University Twitter Analysis

Presented by: Rohit Satishchandra
As Part of MSCA 31013: Big Data Platforms
March 19, 2021

Executive Summary

This project analyzed and **compared Twitter users ("Twitterers")** who tweet about the **University of Chicago** to those who post about three other peer institutions: **Northwestern University, MIT, and Yale University**. The key findings were as follows:

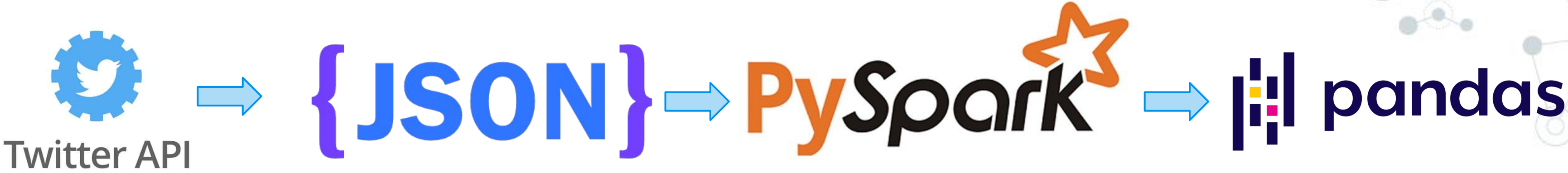
- Optimizing for followers count, total tweet volume or verified status is not fruitful. **Instead target users with high percentage of tweets related to University** and moderate followers.

- **67%** of identified Tweets about UChicago **are retweets**. Post high quality content aligned with University values to generate positive conversation

- Baseline Tweet volume does not display strong annual trend or seasonality. **Don't "over-optimize" on timing of content**

- UChicago Twitterers are comparatively more **localized in Chicago (6%)**

Data Overview



- Data was pulled from Twitter API into JSON files and loaded into a **PySpark** environment within **Google Cloud Platform** (Dataproc Hub)
- Full dataset was **~347M unique Tweets**, **reduced to ~2.2M Tweets** about UChicago and three selected peer institutions (Yale University, Northwestern University, MIT)
    - Tweets covered a **nearly 4 year period (with 56 missing dates)**
- Reduced data comprises **~1.6M unique users** and was analyzed in Python Pandas (in GCP and Kaggle notebook)

Methodology

- Relevant **Tweets** were identified using **keyword search** in PySpark SQL
  - ○ Restricted search space (**Tweet text, hashtags, and original Tweet text/hashtags if retweet**)
  - ○ Text converted to **lowercase**
  - ○ Limited set of possible matches to **minimize false positives** (ex. “uchicago”, university of chicago”, “ u of c ”

- Reduced data was still unique Tweets
- Needed to **ensure** unique set of users for Twitterer analysis
  - ○ **Sort by user_id, tweet_created_at**
  - ○ **Group by university, user_id**
  - ○ **Select first record in each group**
- For each user, compute **weighted sum of original Tweets, replies, and retweets**
- Analyze **volume, location, and timeline**

Political Science professor at UChicago -- high follower count, and percentage of tweets related to UChicago

Influential Users

University	Most Recent Tweet	name	location	followers_count	university_related_tweets	pct_university_tweets
UChicago	2019-07-11	Paul Staniland	Chicago IL	18522	412	91
UChicago	2021-03-14	V.TITOV	Kiev, Ukraine	2028	68	35
UChicago	2018-03-08	Christine Fair	Capital of Jesus-e-Stan	628	9	90

- Influential Twitterers are of potential value to University of Chicago because a **high proportion of their Tweets are related to the University and they have high follower count**
- These profiles' should be further analyzed for sentiment, and if positive, **targeted because of their high interest in University life**, with events and brand messaging

