NYU | Center for Urban Science + Progress

Big Data Management and Analysis
Spring 2017

PULSE OF NEW YORK TIMES

Alexey Kalinin, Enrique Gonzalez, Priyanshi Singh,
Shalmali Kulkarni, Sunny Kulkarni

AGENDA

- ❖ Introduction
- ❖ What’s Popular
- ❖ Project Objectives
- ❖ Data Set
- ❖ Methodology
- ❖ Sentiment Analysis
- ❖ Big Data Challenges
- ❖ Descriptive Statistics
- ❖ Visualization
- ❖ Next Steps
- ❖ Appendix (scripts and job log)

INTRODUCTION

NYU Center for Urban Science + Progress

The New York Times

INTERNATIONAL EDITION | WEDNESDAY, MAY 3, 2017

Uncertainty bordering on dismay

Roger Cohen

OPINION

'It's easy to kill a journalist'

TIERRA BLANCA, MEXICO

The attacks on reporters in Mexico's Veracruz State put chill on press freedom

In France, voters who skip could help Le Pen

STAINS, FRANCE

Many in poor suburbs say they don't trust politics and plan to stay home

"All the News That's Fit to Print"

The New York Times

Late Edition

VOL. CLXVI . . . No. 57,586 NEW YORK, WEDNESDAY, MAY 3, 2017 $2.50

TRUMP AND PUTIN AGREE TO PURSUE SYRIA CEASE-FIRE

LEADERS RESUME TALKS

A 'Constructive Call,' but U.S. and Russian Accounts Differ

By PETER BAKER and NEIL MacFARQUHAR

Ivanka Trump said she intended to act as a moderating influence on her father's administration.

A First Daughter Assembling a Vast Portfolio

Among Mostly Men, a Trump Takes Aim at Gender Issues

'Part of the Resistance'

Patients at Risk Fear the Worst From Waivers

Pre-existing Conditions May Raise Premiums

By ABBY GOODNOUGH and REED ABELSON

G.O.P. SCRAMBLES AS CRUCIAL VOICE SHUNS CARE BILL

CHASING REPEAL VOTES

Concessions to Freedom Caucus Are Called a Deal-Breaker

By THOMAS KAPLAN and ROBERT PEAR

Representative Fred Upton of Michigan criticized the bill.

- **Sentiment Analysis** identifying positive and negative opinions, emotions and evaluations in text.

WHAT'S POPULAR…

Popular Google Search

Popular on Facebook

Popular on Twitter

Popular Emailed

PROJECT OBJECTIVES

- ❖ Understand NYTimes trends on Twitter across US
- ❖ Reveal Spatial clusters for the tweeted subjects
- ❖ Sentiment analysis of most popular NY Times articles tweets
- ❖ Visualizations

