DS 785 Data Science Capstone

Capstone Activities Report

Name: Spencer Swartz

Project Title: Social Media Content Optimization and Data Driven Brand Advocate Identification

Activity Report #1

Agenda/goals:

At the completion of this activity period to remain on track for the project it is expected that I was able to connect to the Twitter API, and pull JSON data related to the Brewers Association related accounts (BrewersAssoc, HomebrewAssoc, and craftbeerdotcom). Additionally, the data is needed in a structured and relational form.

Contacts Made/Method of Contact:

I have contacted Bart Watson (BA) via email with a similar update notifying him of the progress made in the data collection period on 2/17/18.

Resources and Investigation Methods:

In order to complete this portion of the project, a few resources are needed. The first would be requesting access to the Twitter API, this can be done at https://apps.twitter.com. Once access has been granted, access keys and tokens are generated. These are used in Python with the Tweepy package. The Tweepy package has been created to make request to the Twitter API simplistic and straight forward in a Python development stack.

Progress:

(Project progress relative to the entire project)

The projects progress is on track at the moment, I was successful in gaining access to the twitter API and pulling the json data. For this first part of the project the data needed is most specific to BA tweets from its various handles, pulling this data is fairly straight forward as all that is needed to make the request is the handle itself. Bellow you can se an example of this request using Tweepy:

```
1.      public_tweets = api.user_timeline(id = 'BrewersAssoc')
```

This snippet will grab roughly the last 3200 tweets and retweets from the user BrewersAssoc. The data returned looks like this:

```
1.      {
2.          'contributors': None,
3.          'coordinates': None,
4.          'created_at': 'Sat Feb 17 14:30:32 +0000 2018',
5.          'entities': {
6.              'hashtags': [{
7.                  'indices': [84, 92],
8.                  'text': 'brewery'
```

```
    x
            'symbols': [],
            'urls': [{
                'display_url': 'twitter.com/i/web/status/9…',
                'expanded_url': 'https://twitter.com/i/web/status/964869680107225089',
                'indices': [109, 132],
                'url': 'https://t.co/cfoMwoMGLg'
            }],
            'user_mentions': []
        },
        …
    }
```

This is a JSON script with many attributes, the problem is that some of the values are nested types (the urls line is an example) and do not lend them self well to a structured data source. Luckily though most of these nested fields can be converted into a separated data file with the tweet id as a key between files. Additionally some of the fields are repetitive and have no use in the project, we can remove these and create the sub tables that are needed.

The last set of data that was pulled for this part of the project was the user profiles of the users who retweeted a given BA tweet. This is done by looking up each individual tweet id and running a separate script also part of Tweepy:

```
req = api.retweets(id = 1023302430304)
```

This script works well, but is limited by the rate limits that are a part of Twitter API, because of this and given the shear amount of tweets that were collected (~8000) this portion of data collection took nearly a 24 hour period to complete.

With all of this data collected it was possible to create 7 related tables. These include:

1. tweets (contains text, time, # retweets, #likes, etc.)
2. retweeted by (contains the users who retweeted at least one BA's tweets)
3. user mentions (users mentioned in each tweet, and location)
4. symbols (any symbols used in each tweet, and location)
5. urls (urls used in each tweet)
6. hashtages (hastages of each tweet, and location)
7. media (links to the media used in each tweet)

Achievements:

I was successful in obtaining and formatting tweet data for roughly 8000 tweets related to BA as described above.

Questions:

I currently have no questions for the instructor at this time, moving forward I need to be cognizant of the time that may be required to research practices to identify brand advocates and possibly the time that will be needed to collect additional data. This is especially important as some request to the API take more time than others given the rate limits imposed.

Next step:

The next 2 weeks will be focused visually modeling the data within Tableau. Specifically I will be looking to answer the following question:

- What type(s) of content engages BA audiences the most (links, photos, videos, GIFs, etc.)?
- Are there words that are more appealing, certain images (color palette themes, lifestyle, technical, etc.)?
- Do specific days of the week, times, frequency, etc. matter?
- Do we see overlap from the various BA brands (Brewers Association, American Homebrewers Association and CraftBeer.com)?
- What are these brands similarities and differences?

With the data collected in this reporting period it is very likely that the above questions will have some answers. To be able to create these visuals various data joins will be need between the tables created during this reporting period. Most of this will be done within Tableau as the data ingestion portion of the software is quite useful. If there are any needs to manipulate the data further than Tableau is privy to then I will utilize python/pandas to do so. It is outlined within the project timeline that the analysis and development of a dashboard will continue through to the week of March 5, if I believe I am ahead of schedule additional time will be used for researching academic articles of brand advocates, the second phase of this project.