 'down' in blue, 'money' in teal, 'come' in brown, 'end' in yellow, 'countri' in pink, 'illeg' in blue, and 'real' in brown. The words are arranged in a somewhat circular pattern, with 'finish' and 'wall' dominating the center.

## A word cloud featuring various terms related to the US-Mexico border. The most prominent word is 'border' in large green letters. Other significant words include 'mexico' in purple, 'trump' in teal, and 'real donald trump' in blue. Smaller words include 'stop', 'american', 'judg', 'back', 'threaten', 'southern', 'close', 'illeg', 'mexican', 'down', 'aid', 'immigr', 'danger', 'cross', 'shut', 'wall', 'peopl', 'presid', 'countri', 'need', 'migrant', 'come', 'asylum', and 'judg'. The words are arranged in a dense, overlapping manner with various colors.

### Word Cloud on Immigrant



A word cloud visualization of terms related to immigrants. The most prominent word is 'immigr' in a large, teal font. Other significant words include 'border' in green, 'illeg' in green, 'trump' in blue, 'america' in blue, 'white' in orange, 'real' in purple, 'donald' in purple, 'trump' in purple, 'stop' in brown, 'make' in brown, 'work' in green, 'countri' in yellow, 'people' in purple, 'need' in pink, 'law' in pink, 'legal' in blue, 'mexico' in red, 'know' in red, 'american' in blue, 'democrat' in yellow, 'want' in blue, and 'come' in green.

come  
america trump white real donald trump  
want stop make  
border immigr illeg work  
democrat  
american legal mexico know countri  
need law people

