Derek Kedigh

FinalBigData

May 12, 2016

https://github.com/drkgf8/finalProjectBigData.git

Descriptive Statistics

How many tweets are in the collection? 12629072

When do they start? 2014-08-05 09:50:03

When do they end? 2016-04-21 16:37:21

What is the trend for tweet volume?

2014: 1601006

2015: 8487133

2016: 2540933

-It is going up on average

If you look at the most common words over the lifetime of the search, do you notice any particular trends associated with those words?

- They tend to spike during certain periods and then they go down again to normal levels

What external events might correspond with the differences in the trends of most common words?

- Major societal and cultural events

What hashtags show up as most prominent in each month of the lifecycle?

- RIP is very common
- Lmao is very common as well

Which twitter users are the most mentioned?

homo_viator

VeraVanHorne

How frequently is each user mentioned during each month of the lifecycle?

What is the relationship between the volume of tweets you selected and the volume of tweets for other collections in the data set?

- The volume of tweets I selected were less on average then the volume of tweets from other collections in the data set

Answering Research Questions

1) I prepared the dataset for analysis
2) My github link is https://github.com/drkgf8/finalProjectBigData.git
3) I examined the dataset with the command

    sql = 'select * from tweet where job_id=4223 limit 1'

    and I got the following output

```
526385409501061120
4223
2014-10-26 14:50:36
#ritzy cinema workers earn less than the price of popcorn, tel
l @cineworld and @picturehouses to pay the #livingwage: http:/
/t.co/pRDVpMYa7Z
134610989
134610989
soundscloser
Bruce
94
206
20
617
London
None
None
None
<a href="https://dev.twitter.com/docs/tfw" rel="nofollow">Twit
ter for Websites</a>
```

```
None
None
None
en
0
```

4) The Machine Learning directory is called ‘ML’. It has my linear regression plot and my code

5) I wanted to understand the relationship between mentions and followers. I’m not on twitter but I would think there would be a relationship between these two variables. Accordingly I attempted to construct a query that would parse this relationship and output a list joining screen_name from mentions with from_user_followers on tweets. The query took too long and so I had to get creative. Accordingly I first went randomly through the list of screen_names in the table mentioned and retrieved their aggregate count with the command

```
#sql = 'select count(screen_name) from mention where screen_name =
"NinaByzantina"'
```

I likewise got the corresponding followers for each screen_name with the command

```
#sql = 'select from_user_followers, screen_name from tweet join mention on
mention.job_id = tweet.job_id where mention.job_id=4223 group by screen_name
```

I did this randomly for 40 screen_names. I wanted to have enough data for a train and a test set for my linear regression.

```
x_train = np.array([224,152,40,181,2,316,581,2154,720,97,11,692,1503,621,566,385,72,410,199,452])

y_train =
np.array([6455,3525,224,1019,6,11972,15906,40788,20330,2407,84,8635,27599,19049,14188,12711,51
95,14766,

        8010, 19210])

x_test = np.array([50,109,154,78,376, 222, 203, 186, 433, 301, 193, 333, 102, 181, 456, 284, 291, 167,
317, 99])

y_test = np.array([1411,3013,5005,2186,10174,7773,5991, 6291,18044,13864,4001,12310,6322,10931,
16692,5051,

        8612, 4754, 10,321])
```

I fit the model to the training data

```
regr.fit(x_train, y_train)
```

And did some statistics

```
# The coefficients
print('Coefficients: \n', regr.coef_)
print("Residual sum of squares: %.2f"
    % np.mean((regr.predict(x_test) - y_test) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x_test, y_test))
```

And got the following output

```
Coefficients:
 [ 18.40919505]
Residual sum of squares: 13672270.60
Variance score: 0.46
```

I expected there to be a pretty good relationship between the number of mentions and the number of followers. The relationship wasn't as strong as I expected. The sum of squares and the variance are too high given the small amount of data that I input in to the model.

I retried the linear regression method with different data constructed from the query

```
Select user_followers from tweet where from_user_fullname = 'screen_name'
```

And I was surprised to find that from_user_fullname isn't the same as the screen_name. So I did a little investigating to find the relationship between the major tables. I discovered that the relationship between the tweet table and the mention table is one to many. I also discovered that tweet_id_str in the tweet table is identical to the tweet_id in the mention table- which surprised me given their different names.

The plot looks fairly decent for the linear regression. With an obvious linear relationship between the variables.