DATASETS

- **Twitter**

Geocoded tweets

- **NYTimes Articles**

Popular Articles

- **NLTK**

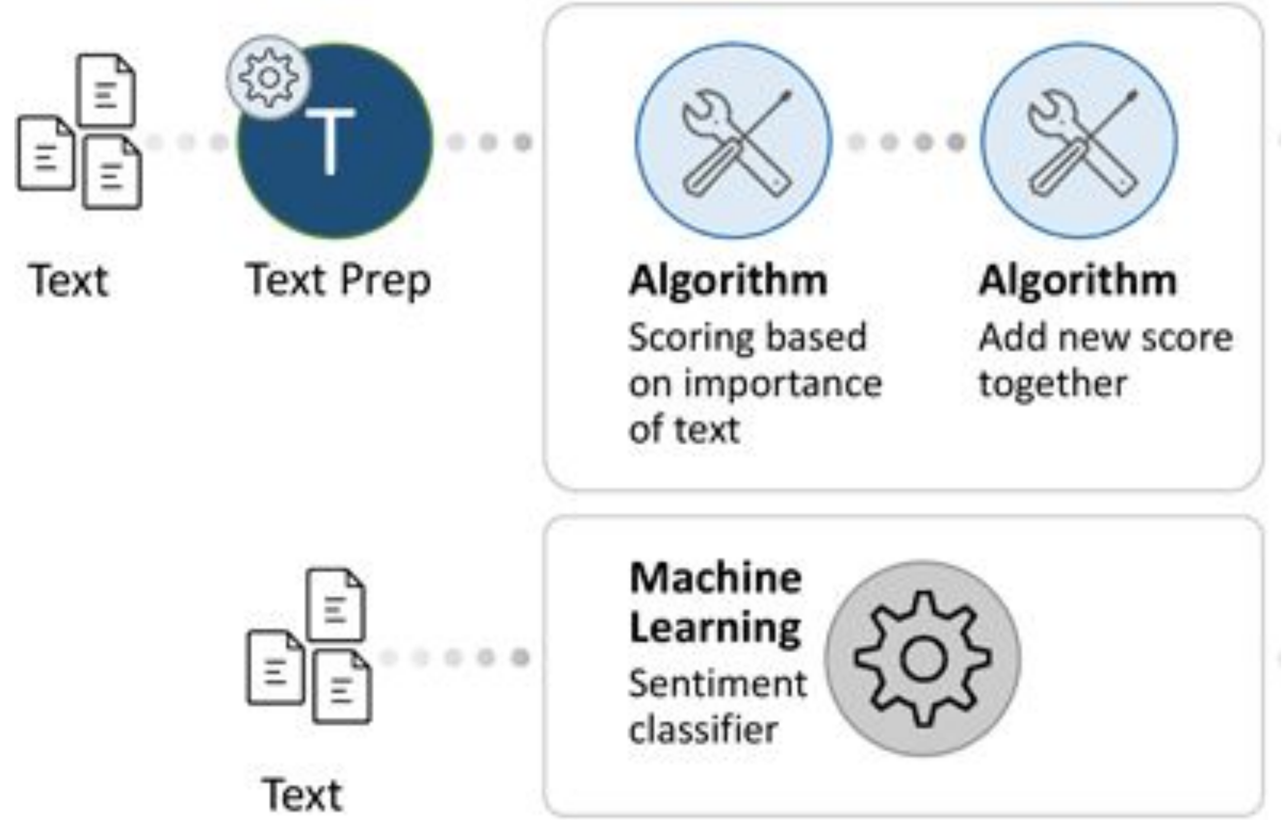
Sentiment Word List

METHODOLOGY

Machine Learning Pipeline

Sentiment Analysis using NLTK

SENTIMENT ANALYSIS