### Word Co-occurrence

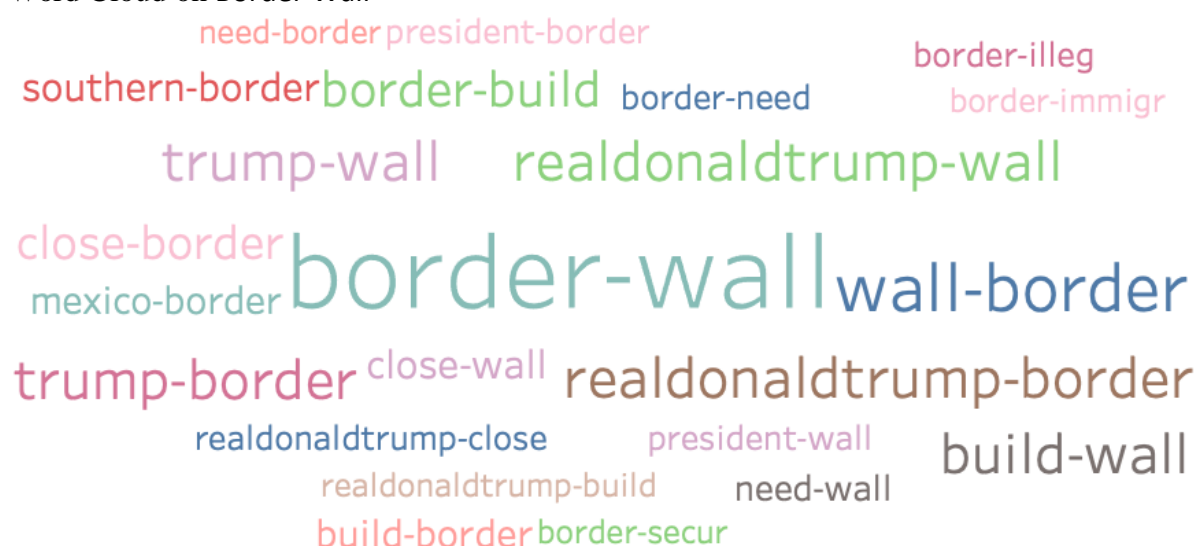
#### Word Cloud on Full Data



A word cloud visualization of terms related to border walls. The most prominent words are 'build-wall' in green, 'border-wall' in orange, and 'wall-build' in orange. Other significant words include 'finish-wall' in blue, 'mexico-border' in red, 'border-mexico' in orange, 'wall-illeg' in orange, 'illegal-immigr' in blue, 'trump-build' in blue, 'close-wall' in yellow, 'wall-border' in blue, 'real donald trump-immigr' in pink, 'real donald trump-build' in pink, 'trump-wall' in blue, 'border-build' in blue, 'real donald trump-border' in pink, 'real donald trump-wall' in pink, 'close-border' in yellow, 'trump-border' in blue, 'southern-border' in blue, 'border-immigr' in blue, 'build-border' in green, 'president-wall' in red, and 'trump-immigr' in blue.

finish-wall wall-build  
mexico-border border-mexico wall-illeg  
illegal-immigr trump-build close-wall wall-border  
real donald trump-immigr real donald trump-build  
trump-wall border-build build-wall border-wall  
real donald trump-border real donald trump-wall  
close-border trump-border southern-border border-immigr  
build-border president-wall trump-immigr

#### Word Cloud on Border Wall

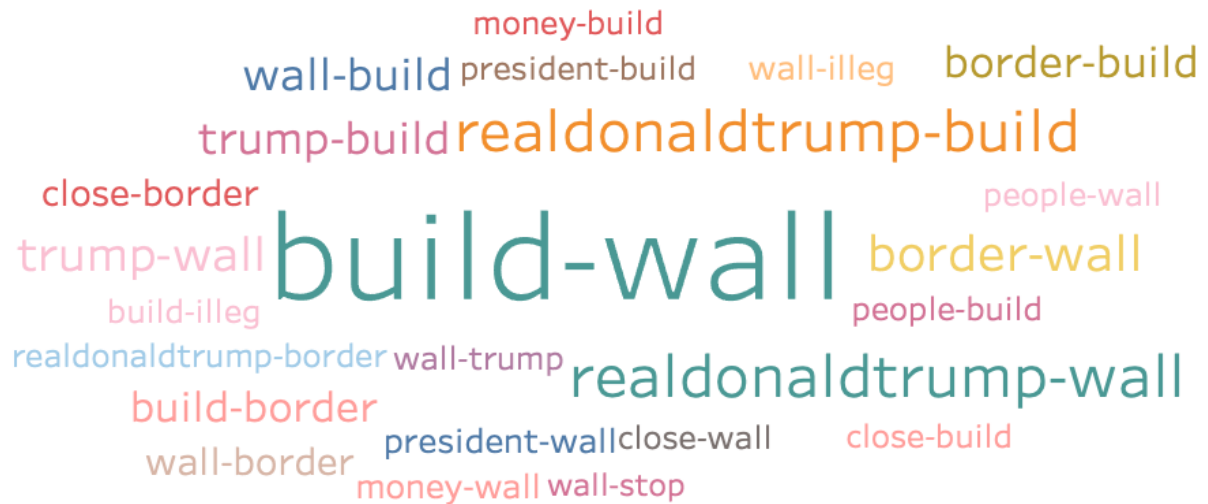


A word cloud visualization of terms related to border walls. The most prominent words are 'border-wall' in teal, 'wall-border' in blue, and 'border-build' in green. Other significant words include 'trump-wall' in purple, 'real donald trump-wall' in green, 'close-border' in pink, 'mexico-border' in blue, 'trump-border' in pink, 'close-wall' in pink, 'real donald trump-border' in brown, 'real donald trump-close' in blue, 'president-wall' in purple, 'build-wall' in brown, 'real donald trump-build' in blue, 'need-wall' in blue, 'build-border' in red, 'border-secur' in green, 'need-border' in pink, 'president-border' in pink, 'border-illeg' in pink, 'border-immigr' in pink, and 'border-need' in blue.

