**Final Project: Data Design**

**Name: Jiayuan Hu**

**UCSC ID:** [**jhu125@ucsc.edu**](mailto:jhu125@ucsc.edu)

1. Proposed project: Have people's views towards Donald Trump changed since his inauguration?

The goal of our project is to create a web crawler of tweets about people's view towards Donald Trump before and after his inauguration.To be specific, we want to show what people are thinking about trump at different states of US in the past several months. We will focus on the results before and after his election with a time horizon of months.

1. Data Design and Process Document

(Q1) Why do I choose Rest API over streaming API?

The data is live, while the REST API returns only data that has already been tweeted.

Streaming APIs are less common, but will become more available in the future as more live data is generated and exposed.

Because we need to analyze the data of the past several months, the live streaming data is not ideal for us.

(Q2) Is advanced Rest API better than simple Rest API in finishing our project?

Yes, absolutely, in simple Rest API, a single datafile returns only about 15 tweets, which is not very useful in analyzing the trends over a long period. The advanced Rest API can get specify the number of tweets we want to get.

(Q3) How can I get the access token of Twitter API?

* Create a Twitter Account by logging on to http://[www.twitter.com](http://www.twitter.com)
* Sign in to apps.twitter.com
* Click the “Create New App” button
* Give your application name and description.
* As we do not have a URL yet,so we are going to put theTwitter URL i.e.https://twitter.com in the box that says Website. Do not forget to include https.
* Agree to the agreement and click “Create Twitter Application.”
* Scroll to the bottom and click on ”Create my access token.” The page is refreshed when you click on it and if you scroll you will be able to see the access token.
* You will have a Consumer key, Consumer secret, access token and Access token secret.

(Q4) Why do I choose to scrap data instead of use existed tweets.csv files?

The existed All\_Tweets\_And\_Stats\_2011\_To\_2013.csv are not suitable for us because the tweets contents contains different areas, which does not correspond to our topic. Also, we want to get the datafiles around Donald Trump's inauguration.

(Q5) Why do I choose to output json files and how to parse them ?

JSON data is one of the most commonly used formats for data transfers. It is preferred because it is clean, easy to read, and easy to parse. Additionally, this data format is officially supported by Twitter API.

In order to parse json files, we need to import the json Python library:

json\_data = open('data-text.json').read()

data = json.loads(json\_data)

1. Python and/or all scripts utilized to scrape the data

**Steps of using xpath and lxml to locate elements find newly posted tweets with**

**tag #**[**DonaldTrump**](https://twitter.com/hashtag/DonaldTrump?src=hash)

* Import basic libraries including lxml, os, urllib2
* Use urllib2.urlopen to get contents of the target page.
* page = html.parse(StringIO(page)) // use html in lxml to parse the page content
* By inspecting the page ''<https://twitter.com/search?q=%23DONALDTRUMP>' we can find most tweets are placed in the <p class = "TweetTextSize js-tweet-text tweet-text"> tag.
* therefore, tweets = page.xpath('//p[@class="TweetTextSize js-tweet-text tweet-text"]')
* Iterate the tweets to get all tweets' contents.

**Steps of advanced data pull with Rest API**

* Import json. Put it in a try block to avoid any import errors.
* Import all the necessary methods from twitter library.
* Assign the keys and access tokens to variables.
* twitter= Twitter(auth=oauth) //Initiates connection to Twitter Streaming API
* iterator=twitter.search.tweets(q=‘#trump',,result\_type='recent',lang='en',count=1000) //gets 1000 recent tweets in English where the tweets contain #trump
* print json.dumps(iterator,indent=4) // On the terminal prints the tweets a JSON format with an indent of 4.
* To save the data extracted in a JSON file, use the command : python “Name of the

Python Script File with the .py extension” > “The name of the file you would want to save it as with an extension .json ”

**Steps of advanced data collection from Twitter’s Streaming API**

* Stream Listener helps create a streaming session and listen to messages
* Stream handles TwitterStream.
* Sub class called Listener created to redefine on\_data method.
* with open("scraping\_tweets4.json","w") as data\_file4:

data\_file4.write(json.dumps(data,indent=4))

// write the data into scraping\_tweets4.json file.

* Later the authentication handlers are added.
* Finally, pass the Listener and auth to the Stream and start filtering with a search term. In this case, we are going to look at "trump".

1. Input and output data

github url: <https://github.com/jiayuancc/twitter_crawler>

output data: xpath\_tweets.txt

scraping\_tweets1.json

scraping\_tweets2.json

scraping\_tweets3.json

scraping\_tweets4.json

Some available datafiles:

All\_Tweets\_And\_Stats\_2011\_To\_2013.csv

geolocation\_x\_usa\_x\_filter\_nativeretweets.csv

Some available data links:

* Social Computing Data Repository at ASU

<http://socialcomputing.asu.edu/datasets/Twitter>

* Send secure authorized requests to the Twitter API

<http://dev.twitter.com/oauth>

* All Tweets In PDF or XML Format <http://www.clickonf5.org/5438/download-tweets-pdf-xml-format-local-machine-server/>
* Twitter search with tag #DONALDTRUMP

<https://twitter.com/search?q=%23DONALDTRUMP>'

* Twitter API for keyword 'trump' with json output files

https://api.twitter.com/1.1/search/tweets.json?q=%23donald+trump&src=typd