**SMA Assignment #2 (group work, due 2/15 by 11:59 p.m.)**

The learning objectives for this assignment are to learn ways to

* 1. Detect social “influencers” using analytics
  2. Quantify the financial value of influence

(iii) Identify and leverage influencers

The assignment has two parts: I and II. In Part 1, you will use data on social influence to find out the important predictors of influence, and to quantify the financial value of influence. In Part II, you will collect tweets, and use the predictors from Part I to identify 100 top influencers in a domain of your choice.

**Part I: Find predictors of influence**

The dataset for Part I can be http://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network

Each observation describes two individuals, A and B. There are 11 variables for each person based on Twitter activity, e.g., number of followers, retweets, network characteristics, etc. Each observation shows whether A > B (Choice = “1”) or B > A (Choice = “0”).

**Create an analytic model for pairs of individuals to classify who is more influential**

* + Check if you should use all variables
  + Perhaps a transformation of A/B or A-B variables will be better than using A and B variables separately. This may also be easier to interpret.
  + Report the confusion matrix of your final model (provide screenshot)

From your model, which factors are best predictors of influence? (Provide screenshots). Are there any surprises here? How can a business use your model/results?

**Calculate the *financial value* of your model**

A retailer wants influencers to tweet its promotion for a product. If a non-influencer tweets, there is no benefit to the retailer. If an influencer tweets once, there is a 0.10% chance that his/her followers will buy one unit of a product. Assume the retailer has a profit margin of $50 per unit, and that one customer can buy only one unit. If an influencer tweets twice, the overall buying probability will be 0.15%. Without analytics, the retailer offers $5 to each person (A and B) to tweet once. With analytics, the retailer offers $10 to those identified as influencers by the model to send two tweets each. If the model classifies an individual as a non-influencer, s/he is not selected/paid by the retailer to tweet.

What is the lift in expected net profit from using your analytic model (versus not using analytics)? Show all calculations. What is the lift in net profit from using a perfect analytic model (versus not using analytics)?

**Assumption: Each user appears only once in the data**

|  |  |  |
| --- | --- | --- |
| **A** | **B** | **A>B?** |
| **John** | **Ted** | **Yes** |
| **Sue** | **Ron** | **Yes** |
| **Fred** | **Sandy** | **No (Sandy > Fred)** |
| **Alex** | **Moe** | **No (Mo > Alex)** |

The Influencers in the above table are John, Sue, Sandy & Moe, but no ordered ranking is possible (or needed in this case).

**Part II: Finding influencers from Twitter**

Collect about 5000 tweets on any topic (e.g., politics, sports, current events, etc.). In addition to the tweet itself, the Twitter API provides a large quantity of information about the tweet as well as the author. Fetch all of this additional information along with the tweets.

Many of you may already know how to get authorization from Twitter, but for others, details are provided at the end of this document.

**Using the results from Part I, create a list of top 100 influencers from the tweets. Here is one way to do it.** Suppose three factors – retweets, listed count and # followers turned out to be the most important indicators of influence in assignment 1. Now create a score for each author who tweeted as:

Score = w1\*retweets + w2\*listed\_count + w3\*#followers where w1+w2+w3 =1

Choose the weights (it is subjective) such that bigger weights are given to factors that were more significant (as judged by coefficients and *p* values in Part I). You should normalize or standardize your data before creating the overall scores.

Note that some factors that may have turned out as important in Part I may not be available in the tweets (e.g., network characteristics).

**Write a script in Python (or R) that parses through the tweets and does the following: For each tweet:**

Any **retweet** (RT), **mention** or **reply** should result in an arrow from the person retweeting to the person retweeted, mentioned or replied to. However, you don’t need to draw the actual arrows and the network. Instead, create a three-column .CSV file as follows: If @XYZ retweets a tweet by @ABC, then put the following in the .CSV file:

Column 1 Column 2 Column 3 (type of content)

@ABC @ABC Tweet

@XYZ @ABC RT

Most social network analysis tools (e.g., NodeXL, which you don’t need for this assignment) will take the first two two columns and draw arrows from the user in the left column to the one in the right. Note that in most cases the set of tweets you may fetch will not have the original tweet that is being retweeted by someone else. E.g., a tweet in your data (tweeted by, say, @XYZ) may be: “RT @ABC Working on my social media assignment.”

