## Team 14

Henry Garcia

Amr Elsisy

John Almarez

## Crawler Report

## Requirements

* Java Requirements
  + json­simple 1.1.1
  + twitter4j 4.0.2
  + JavaSE 1.7
* Python Requirements
  + git clone https://github.com/tweetbot/tweetbot.git
  + #On mac, to install pip, run  
    sudo easy\_install pip  
    #update  
    sudo pip install --upgrade pip  
      
    #On Linux, run  
    sudo apt-get install python-pip  
    sudo apt-get install libxml2-dev libxslt1-dev python-dev  
    #Then update pip using  
    sudo pip install --upgrade pip
  + #install virtual env  
    sudo pip install virtualenv  
      
    #Create a new virtual environment in the current folder.  
    #This will make sure that the dependencies for tweetbot   
    #do not cause a conflict with the dependencies for your  
    #own projects.   
    virtualenv tweetbot\_env  
      
    #To use the virtual environment, first we need to #activate it  
    source tweetbot\_env/bin/activate  
      
    #install dependencies  
    pip install -r requirements.txt  
      
    #run tweetbot with help flag to learn more  
    python tweetbot.py -h  
      
    #to deactivate the virtualenv  
    deactivate

## Design & Implementation

* Architecture
  + Java Crawler
    - The Crawler connects to the twitter streaming API and downloads a number of tweets (specified by the user), writing them to disk in batches of files, 100 by default. The size of the files are determined by the user.
  + Python Crawler
    - The Crawler connects to tweepy and downloads tweets and stores them in files of size 10Mbs by default, the user can change the file size if they want, writing them to disk as soon as they are fetched. By default the files will be stored in the User’s current directory, unless the user specifies a different directory path
* Crawling/data collection strategy
  + Crawler
    - The crawler first checks the directory for a file containing saved tweet ids. If the file exists the hashmap will be loaded with the tweet ids, otherwise the hashmap will start empty.
    - The tweet crawler first creates a filter in order to grab tweets that are in English and located in the US. After creating the filter it establishes a connection with twitter’s stream, using account and consumer tokens provided through https://dev.twitter.com/
* After establishing a connection with the twitter stream the system will start to grab tweets. Only unique tweets are grabbed, duplicates are ignored.
* When saving the tweets the crawler will check if the file is larger than that specified by the user. The crawler will continue to increment the tweet file until it finds a file that is no longer larger than the file size provided by the user (useful if the user has already crawled tweets and doesn't want them to be placed in already crawled files).

## Evaluation

* statistics on amount and properties of data, time it took you to collect it,…
* We stored ~10GB of data
  + Each file is ~100MB (which is approximately ~20,000 tweets)
* Tweets were stored as json objects, with one tweet per line
* Time to collect 10GB of data:
  + Around 24 hours per 2GB of tweets, due to twitter restriction
  + In total ~5 days

## Contribution

Henry Garcia

* Crawled twitter using twitter4j <Java Crawler>
  + Added robustness for command line inputs
  + Added hashmap to keep track of duplicate tweets
  + Added hashmap to store tweets in memory and then save in bulk
  + Added json fields to store a tweet
  + Added functionality to import an already found hashmap for duplicate tweets
  + Added functionality in order to stop crawler
  + User decides size of each file
  + User decides how often they save tweets
  + Decides how many tweets to crawl
  + User inputs directory to store tweets

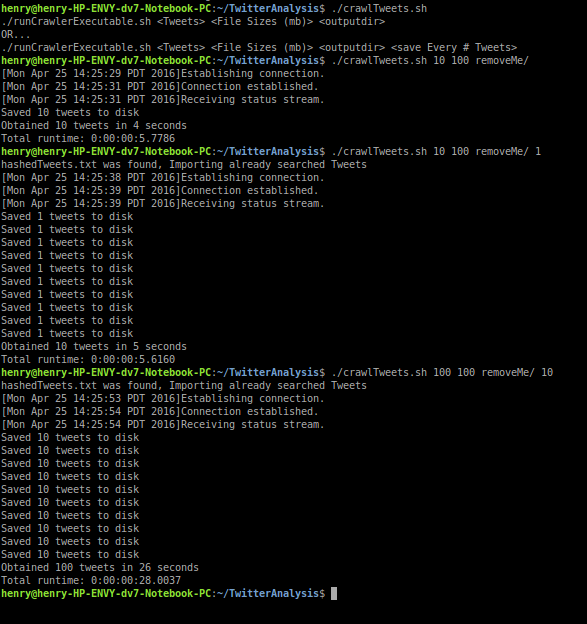
Amr Elsisy

* Crawled twitter using tweepy <Python Crawler>
  + Added robustness for command line inputs
  + Added json fields to store a tweet
  + Added functionality in order to stop crawler
  + User decides size of each file
  + Decides how many tweets to crawl
  + User inputs directory to store tweets

John Almaraz

* Did the report
* Helped us determine which tweet attributes to save for future use
* Proposed the idea of two different crawler, to match user preference, and increase the rate of crawling

Java Screenshots:



Python Screenshots:



