

Sentiment Analysis on Big Data of Social and Political events

By
Rowan Menezes
Suraj Jayakumar

Project Guide: Prof. Dr. Indiramma M



Introduction

What is Big Data?

Big data is a broad term for data sets so large that traditional data processing applications are inadequate

Why is there a boom?

Organizations create and store more transactional data in digital form

Big Data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services

What is sentiment Analysis? (Opinion Mining)

Process of computationally identifying and categorizing opinions expressed in a piece of text in order to determine the writer's attitude towards a particular topic is positive, negative or neutral.



Problem statement and Target

Issues we are trying to solve?

To build a comprehensive and efficient sentiment analyzer for social and political topics such that given a message, classify whether the message is of positive, negative or neutral sentiment.

For messages conveying both positive and negative sentiments, whichever is the stronger sentiment is chosen.

Target Users?

Politicians

Firms and Organizations

Normal end users



USA 2012 election – Obama Victory

How President Obama's campaign used big data to rally individual voters

Core team of Data Scientists

Stocks price prediction based on market movements

The correlation between news articles and stock variations is already proved

Personalized Recommendations by e-Commerce sites



Social Impact

Politicians will get a feedback on public opinion on their actions

Example – Recent Comment on SRK by Kailash Vijayvargiya which said his “soul” was in Paksitan which resulted in a public outrage (could be avoided using this)

People can monitor public view about their MLA/MP in that constituency

Example – People get a holistic opinion about their elected representative

Social Impact



Firms/Companies will get a comprehensive analysis of the event they organized

Recent Flipkart's Big Billion Day vs Amazon's Sale

We can predict the maximum reach of a Politician's social media message based on past history

Example - Maximum reach to public on Thursday 8pm

Feedback mechanism for any new law passed and new movie released



Challenges



Sarcasm

My hostel WiFi speed clearly has to be fastest in the world.

Context / Domain Dependence

The story of the movie was really **unpredictable**.

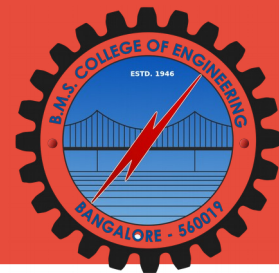
Actions of my MLA are totally **unpredictable**.

Thwarted Expectations

Start one direction and end it with opposite meaning

Media attached to tweets/ social media message

@Sachin_RT: Virat's batting today has been t.co/alPoLX



Challenges

Pragmatics (Awwwesome, BAD)

Awwwesomeeee vs Awesome

BAD vs bad

Chat corpus (diz, wer, der)

Word Knowledge

He is a Hitler

Comparative

Modi is much better than Rahul Gandhi

External Links

Breaking News: Modi says <link>



So Whats New?

Machine Learning

Naive-Bayes Classifier

Classification using Maximum Entropy

Support Vector Machines (SVM)