BIG DATA CHALLENGES

❖ **Scalability**

❖ **Data Volume & Quality**

❖ **Data Structure**

DESCRIPTIVE STATISTICS

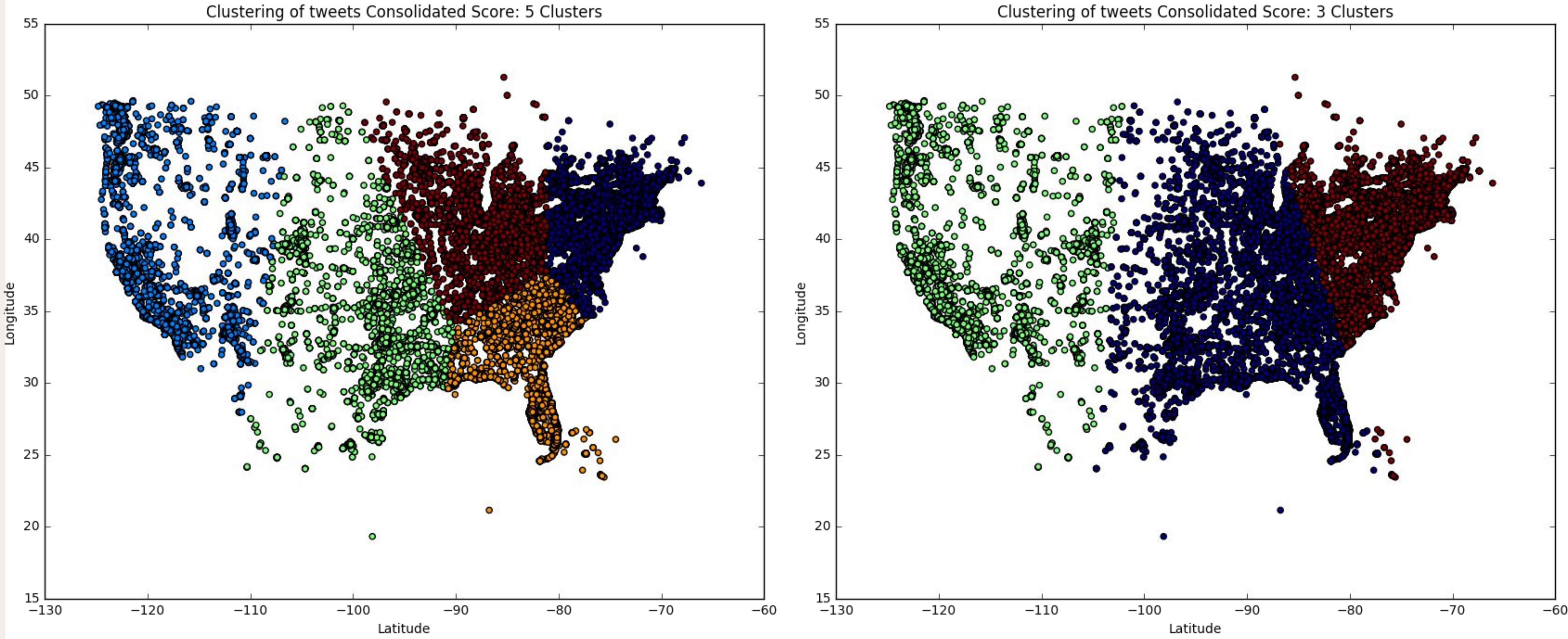
Time frame: April 28th to May 3rd 2017

No. of Tweets - approx. 600,000 records (3.2 GB)

