University Twitterers Analysis

By Duo Zhou

Executive Summary, Methodology and Source Data Overview

Executive Summary

- *Tweets From 4 Universities (UChicago, Cornell, UMD and UVM) are chosen for this Analysis*
- *The most prolific Twitterers are not the same as the most influential Twitterers for all the Universities*
- *Tweets on Universities located in big cities tend to have higher percentage local Twitterers than tweets on Universities located in small towns.*
- *UChicago Twitterers are more active on twitter that other Universities Twitterers.*
- *Big University news tend to cause peak in tweets volume and there are more tweets valleys in 2019 and 2020.*
- *The collection gap dates are continuous from 2018-01-06 to 2018-02-12 and from 2018-04-03 to 2018-4-21.*
- *Big city universities have a much higher percentage of near-duplicated tweets than small town universities.*

Methodology and Data Overview

- *The source data contained over 300M+ records and tweets about the 4 chosen universities are selected for the analysis by matching key words against tweet texts (Total 1.6M+tweets after filtering).*
- *Descriptive analysis are conducted for all universities (prolific/influential users, user locations, timeline of tweets, UChicago vs Other Univ. Twitterers' statuses)*
- *Similarity analysis of tweets using MinHashLSH. Reviewed results based on various thresholds and 0.01 Jaccard Distance was chosen as the threshold.*

Content Filtering, Feature Selection and EDA

- *A set of key words are selected for each University to obtain records that are related chosen universities. Those key words sets are built based on limitations of computational resource and google search. (Total 1.6M+ records are selected)*

- *Only the features that are useful and relevant for our analysis are chosen based on the final project requirements. This step is applied to before EDA to save space and memory during analysis.*

- *A statistical summary of all chosen columns was generated to spot poorly populated feature. All relevant columns contain enough records and no feature dropping was needed.*

Chosen Features

```
df.printSchema()
```

```
root
 |-- created_at: string (nullable = true)
 |-- favorite_count: long (nullable = true)
 |-- id: long (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- text: string (nullable = true)
 |-- user_created_at: string (nullable = true)
 |-- user_followers_count: long (nullable = true)
 |-- user_friends_count: long (nullable = true)
 |-- user_id: long (nullable = true)
 |-- user_listed_count: long (nullable = true)
 |-- user_location: string (nullable = true)
 |-- user_statuses_count: long (nullable = true)
 |-- user_verified: boolean (nullable = true)
```

Statistical Summary

```
df.summary()
```

summary	created_at	favorite_count	id	retweet_count	text	user_created_at	user_followers_count	user_friends_count	
count	1618667	1618667	1618667	1130058	1618667	1618667	1618667	1618667	
mean	null	0.0	1.132129981807879...	2034.2602804457824	null	null	9510.111565257092	1889.3103393100619	2.
stddev	null	0.0	1.394872691894534...	6936.935552834207	null	null	292109.17424150254	9465.449179322744	4.
min	Fri Apr 03 00:00:...	0	877892761638318080	1	I @ Cornell Unive...	Fri Apr 01 00:01:...	0	0	
25%	null	0	1007622557711896576	5	null	null	154	206	
50%	null	0	1140417883446697984	37	null	null	505	527	
75%	null	0	1251733293105647616	575	null	null	1701	1375	78
max	Wed Sep 30 23:59:...	0	1371223307329081358	62301	PGY-4 PROFIL...	Wed Sep 30 23:58:...	67031966	1615588	137

Prolific/Influential Tweeters and Tweeters' Behavior

The twitterers that tweet the most about a given university are not the most re-tweeted

The prolific and influential twitterers tweet more about other-topics than university related topics

Top Twitterers By Tweets Volume

University	user_id	message_volume
UChicago	131144285	3444
Cornell University	1141158911946309632	3073
UMD	47364511	5358
UVM	469853729	388