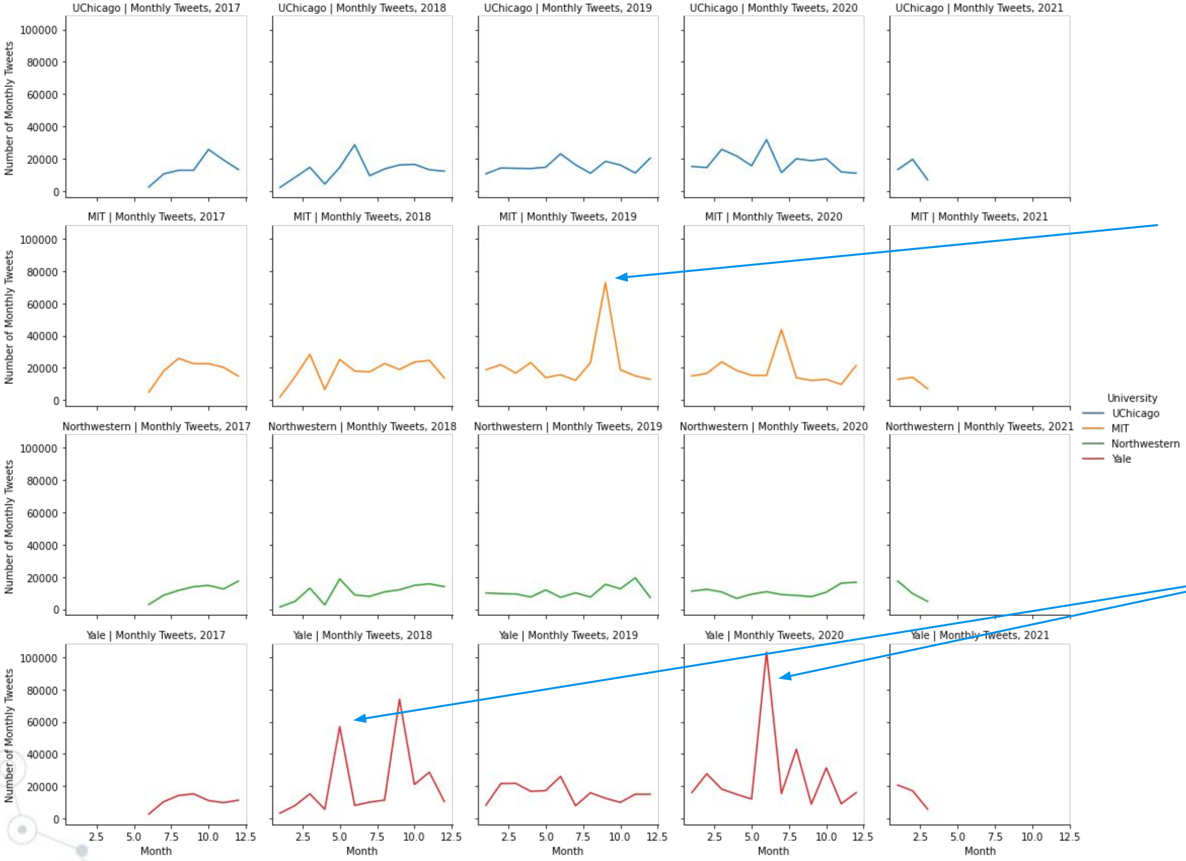
Location Analysis

- For each university, approximately **30% of users do not provide a location** in their profile
  - Major US Cities (New York, Chicago, Los Angeles, Washington DC) often appear in **top 10 most common cities**
- Among **users who tweet about UChicago, 6% (~21K) list Chicago** as a location
  - ~83K unique locations listed (not including nulls)
- Most other users who tweet about other universities are not based in same city as the school
  - **0%** of those who tweet about Yale and provide location list **New Haven or Connecticut**
  - **0%** of those who tweet about Northwestern and provide location list **Evanston (3% Chicago)**
  - **3%** of those who tweet about MIT and provide location list **Cambridge or Boston**

Timeline Analysis

Rows represent Universities, columns represent years

No strong trend or consistent seasonal pattern. Apparent spikes correlated with spring commencement and slightly higher volume in summer months



Possibly tied to “MIT Media Lab Crisis,” August 2019

Yale Outliers Corresponding to Annual Commencement

Uniqueness of Tweets

- **Greater weight assigned to “original” Tweets** compared to replies and retweets (lowest weight)
  - For each user:
    - **weighted_score** = $$\frac{\text{num\_originals} + 0.9 \text{ x } \text{num\_replies} + 0.75 \text{ x } \text{num\_retweets}}{\text{total\_university\_tweets}}$$
- All 4 Universities have **significantly more retweets than originals/replies (average 65% retweets)**
  - **MIT has most total originals and replies, Yale has fewest originals** and most retweets

