**Data Mining**

**&**

**Meaningful Visualization**

by

Usman Ahmad Khwaja

Chew (Jason) Kuang Sheng

Mohamed Helal Elsayed

# Abstract

***Data Mining* is a data extraction technique used in many *Artificial Intelligence* applications. The purpose of data mining is to obtain data sets from large volumes of complex data. The data sets can be used to generate meaningful information. Functions can be performed on such data for classification, approximation, or prediction purposes that would otherwise be impossible to run on big data. The data sets obtained however are still large and incomprehensible by human eye. This report will explore some scripting methods by which large informative data can be mined using social and market media applications’ API. Then we should be able to demonstrate how visually readable models can be generated from the extracted data. This can help in human analysis and visualization of large amounts of data.**

### Introduction

Text mining is the application of natural language processing techniques and analytical methods to text data in order to derive relevant information. Text mining is getting a lot attention these last years, due to an exponential increase in digital text data from web pages, google's projects such as google books and google ngram, and social media services such as Twitter. Twitter data constitutes a rich source that can be used for capturing information about any topic imaginable. This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products and services. Once the data is collected, we will also use the Google Charts library for python to display the information in a graph

In this report, we will use Twitter data to compare the popularity of 3 programming languages: Python, Javascript and Ruby. First, we will explain how to connect to Twitter Streaming API and how to get the data. Secondly, we will explain how to structure the data for analysis, and lastly paragraph, we will explain how to filter the data and plot the information collected in a chart.

### System

Operating system Ubuntu 14.04 'Trusty' (x86-64)

Linux Kernel 3.13.0-37-generic

Processor Intel Core i5-3317U CPU @ 1.70GHz x 2

Memory 3.4 GiB

### Data Mining

API stands for Application Programming Interface. It is a tool that makes the interaction with computer programs and web services easy. Many web services provides APIs to developers to interact with their services and to access data in programmatic way. For this report, we will use Twitter Streaming API to download tweets related to 3 keywords: "python", "javascript", and "ruby". We will be using a Python library called *Tweepy* to connect to Twitter Streaming API and downloading the data.

We created a file called twitter\_streaming.py, and wrote the following script to collect data.

#Import the necessary methods from tweepy library

from tweepy.streaming import StreamListener

from tweepy import OAuthHandler

from tweepy import Stream

#Variables that contains the user credentials to access Twitter API

access\_token = "\*\*\*\*\*"

access\_token\_secret = "\*\*\*\*\*"

consumer\_key = "\*\*\*\*\*"

consumer\_secret = "\*\*\*\*\*"

#This is a basic listener that just prints received tweets to stdout.

class StdOutListener(StreamListener):

def on\_data(self, data):

print data

return True

def on\_error(self, status):

print status

if \_\_name\_\_ == '\_\_main\_\_':

#This handles Twitter authetification and the connection to Twitter Streaming API

l = StdOutListener()

auth = OAuthHandler(consumer\_key, consumer\_secret)

auth.set\_access\_token(access\_token, access\_token\_secret)

stream = Stream(auth, l)

#This line filter Twitter Streams to capture data by the keywords: 'python', 'javascript', 'ruby'

stream.filter(track=['python', 'javascript', 'ruby'])

We ran the script for about an hour on a Saturday afternoon. The data was extracted onto a text file so it can be used later in the program. The results yielded were 128 tweets for python, 302 for javascript, and 258 for ruby. From this we can conclude that in that hour, javascript was the most popular language tweeted about. The file size was about 3 mb.

python twitter\_streaming.py > twitter\_data.txt

### Reading the data

The data that we stored twitter\_data.txt is in JSON format. JSON stands for JavaScript Object Notation. This format makes it easy to humans to read the data, and for machines to parse it. Below is an example for one tweet in JSON format. You can see that the tweet contains additional information in addition to the main text which in this example: "Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #hashbrown #hashtag".

{"created\_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id\_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #hashbrown #hashtag","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in\_reply\_to\_status\_id":null,"in\_reply\_to\_status\_id\_str":null,"in\_reply\_to\_user\_id":null,"in\_reply\_to\_user\_id\_str":null,"in\_reply\_to\_screen\_name":null,"user":{"id":2301702187,"id\_str":"2301702187","name":"Toni Barlettano","screen\_name":"itsmetonib","location":"Greater NYC Area","url":"http:\/\/www.tonib.me","description":"So Full of Art | \nToni Barlettano Creative Media + Design","protected":false,"followers\_count":8,"friends\_count":25,"listed\_count":0,"created\_at":"Mon Jan 20 16:49:46 +0000 2014","favourites\_count":6,"utc\_offset":null,"time\_zone":null,"geo\_enabled":false,"verified":false,"statuses\_count":20,"lang":"en","contributors\_enabled":false,"is\_translator":false,"is\_translation\_enabled":false,"profile\_background\_color":"C0DEED","profile\_background\_image\_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile\_background\_image\_url\_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile\_background\_tile":false,"profile\_image\_url":"http:\/\/pbs.twimg.com\/profile\_images\/425313048320958464\/Z2GcderW\_normal.jpeg","profile\_image\_url\_https":"https:\/\/pbs.twimg.com\/profile\_images\/425313048320958464\/Z2GcderW\_normal.jpeg","profile\_link\_color":"0084B4","profile\_sidebar\_border\_color":"C0DEED","profile\_sidebar\_fill\_color":"DDEEF6","profile\_text\_color":"333333","profile\_use\_background\_image":true,"default\_profile":true,"default\_profile\_image":false,"following":null,"follow\_request\_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet\_count":0,"favorite\_count":0,"entities":{"hashtags":[{"text":"thatwasntsohard","indices":[40,56]},{"text":"yesitwas","indices":[57,66]},{"text":"stoptalkingtoyourself","indices":[67,89]},{"text":"hashbrown","indices":[90,100]},{"text":"hashtag","indices":[101,109]}],"symbols":[],"urls":[],"user\_mentions":[]},"favorited":false,"retweeted":false,"filter\_level":"medium","lang":"en"}

For the remaining of this report, we will be using 4 Python libraries json for parsing the data, pandas for data manipulation, matplotlib for creating charts, and re for regular expressions. The json and re libraries are installed by default in Python. We had to install pandas because it was not installed in our machine.