Deep Learning

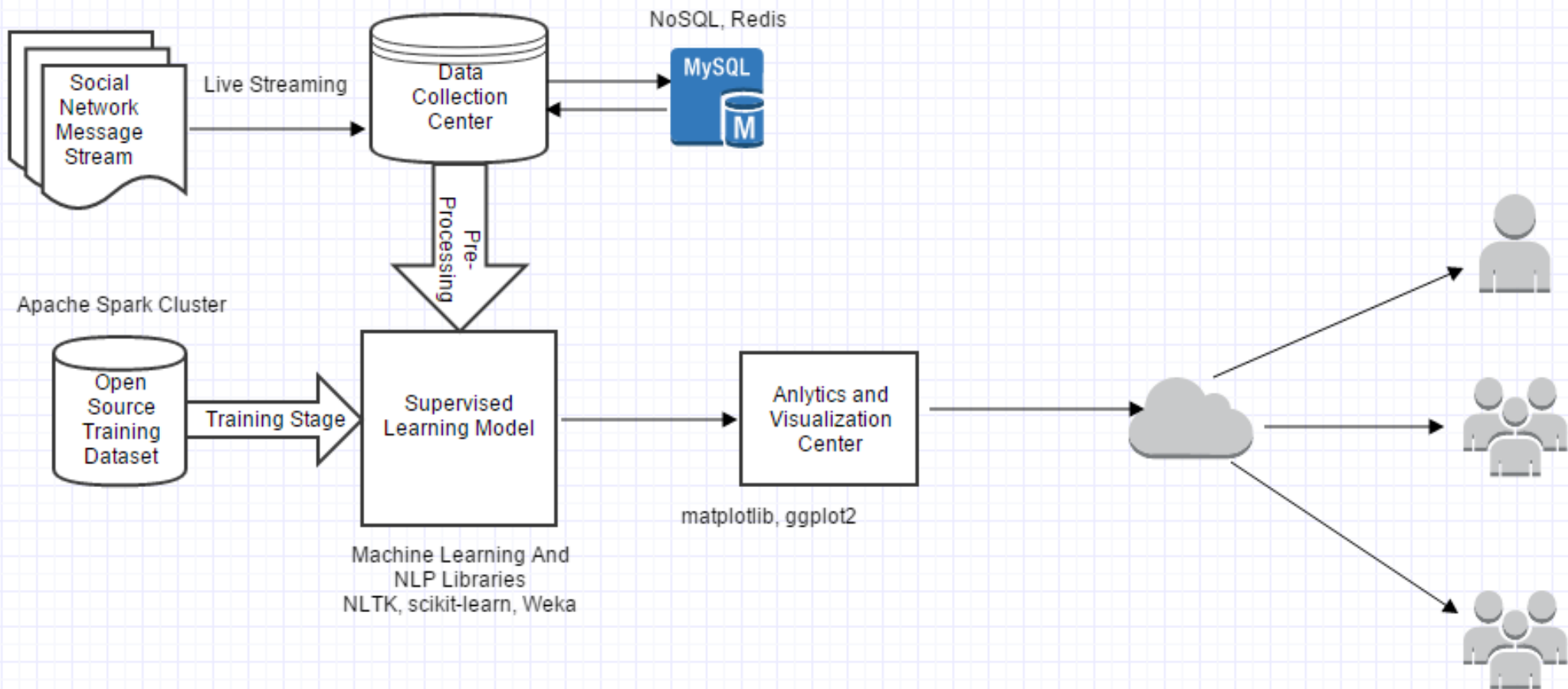
Neural Networks



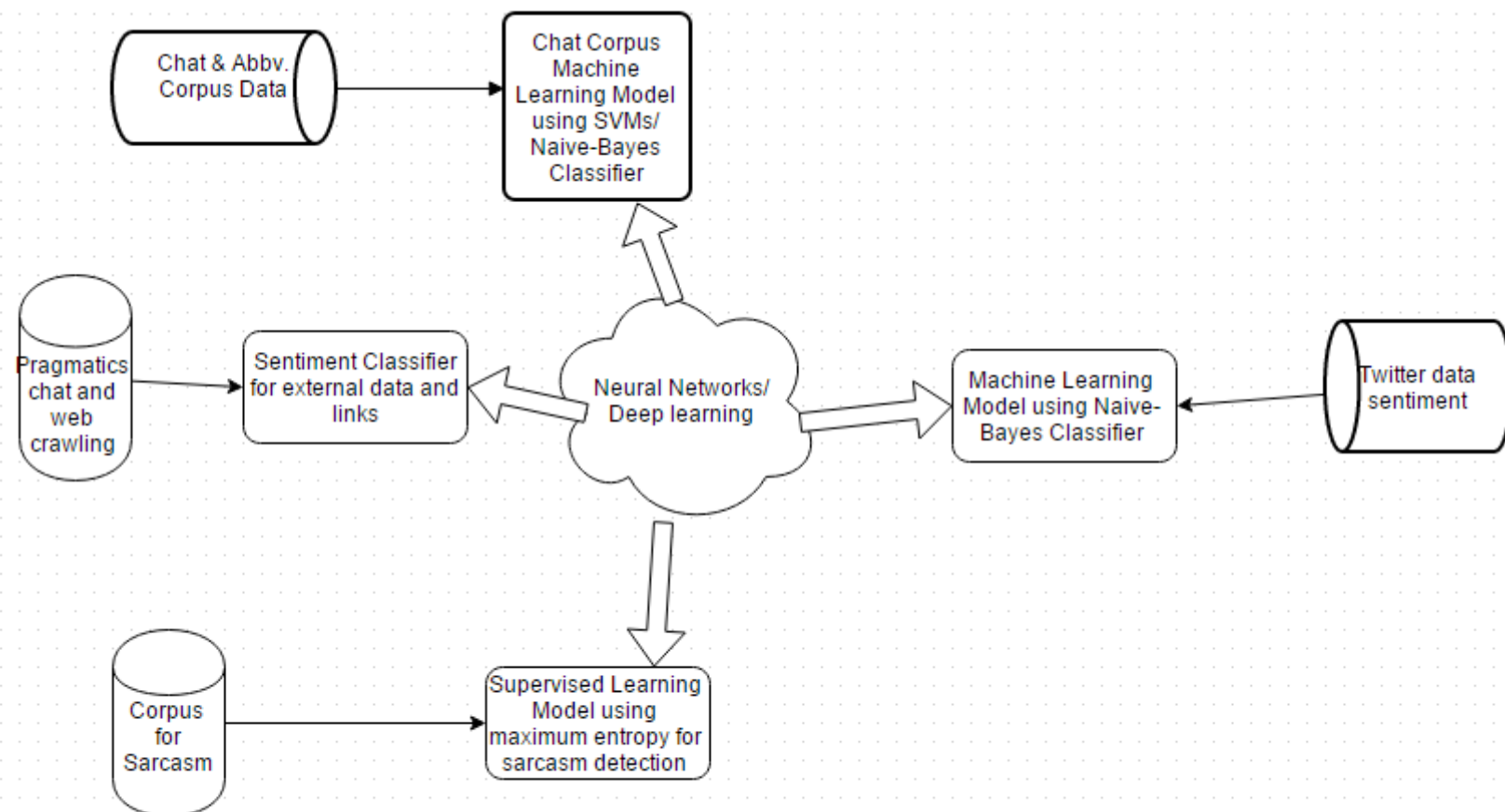
SOFTWARE REQUIREMENTS

- Big data processing engine – ***Apache Spark with pySpark***
- Machine Learning tool kit – ***scikit-learn, mllib, weka***
- Neural Networks – ***pyBrain***
- Querying Language – ***SparkSQL, Hive***
- Data Collection – ***NoSql, MongoDB***
- In-Memory Data Structure Store – ***redis***
- Visualization Tools – ***matplotlib, bokeh, ggplot, pandas***