University	Mean Weighted “Originality” Score
Northwestern	0.822
MIT	0.812
UChicago	0.803
Yale	0.787

Recommendations

- Twitterers with high tweet volume, follower count, or “verified” status are often major news organizations, billionaires, or accounts already affiliated with University. **Instead, target “influential” users with high percentage of tweets related to University** and high/moderate follower count.

- **67%** of identified Tweets about UChicago **are retweets**. Twitter is already a major online hub of discourse: **take firm stances and post high quality media** (video spotlights, student/faculty profiles, etc.) to **generate positive brand impression.**

- UChicago Twitterers are comparatively more **localized in Chicago (6%). Analyze sentiment of original Tweets and tailor programming/events.**

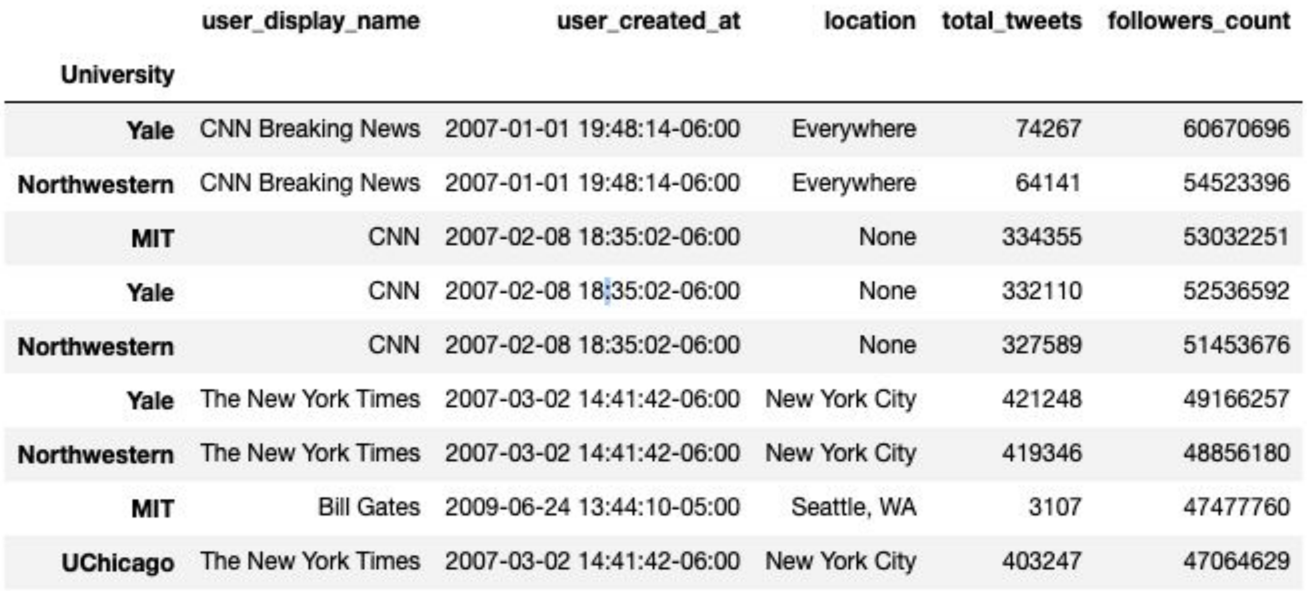
- Baseline Tweet volume does not display strong annual trend or seasonality there is **no need to “over-optimize” on timing of content**