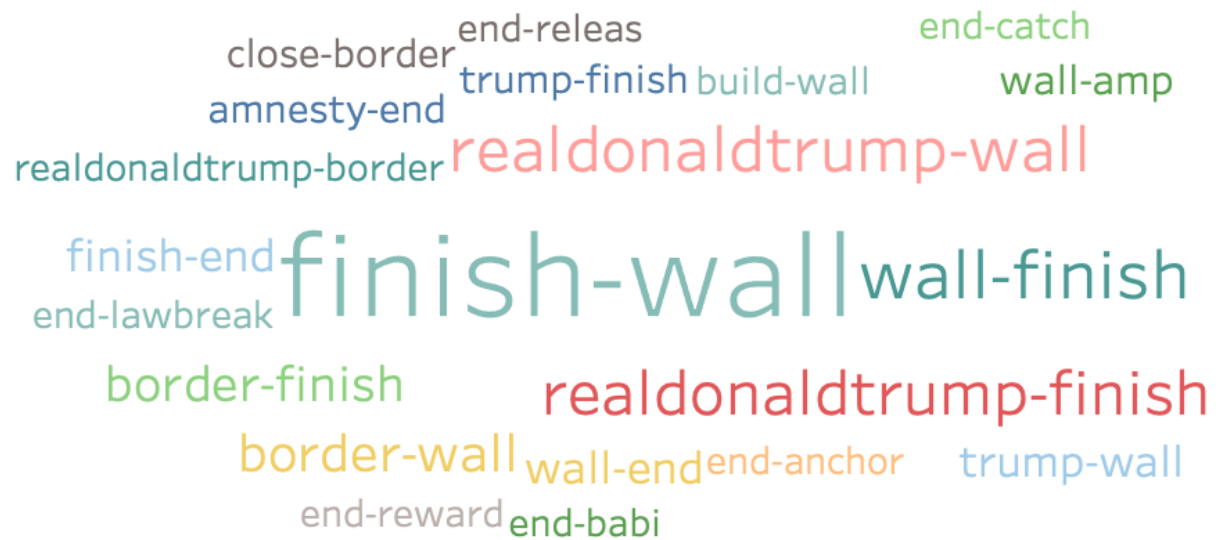
need-border president-border border-illeg  
southern-border border-build border-need border-immigr  
trump-wall real donald trump-wall  
close-border mexico-border border-wall wall-border  
trump-border close-wall real donald trump-border  
real donald trump-close president-wall build-wall  
real donald trump-build need-wall  
build-border border-secur