No.of Geocoded Tweets - 100,000 records

NYTimes - 10 popular Articles per day

VISUALIZATION - SPATIAL CLUSTERING

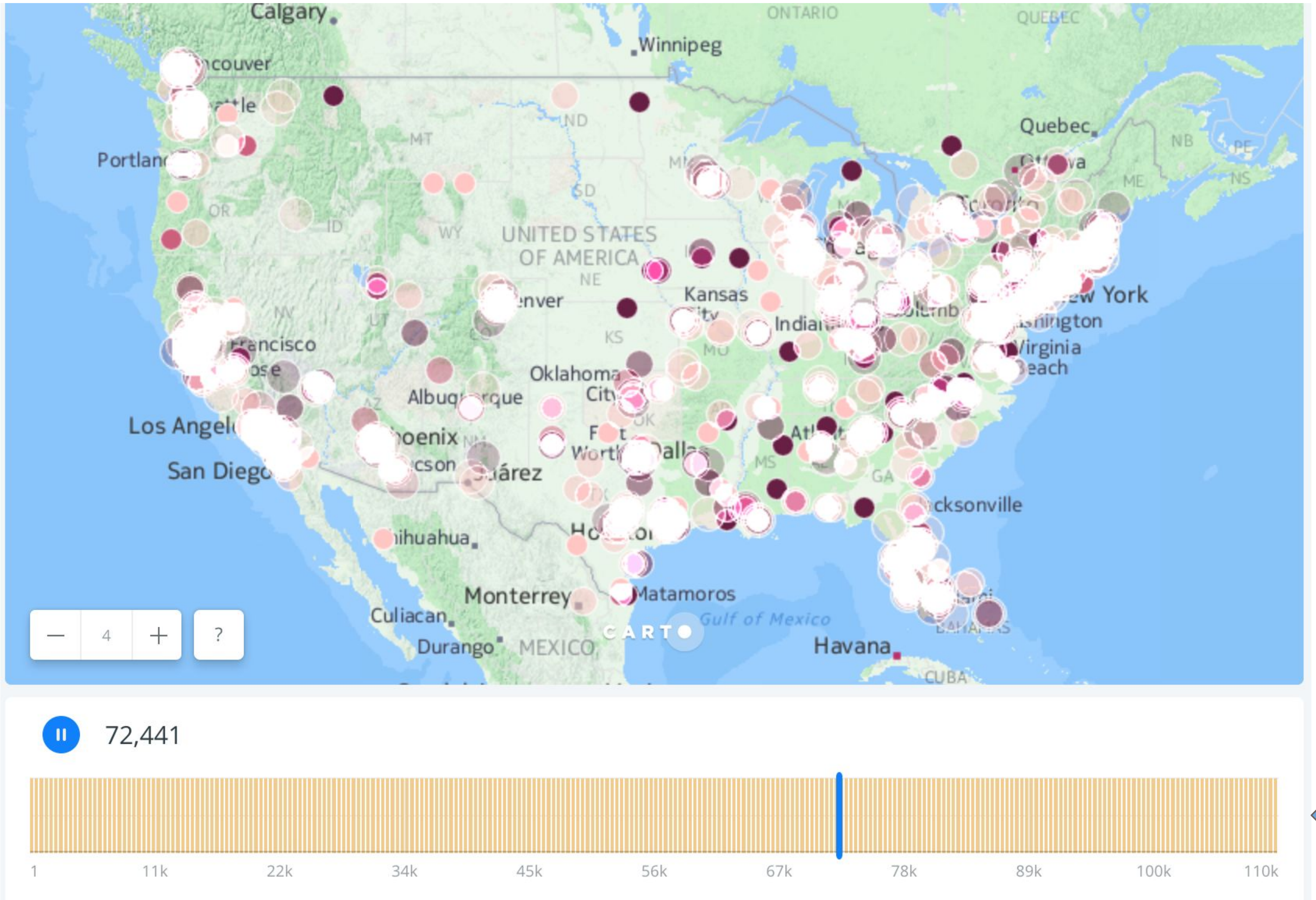


Clustering of tweets Consolidated Score: 5 Clusters

Clustering of tweets Consolidated Score: 3 Clusters

Clusters found after spatial clustering

VIS. - SENTIMENTS ACROSS USA



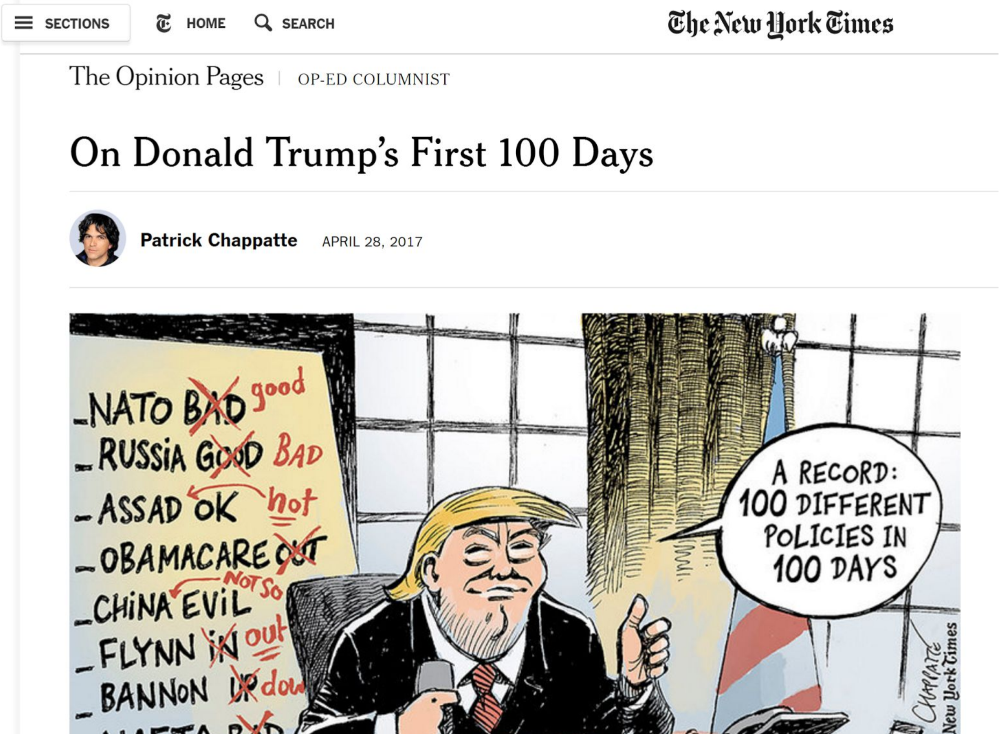
Sentiments attached to the tweets across US with Time