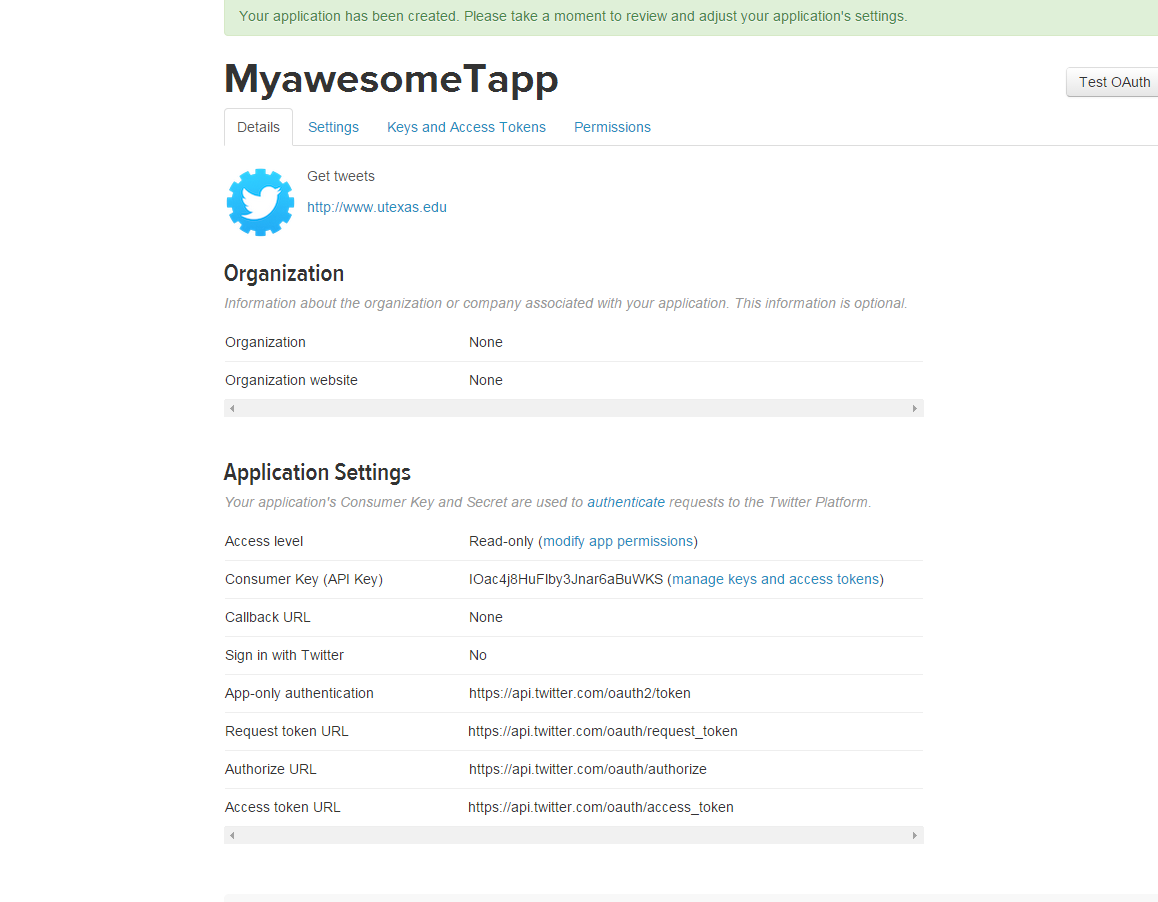
It is quite possible that you will not have the original tweet by @ABC in your data. Still an arrow should go from @XYZ to @ABC. Therefore, even if you have fetched 5000 tweets by 5000 unique users, your network may consist of a much larger set of users.