### Word Cloud on Build Wall



### Word Cloud on Finish Wall



### Word Cloud on Mexico Border



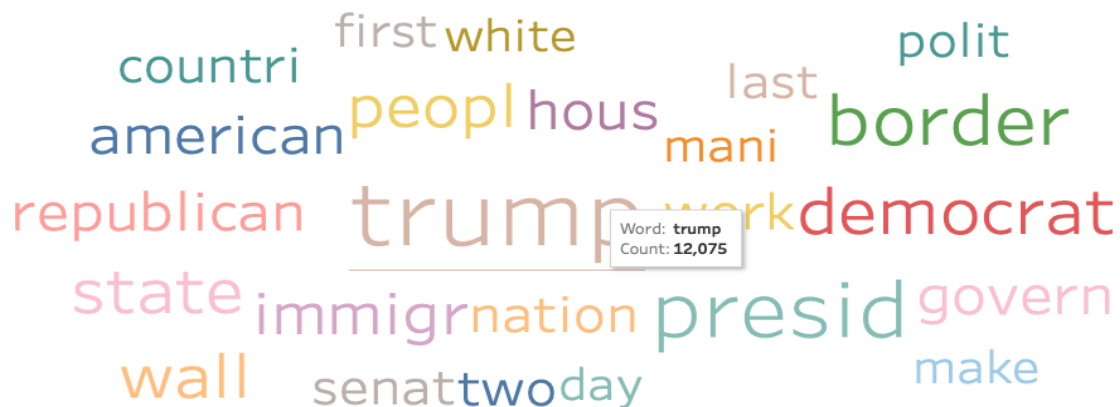
## Word Cloud on Immigrant



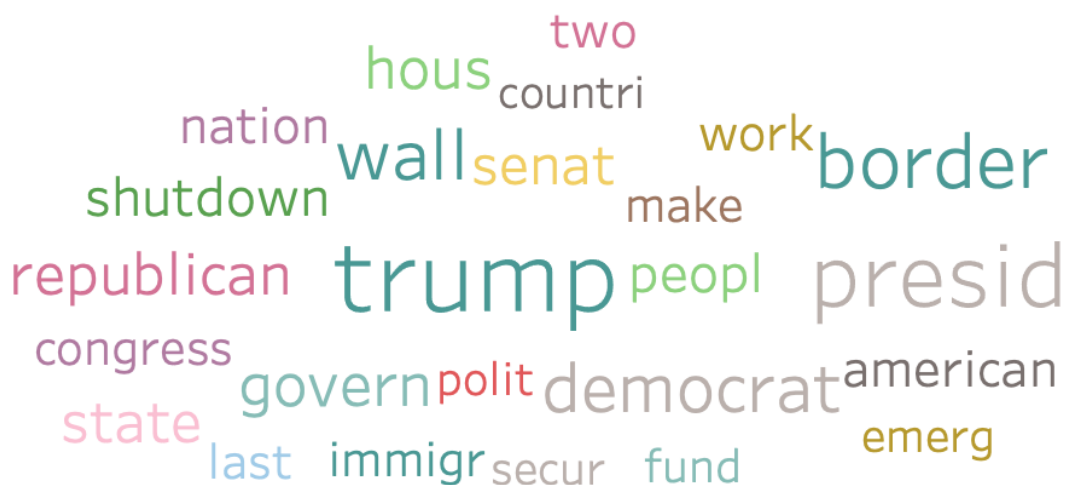
## New York Times

### Word Count

#### Word Cloud on Full Data



## Word Cloud on Border Wall



Word Cloud on Build Wall

build think support congress  
two wall nation fund democrat hous  
republican trump thi presid  
govern american immigrer senat  
state peopl emerg countri polit  
border year

Word Cloud on Mexico Border

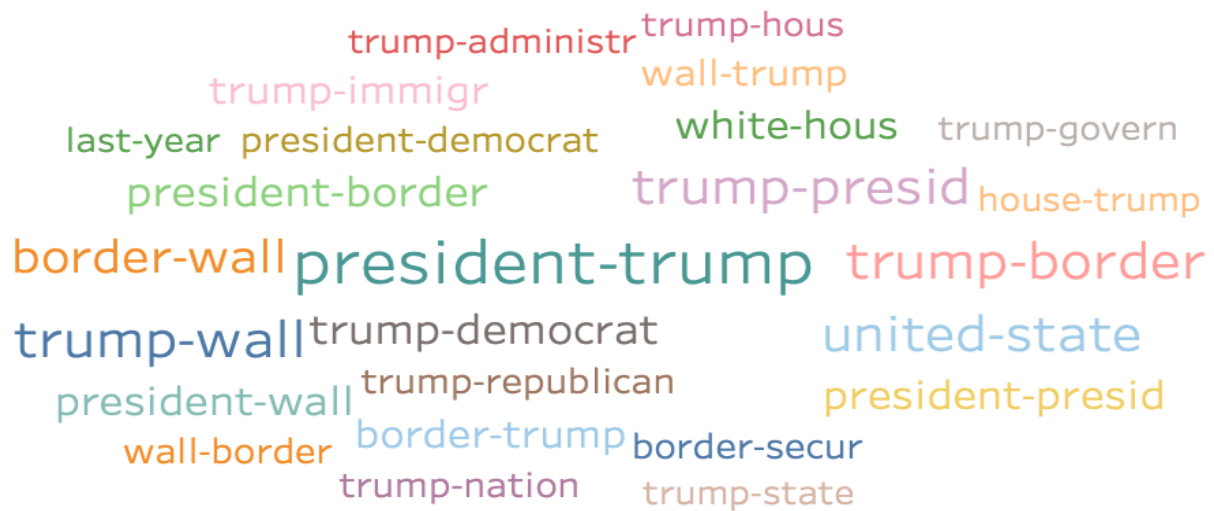
peopl countri wall last  
migrant mexico year emerg  
republican border nation trump  
american govern democrat immigr  
two state mexican work offici  
secur senat