Top Twitterers By Retweets Volume

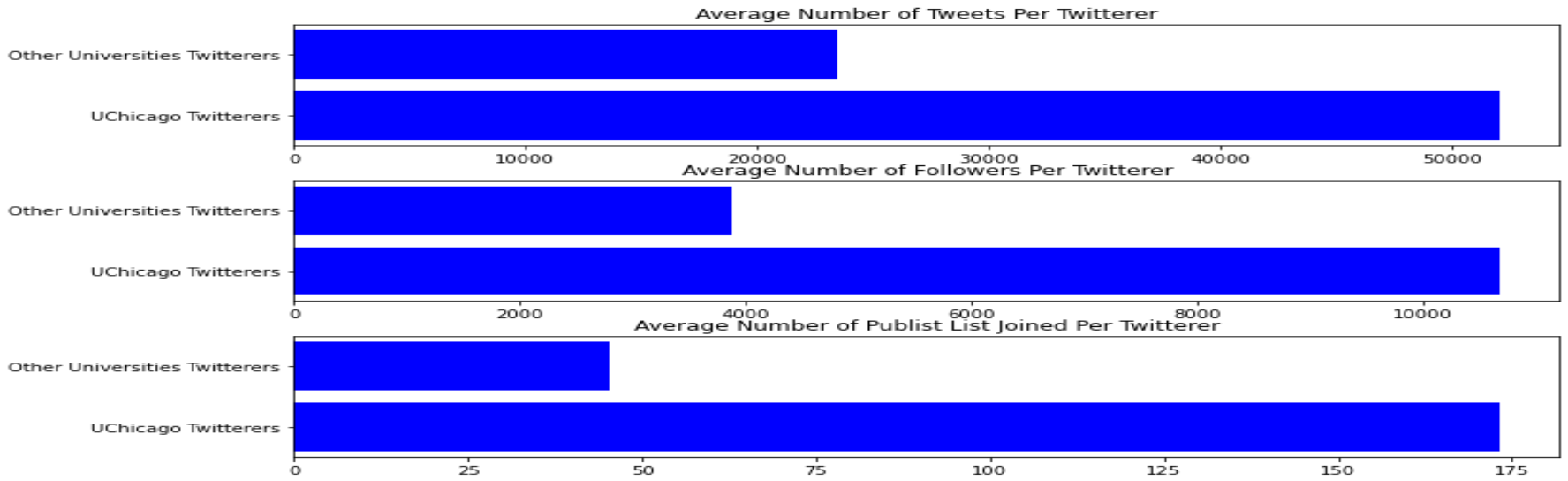
University	user_id	message_retweets
UChicago	1329526295709970433	108780
Cornell	1063894687709200384	143691
UMD	43638469	304101
UVM	555811765	73432

Topics	Message_Volume
Universities	33
Other_Topics	12270

Uchicago Twitterers tend to be more atvie on twitter than other Universities Twitterers.

On average UChicago twiterers:

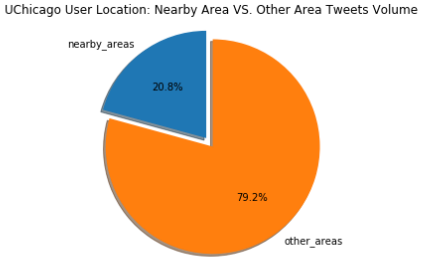
1. *Tweet more*
2. *Have more followers*
3. *Join more public lists*

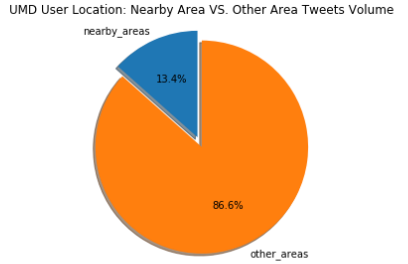


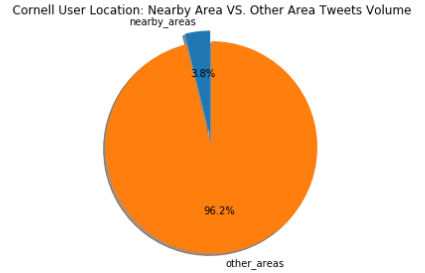
Twittters' Location and Tweets Content Similarity Analysis