**Submit the following to Canvas:**

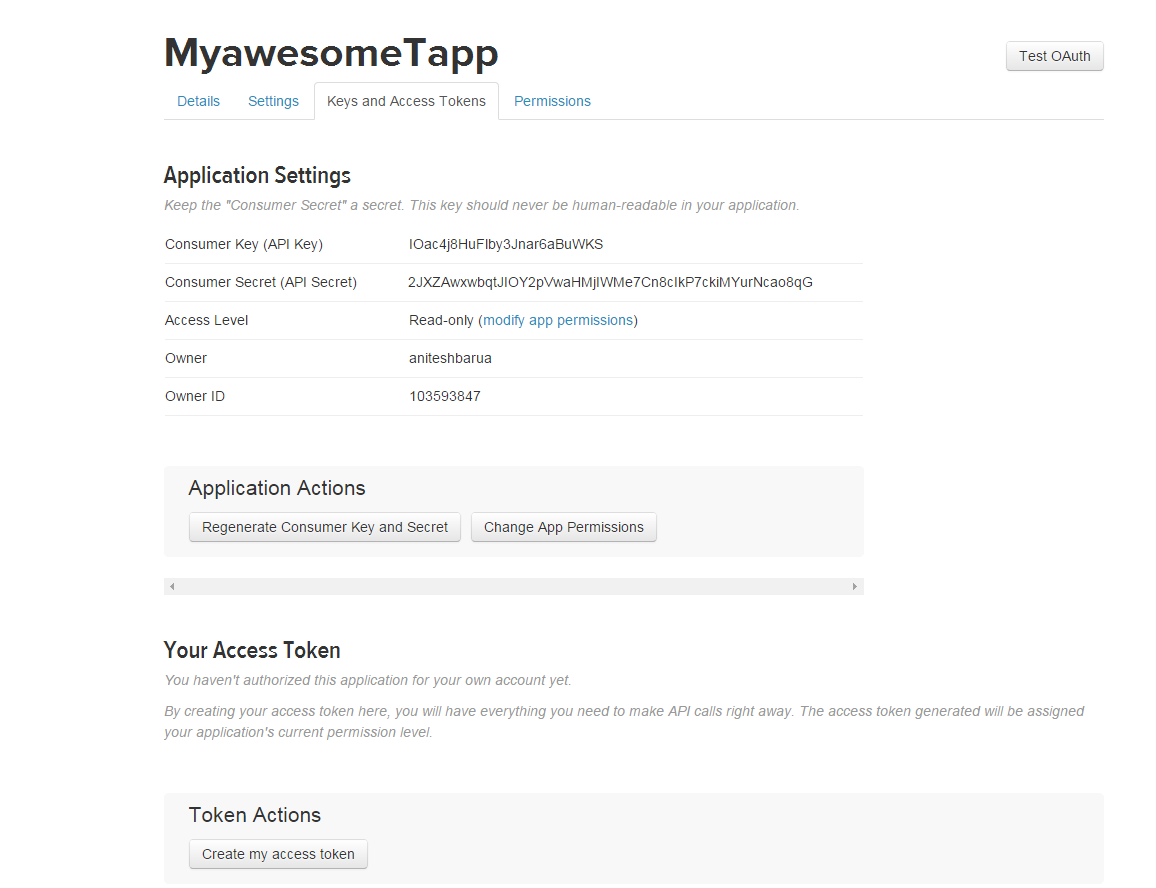
1. Python script (if you used the utility in R, just mention so in the answer document)
2. A .pdf file with answer to Parts I and II.
3. A .CSV file with three columns as described in Part II

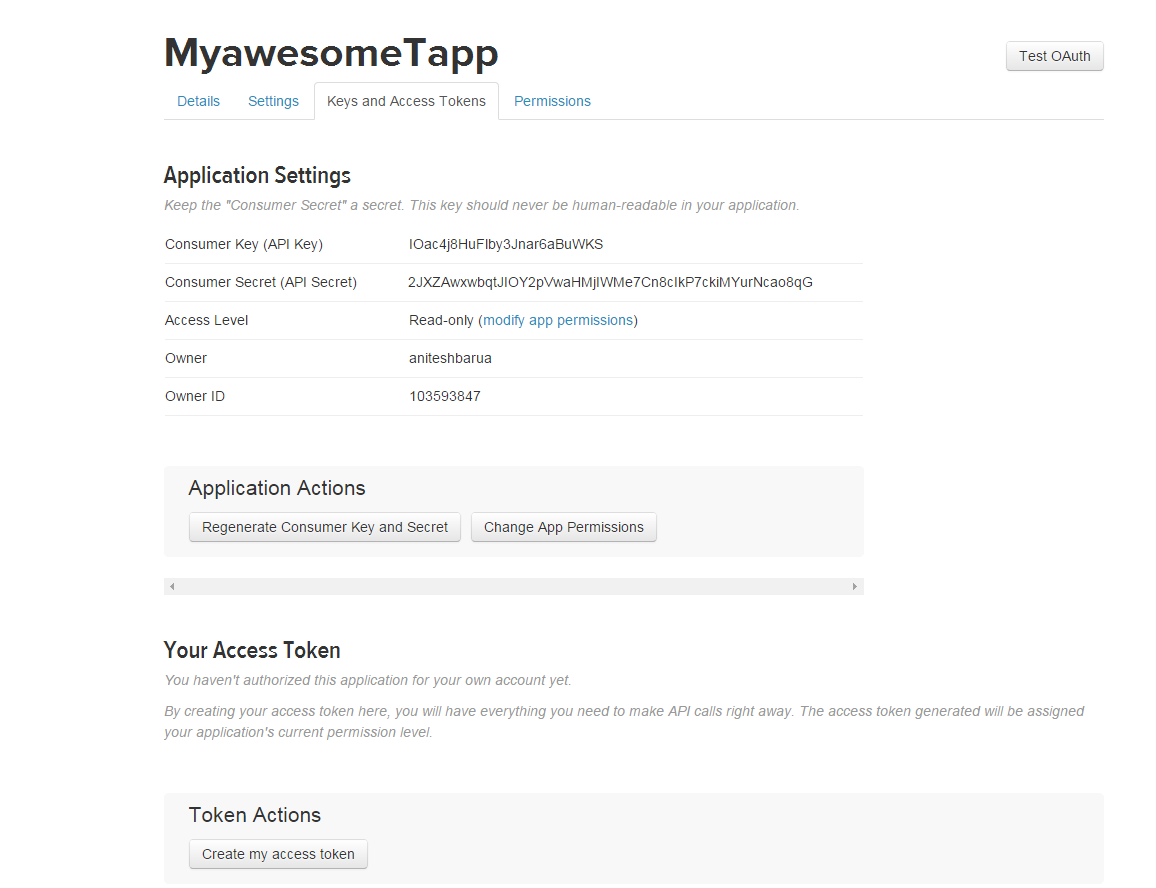
**Procedure to obtain authorization from Twitter (Ignore this if you are familiar with these steps)**