TECHNOLOGY STACK



Supervised Learning Model Architecture



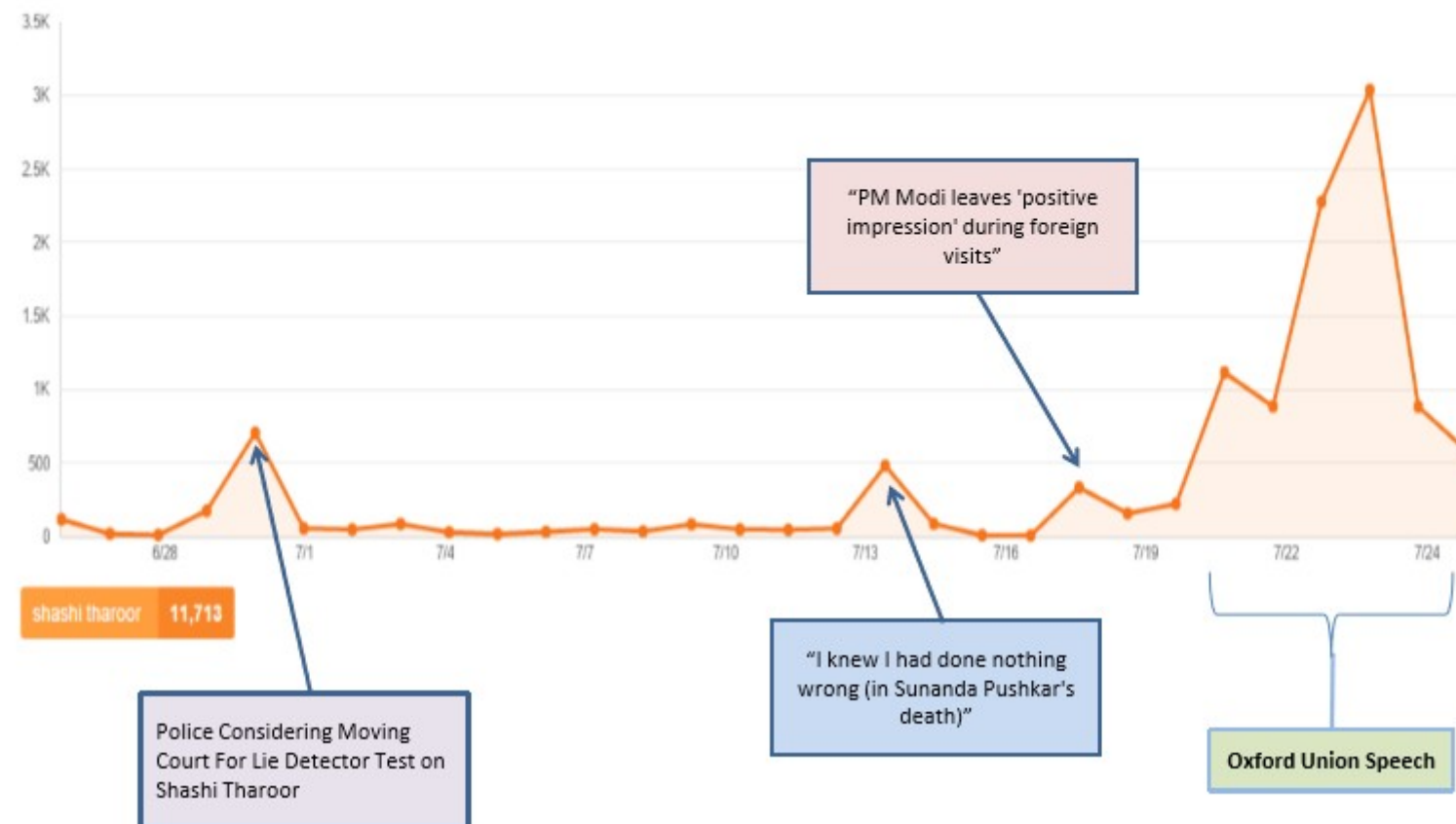
Rough Estimate

PRODUCT	COST
Distributed Cluster Server + 96GB Ram (Optional)	₹1,00,000/-
Cloud Storage	₹2,000/-
Input Training Dataset	₹5,000/-
Data Streams	₹4,500/-