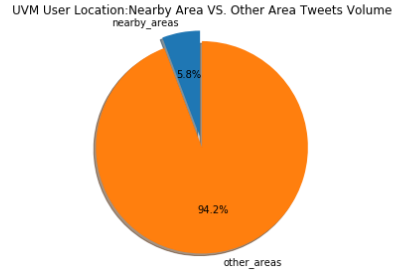
- *Tweets on Universities located in big cities tend to have higher percentage local Twitterers than tweets on Universities located in small cities.*
- *UChicago and UMD are located in big metropolitan areas whereas Cornell and UVM are located in small towns.*
- *The top Twitterer location for each university is the same as the university location for all 4 universities.*

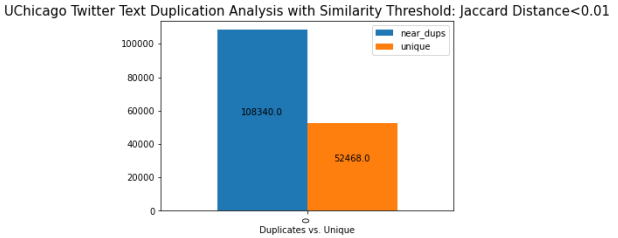
- Big City Universities tend to have higher percentage of near duplicates tweets indicating that big universities tweets get more retweets.
- Using Jaccard Distance = 0.01 as the threshold, the ranking on the percentage of near duplicated tweets are:
  1. UChicago-67.4%
  2. UMD- 61.2%
  3. Cornell-54.7%
  4. UVM-45.8%