To access tweets through the Twitter API, first you need to get authorization from Twitter and then use the R utility or a custom Python script to collect tweets on any topic of your choice. To get authorization from Twitter, create a Twitter account on Twitter.com. Even if you have one already, I suggest creating a second account with a different email address. Once your email address is confirmed by Twitter, go to <https://dev.twitter.com/docs> and click on “manage my apps”. Create a new app by giving it a distinctive name, a short description and a url (this can be any other url like <http://www.utexas.edu>, which does not belong to the twitter domain). Once you create the app, you should see something like this:

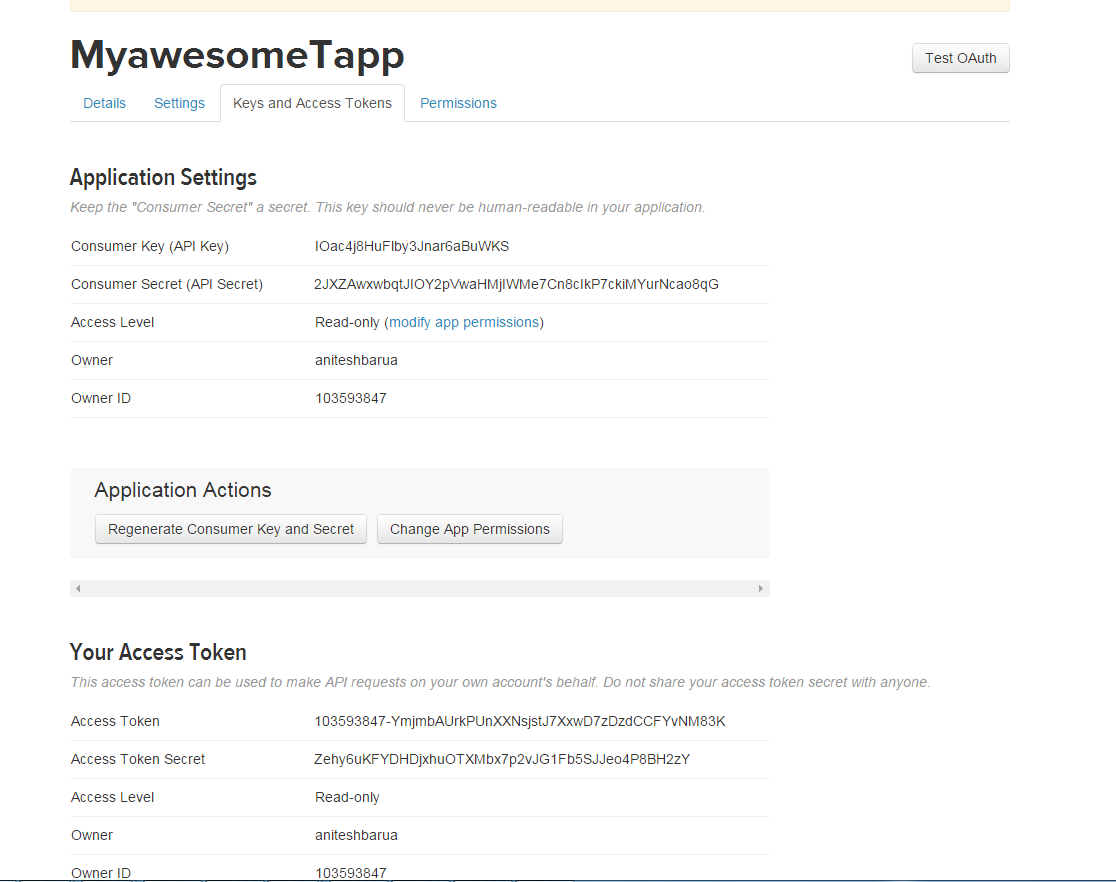


Click on **manage keys and access tokens**, and you should see a page like this:





So now you have the API key and API secret. But you also need access tokens. To do so, click on the button **Create my access token**.



Now you have the Access Token as well as the Access Token Secret. You will copy these 4 keys, tokens and secrets and paste in the right places in the Python script or the R utility that we will use to extract tweets.