Real Time

Timeline

TRENDING TOPICS

NYU Center for Urban Science + Progress



≡ SECTIONS HOME SEARCH The New York Times

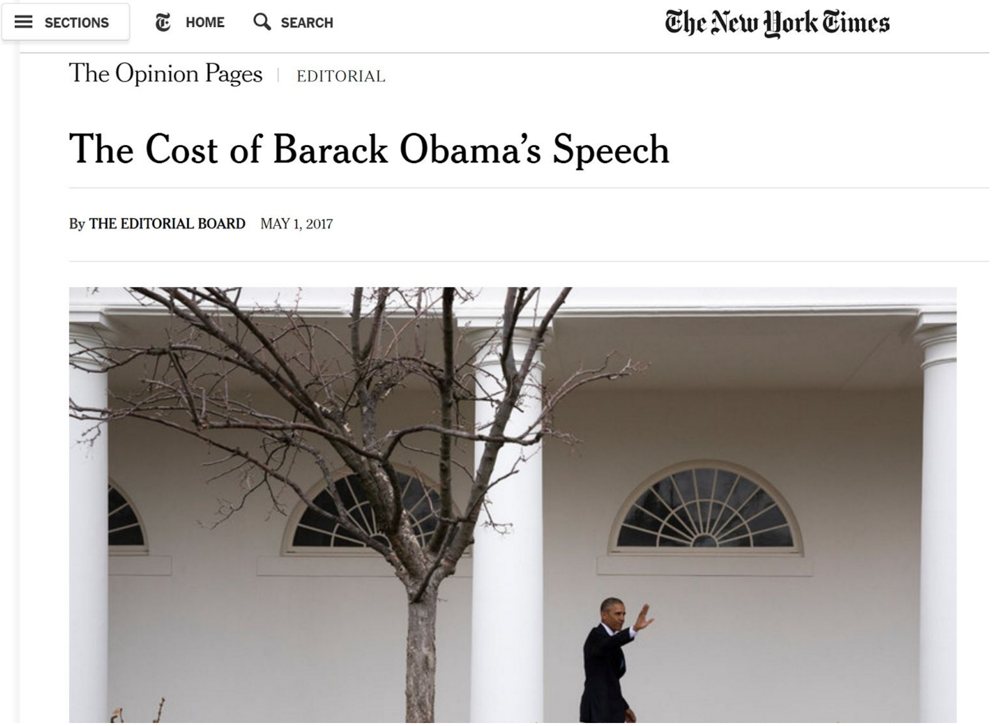
The Opinion Pages | OP-ED COLUMNIST

On Donald Trump's First 100 Days

Patrick Chappatte APRIL 28, 2017

28th April, 2017

1st May, 2017



≡ SECTIONS HOME SEARCH The New York Times

The Opinion Pages | EDITORIAL

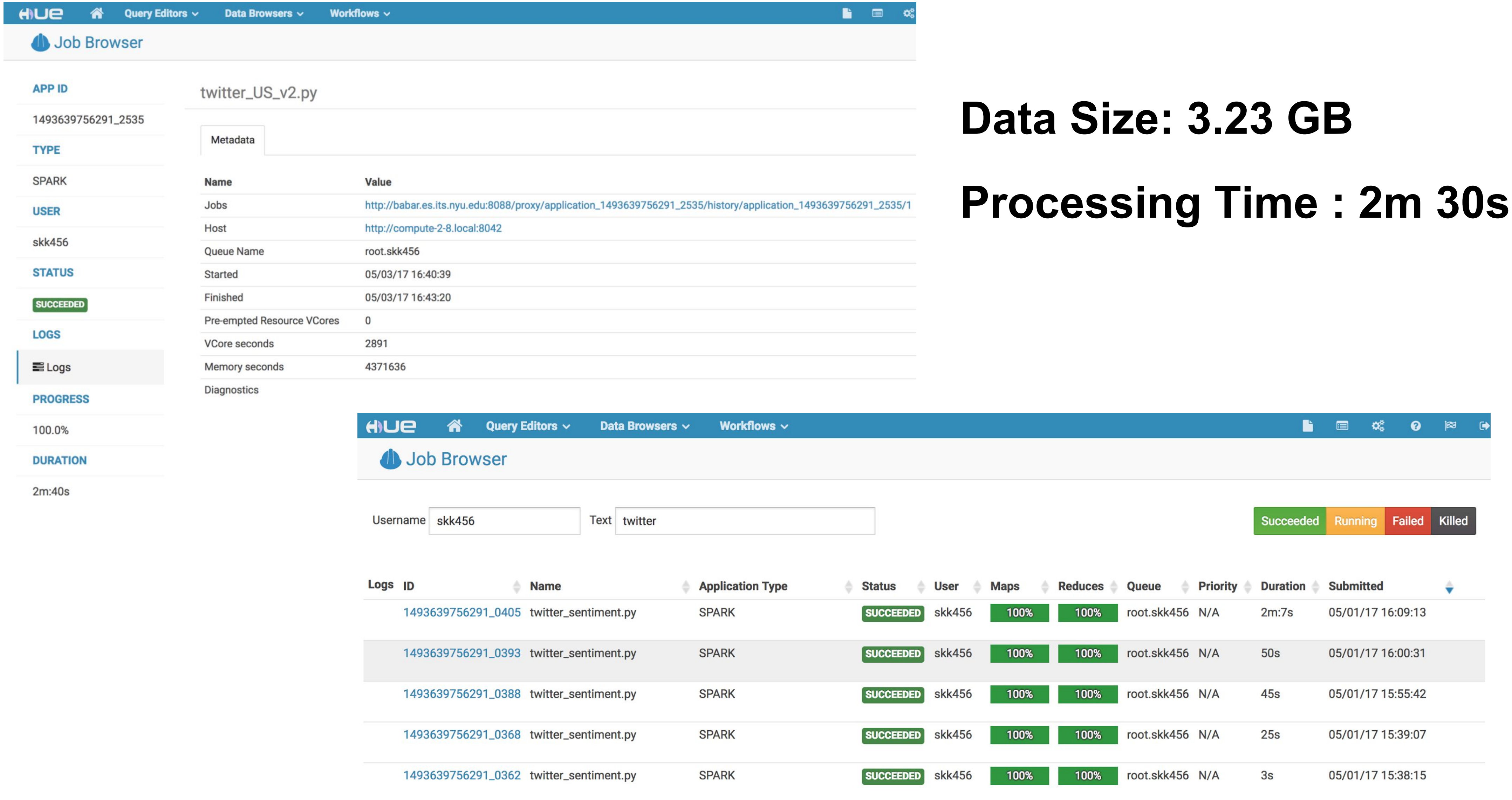
The Cost of Barack Obama's Speech

By THE EDITORIAL BOARD MAY 1, 2017

PROCESSING TIME

Data Size: 3.23 GB

Processing Time : 2m 30s



APP ID

1493639756291_2535

TYPE

SPARK

USER

skk456

STATUS

SUCCEEDED

LOGS

Logs

PROGRESS

100.0%

DURATION

2m:40s

twitter_US_v2.py

Metadata

Name	Value
Jobs	http://babar.es.its.nyu.edu:8088/proxy/application_1493639756291_2535/history/application_1493639756291_2535/1
Host	http://compute-2-8.local:8042
Queue Name	root.skk456
Started	05/03/17 16:40:39
Finished	05/03/17 16:43:20
Pre-empted Resource VCores	0
VCore seconds	2891
Memory seconds	4371636
Diagnostics	

Username skk456 Text twitter

Succeeded Running Failed Killed

Logs	ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1493639756291_0405	twitter_sentiment.py	SPARK	SUCCEEDED	skk456	100%	100%	root.skk456	N/A	2m:7s	05/01/17 16:09:13
	1493639756291_0393	twitter_sentiment.py	SPARK	SUCCEEDED	skk456	100%	100%	root.skk456	N/A	50s	05/01/17 16:00:31
	1493639756291_0388	twitter_sentiment.py	SPARK	SUCCEEDED	skk456	100%	100%	root.skk456	N/A	45s	05/01/17 15:55:42
	1493639756291_0368	twitter_sentiment.py	SPARK	SUCCEEDED	skk456	100%	100%	root.skk456	N/A	25s	05/01/17 15:39:07
	1493639756291_0362	twitter_sentiment.py	SPARK	SUCCEEDED	skk456	100%	100%	root.skk456	N/A	3s	05/01/17 15:38:15

