Xiang Li
MIS 3640
Assignment 3

Analyzing tweets of Manchester United on 11/05/2017

I am a fan of Manchester United. In this assignment, I want to use python to analyze the tweets containing with my team's name to get a sense of the audience's mood towards the outcome. In those data I collected from Twitter, I find out that even Manchester United lost a game with Chelsea today, fans' moods of MU are still be very positive.

Implementation

There are two major ways that I did for this case. First of all, I picked up Twitter Apps as my major tool of analyzing data. In this part of code, I will get 200 results about "Manchester United" tweets on Twitter.

```
def process_data():
    # Replace the following strings with your own keys and secrets
    TOKEN = '2176923597-qdCRmTUOtKpxJlXCgQtOxXCsxogwISEsrL95SrJ'
    TOKEN_SECRET = 'TRTY85MwEYyf2Hi1KocCmul4tiCObrRodOeQJSn6hNB40'
    CONSUMER_KEY = 'WTjWq39CmXN88VPX5OJag277k'
    CONSUMER_SECRET = 'WfpiLlItfqZIynTTLVqYeigs1ExhlpqpY4hBZFvKZhB0IpjOwR'

    t = Twython(CONSUMER_KEY, CONSUMER_SECRET,
    TOKEN, TOKEN_SECRET)

    data = t.search(q="Manchester united", count=200)
    return data
```

Xiang Li
MIS 3640
Assignment 3

After that, I want to summarize those 200 tweets in to some short key words. I find out that, one of the most common word in the result is “rt,” which means “re-tweet.” So I start thinking that in this common-word list, it should have a lot of stop words, and names, etc. therefore, I want to use nltk to do run the sentiment analysis, and try to remove those “mean-less” words. However, before I run the nltk, I find out that in the “most-common-word” list, it is not hard to identify the most common significant wads. Therefore, I choose to pick the top 6 common words in those tweets to run the nltk package, which gives me the results of “Chelsea”, “Manchester,” “united,” “europaleague,” “mourinho,” and “joes.”

```
def most_common(hist):
    t = []
    for key, value in hist.items():
        t.append((value, key))

    t.sort()
    t.reverse()
    return t


def print_most_common(hist, num=30):
    print('Most common words are:')
    for freq, word in t[:num]:
        print(word, '\t', freq)


def main():
    data = process_data()
    hist = process_file(data)
    print(hist)

    t = most_common(hist)
    print('Most common words are:')
    for freq, word in t[0:30]:
        print(word, '\t', freq)
```

Results:

After completing the sentiment analysis for tweets about Manchester United, I got an output: {'neg': 0.0, 'pos': 0.569, 'neu': 0.431, 'compound': 0.9996}; It shows that after the defeat of the game with

Chelsea today, it is still the most popular topic in Twitter, and on the other hand, as the ex-manager of Chelsea, and the current manager of Manchester United, Joes Moruinho is another popular topic in Twitter. Even we lost this game today, people are still willing use positive words rather than use bad words on the team.

Reflection:

I think the time of digging data is very important. When I was testing my code, I got different groups of “most-common-word” from Twitter, and nearly 1/3 of them were quite different. So in order to be more accurate in analyzing, I can do this text mining repeatedly to get the real “common-word.” Also, people love using emoji to comment or retweet now, but I cannot generalize those emoji face into my list; therefore, for some tweets, which only commented with emoticons, the nltk cannot understand what that means.