Visualization

Tweets per day: shashi tharoor

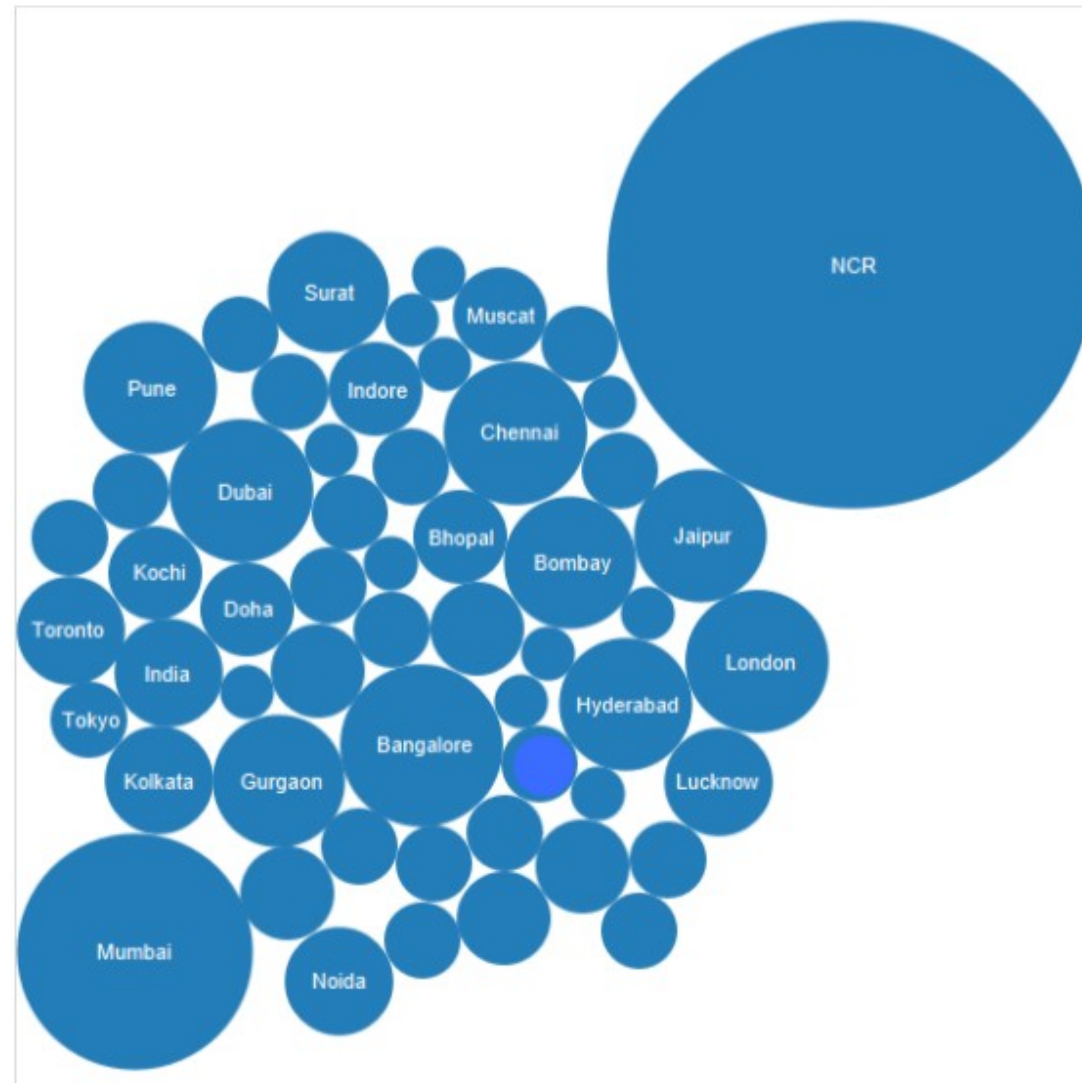
June 26th — July 26th



Visualization

Location Split Up

Specifies the different locations from where tweets about Shashi Tharoor have originated. The bigger the circle, the more the number of tweets from that region