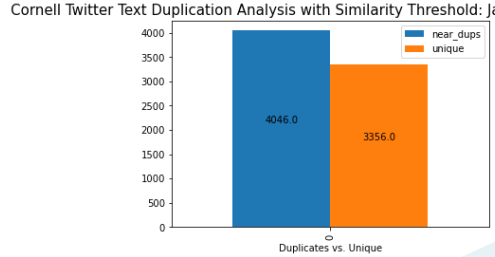


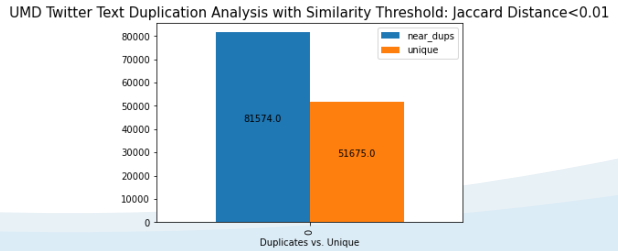


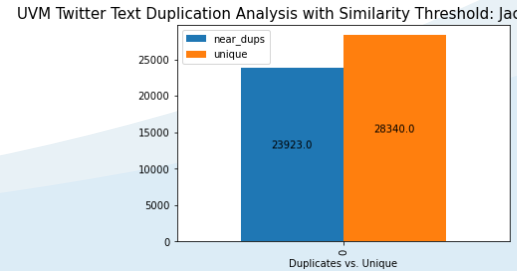








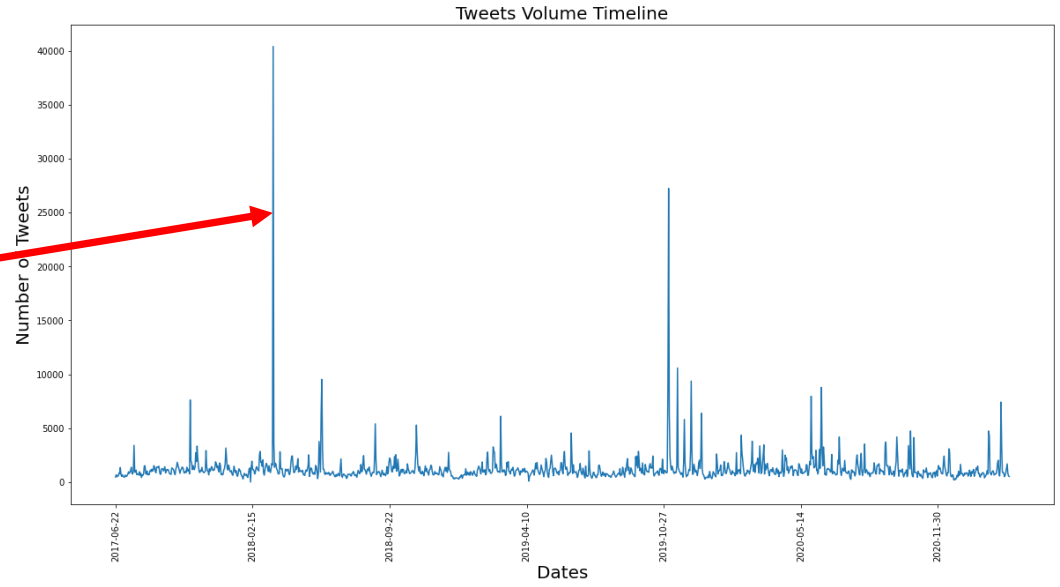




Timeline of Tweets

Big university related news tend to cause peak in tweets volume.

```
'RT @NorrisAG: UChicago dropping
SAT/ACT scores from admission
requirements is monumental.
```



The collection gap dates are continuous:

1. *From 2018-01-06 to 2018-02-12*
2. *From 2018-04-03 to 2018-4-21.*

```
198   2018-01-06
199   2018-01-07
200   2018-01-08
201   2018-01-09
202   2018-01-10
203   2018-01-11
204   2018-01-12
205   2018-01-13
206   2018-01-14
207   2018-01-15
208   2018-01-16
209   2018-01-17
210   2018-01-18
211   2018-01-19
212   2018-01-20
213   2018-01-21
214   2018-01-22
215   2018-01-23
216   2018-01-24
217   2018-01-25
218   2018-01-26
219   2018-01-27
220   2018-01-28
221   2018-01-29
222   2018-01-30
223   2018-01-31
224   2018-02-01
225   2018-02-02
226   2018-02-03
227   2018-02-04
228   2018-02-05
229   2018-02-06
230   2018-02-07
231   2018-02-08
232   2018-02-09
233   2018-02-10
234   2018-02-11
235   2018-02-12
285   2018-04-03
286   2018-04-04
287   2018-04-05
288   2018-04-06
289   2018-04-07
290   2018-04-08
291   2018-04-09
292   2018-04-10
293   2018-04-11
294   2018-04-12
295   2018-04-13
296   2018-04-14
297   2018-04-15
298   2018-04-16
299   2018-04-17
300   2018-04-18
301   2018-04-19
302   2018-04-20
303   2018-04-21
dtype: datetime64[ns]
```

There are more tweets valleys in 2019 and 2020.

	created_day	counts
186	2017-12-25	288
197	2018-01-05	9
501	2019-01-01	287
603	2019-04-13	99
860	2019-12-26	286
1073	2020-07-26	267
1223	2020-12-23	224
1224	2020-12-24	207
1226	2020-12-26	252

Conclusions and Recommendations

- *Most prolific and influential twitterers tweet more about other topics than university related topics for all 4 universities. More outreach to those top twitterers regarding increasing their tweets about universities can help to expand universities' publicity.*

- *UChicago Twitterers tend to be more active on tweets, other 3 Universities should create more twitter public lists to encourage alumni to be more active and tweet more about their universities.*

- *Higher percentage of tweets on Big City Universities(BCU) come from local twitterers, because graduates from BCUs get more local employment opportunities. Small town universities should conduct more non-local outreach to alumni to increase university's influence.*

- *Big university news tend to cause tweets peak, universities should increase the publicity of good news and apply twitter damage control on bad news.*

- *Tweets on Big City Universities(BCUs) get more retweeted. Small town universities need to increase their influence/publicity on twitter.*