Word Cloud on Immigrant

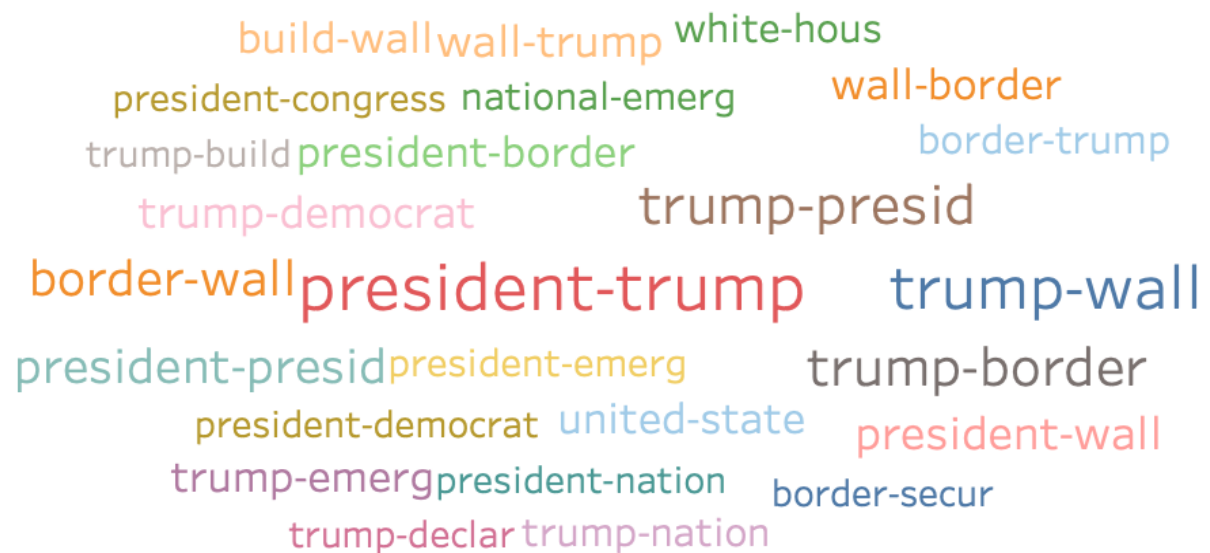
nation work first unit wall hous  
citi year peopl govern last american  
presid polit trump mani democrat  
republican state immigr border  
countri make two  
york