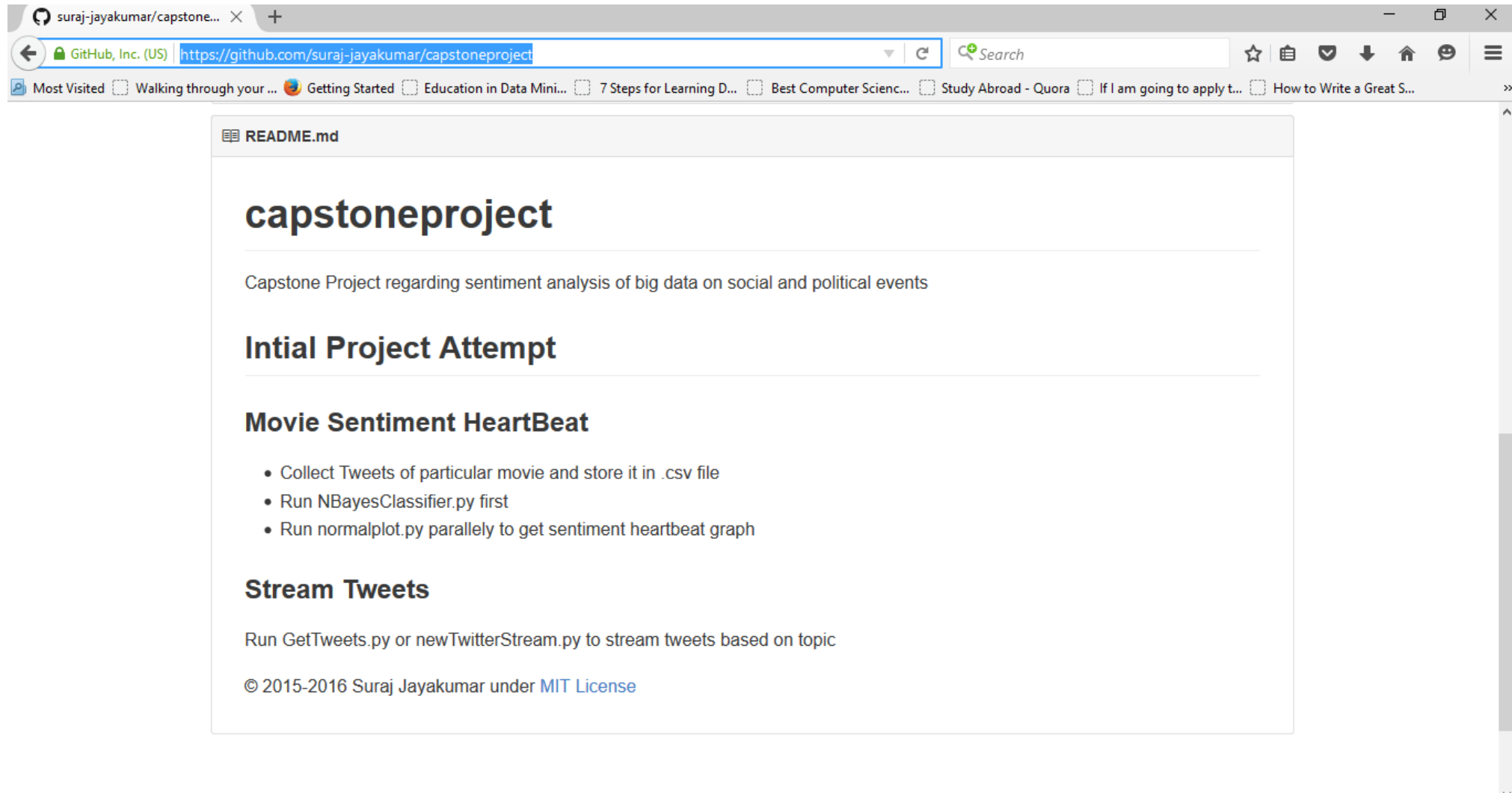


[illegible]

A word cloud visualization of terms related to the Irish famine. The most prominent word is "atrocities". Other significant words include "missed", "fault", "lout", "debt", "famine", "sorry", "emergency", "poor", "wrong", "blame", "loss", "difficulties", "decline", "opposition", "death", "omits", "scolding", "worse", "criticism", "funny", "oppress", "murder", "censure", "bad lie", "utterly", "harm", "issues", "panickers", "corruption", and "issue".

Negative Sentiments

Open-Source



The screenshot shows a web browser window displaying a GitHub repository. The address bar shows the URL `https://github.com/suraj-jayakumar/capstoneproject`. The repository page is titled "capstoneproject" and includes a description: "Capstone Project regarding sentiment analysis of big data on social and political events". Below the description, there is a section titled "Intial Project Attempt" (note the typo) which contains a sub-section "Movie Sentiment HeartBeat". This section lists three bullet points: "Collect Tweets of particular movie and store it in .csv file", "Run NBayesClassifier.py first", and "Run normalplot.py parallely to get sentiment heartbeat graph". Below this, there is a section titled "Stream Tweets" with the instruction "Run GetTweets.py or newTwitterStream.py to stream tweets based on topic". At the bottom, the copyright notice reads "© 2015-2016 Suraj Jayakumar under MIT License". The browser's address bar and tabs are visible at the top, and a sidebar with navigation links is on the right.

suraj-jayakumar/capstone... X +

GitHub, Inc. (US) `https://github.com/suraj-jayakumar/capstoneproject` Search

Most Visited Walking through your ... Getting Started Education in Data Mini... 7 Steps for Learning D... Best Computer Scienc... Study Abroad - Quora If I am going to apply t... How to Write a Great S...

README.md

capstoneproject

Capstone Project regarding sentiment analysis of big data on social and political events

Intial Project Attempt

Movie Sentiment HeartBeat

- Collect Tweets of particular movie and store it in .csv file
- Run NBayesClassifier.py first
- Run normalplot.py parallely to get sentiment heartbeat graph

Stream Tweets

Run GetTweets.py or newTwitterStream.py to stream tweets based on topic

© 2015-2016 Suraj Jayakumar under [MIT License](#)

Feasibility



Video of Existing work on Movie Review



Thank you