NEXT STEPS…

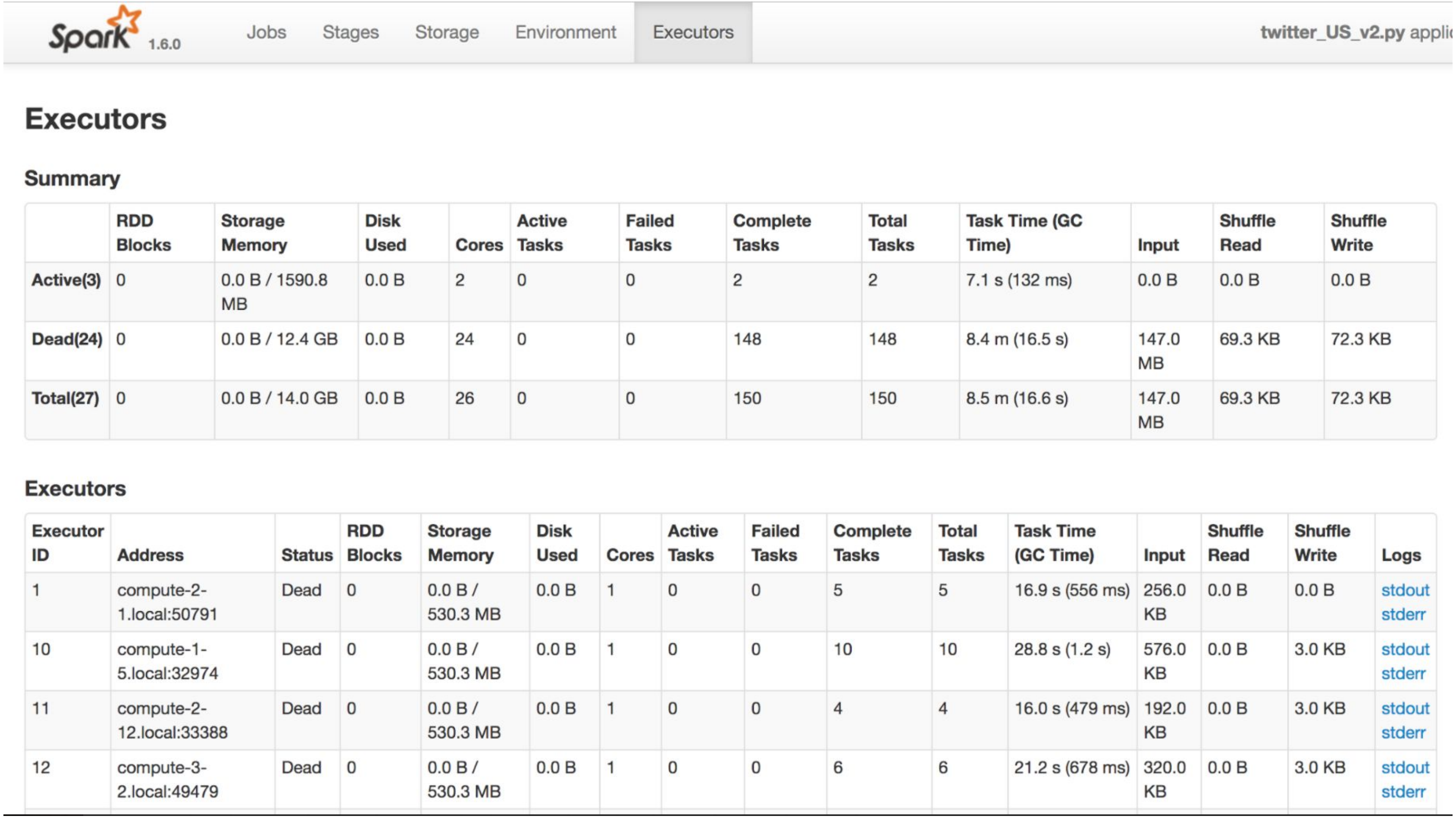
Process Pipeline On Live Data Using Spark Streaming.

Spark Streaming

Spark.MLlib to Train Models and Perform Prediction

Spark MLlib

APPENDIX – Spark Job Log



Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(3)	0	0.0 B / 1590.8 MB	0.0 B	2	0	0	2	2	7.1 s (132 ms)	0.0 B	0.0 B	0.0 B
Dead(24)	0	0.0 B / 12.4 GB	0.0 B	24	0	0	148	148	8.4 m (16.5 s)	147.0 MB	69.3 KB	72.3 KB
Total(27)	0	0.0 B / 14.0 GB	0.0 B	26	0	0	150	150	8.5 m (16.6 s)	147.0 MB	69.3 KB	72.3 KB

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
1	compute-2-1.local:50791	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	5	5	16.9 s (556 ms)	256.0 KB	0.0 B	0.0 B	stdout stderr
10	compute-1-5.local:32974	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	10	10	28.8 s (1.2 s)	576.0 KB	0.0 B	3.0 KB	stdout stderr
11	compute-2-12.local:33388	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	4	4	16.0 s (479 ms)	192.0 KB	0.0 B	3.0 KB	stdout stderr
12	compute-3-2.local:49479	Dead	0	0.0 B / 530.3 MB	0.0 B	1	0	0	6	6	21.2 s (678 ms)	320.0 KB	0.0 B	3.0 KB	stdout stderr

QUESTIONS?

THANK YOU!