## Word Co-occurrence

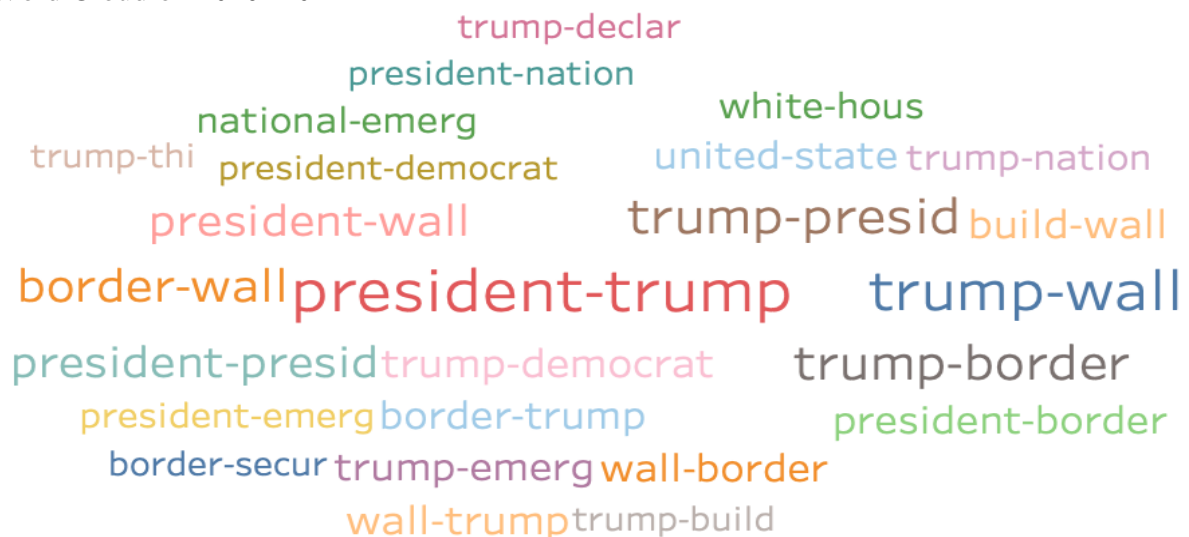
Word Cloud on Full Data



Word Cloud on Border Wall



Word Cloud on Build Wall



### Word Cloud on Mexico Border

border-secur wall-border  
trump-administr states-border border-state  
border-trump border-migrant president-wall trump-immigr  
mexico-border border-wall united-state  
trump-wall president-trump trump-border  
trump-presid president-presid president-border  
border-mexico mexican-border migrants-border  
border-cross trump-emerg trump-nation  
trump-state

### Word Cloud on Immigrant

border-secur  
president-border border-trump  
anti-semit president-democrat students-school border-immigr  
trump-immigr democrats-trump trump-presid  
border-wall president-trump united-state  
trump-border high-school trump-democrat trump-govern  
trump-republican house-democrat trump-wall  
trump-administr president-wall trump-state  
last-year white-hous

### Common Crawl

#### Word Count

#### Word Cloud on Full Data

white  
law national history news visa  
world business information trump  
department immigration government  
international march state war american  
president states united make best social  
wall

A word cloud of terms related to the 2017 US inauguration. The word 'trump' is the largest and most central. Other prominent words include 'march', 'news', 'tax', 'watch', 'national', 'american', 'best', 'economic', 'donald', 'decor', 'state', 'good', 'wall', 'president', 'feb', 'border', 'business', 'diego', 'room', 'government', 'economy', 'san', 'white', 'world', 'war', 'down', and 'house'. The words are in various colors and orientations, creating a dynamic visual representation of the event's associated vocabulary.

Word cloud visualization of the German text: "Der White Bersicht und der Jahrgang Wahlmodul nach Medien Materialien die Window Holz Design Light Storage Fragen". The words are arranged in a circular pattern, with "Bersicht" and "und" being the largest and most prominent. Other visible words include "Jahrgang", "Wahlmodul", "nach", "Medien", "Materialien", "die", "Window", "Holz", "Design", "Light", "Storage", "Fragen", "Blick", "Blickpunkt", "Blickfeld", "Blickwinkel", "Blickrichtung", "Blickhöhe", "Blicktiefe", "Blickweite", "Blicklänge", "Blickbreite", "Blicktiefe", "Blickweite", "Blicklänge", "Blickbreite", "Blicktiefe", "Blickweite", "Blicklänge", "Blickbreite".

A word cloud featuring various terms related to academic writing and research. The most prominent words are 'viagra' and 'essay', both in large orange font. Other significant words include 'thesis' in blue, 'essays' in teal, 'study' in brown, 'writing' in green, 'research' in purple, 'paper' in pink, 'series' in brown, 'workshop' in orange, 'download' in green, 'answers' in blue, 'college' in orange, 'levitra' in orange, 'online' in grey, 'edition' in yellow, 'life' in pink, 'help' in teal, 'write' in orange, 'owners' in pink, 'ebook' in green, 'manuals' in brown, 'buy' in blue, and 'user' in blue. The words are arranged in a dense, overlapping manner on a white background.

john war  
national united march public make  
history business information world  
american immigration government  
department states visa international  
law state country social best trump  
news

Unable to run even with 20 instances on AWS.