Appendix

Top 5 Influential Users

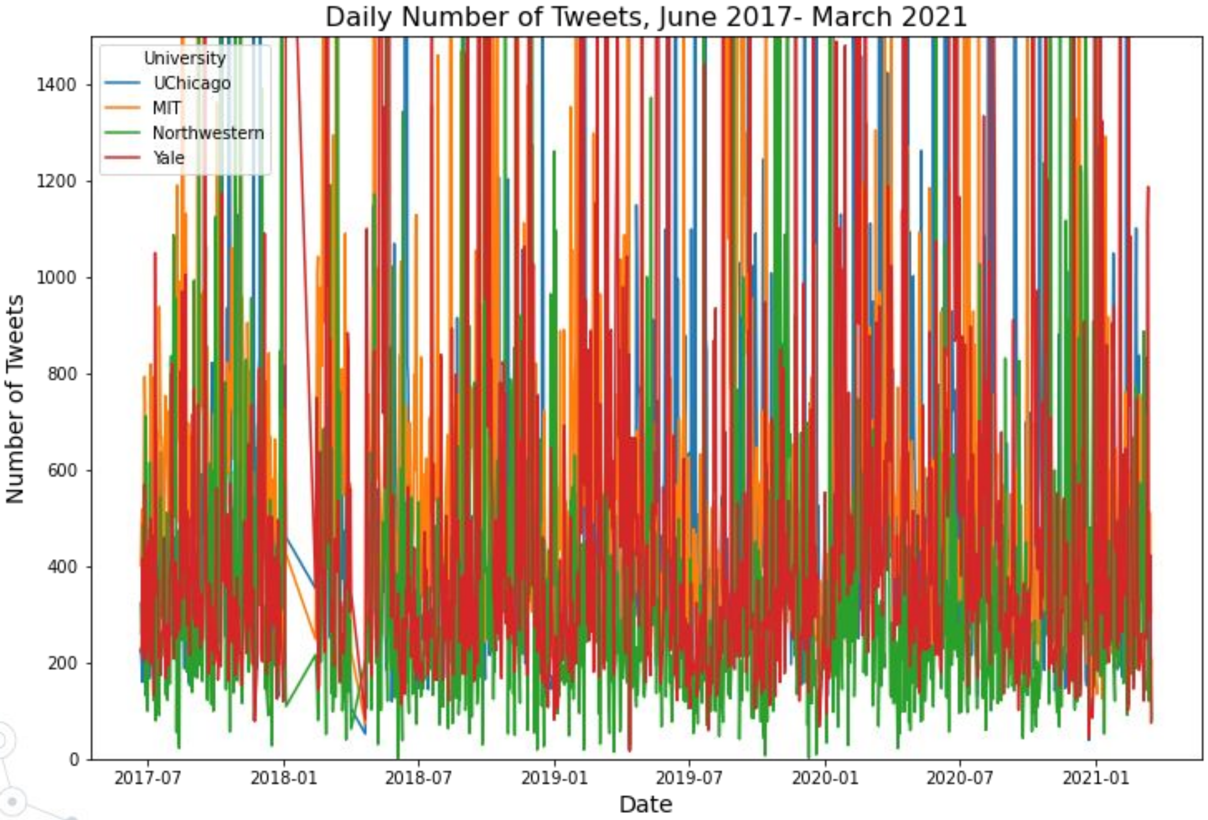
University	Most Recent Tweet	name	location	university_related_tweets	pct_university
UChicago	2019-07-11	Nasir Smith	None	26	93
UChicago	2021-03-14	Paul Staniland	Chicago IL	412	91
UChicago	2018-03-08	Sophia Vlahakis	None	9	90
UChicago	2017-07-22	Art Of The Cure	Chicago IL	19	90
UChicago	2017-07-22	Scott Wilson	Los Angeles CA	245	84

News Accounts Dominate Follower Count



	user_display_name	user_created_at	location	total_tweets	followers_count
University					
Yale	CNN Breaking News	2007-01-01 19:48:14-06:00	Everywhere	74267	60670696
Northwestern	CNN Breaking News	2007-01-01 19:48:14-06:00	Everywhere	64141	54523396
MIT	CNN	2007-02-08 18:35:02-06:00	None	334355	53032251
Yale	CNN	2007-02-08 18:35:02-06:00	None	332110	52536592
Northwestern	CNN	2007-02-08 18:35:02-06:00	None	327589	51453676
Yale	The New York Times	2007-03-02 14:41:42-06:00	New York City	421248	49166257
Northwestern	The New York Times	2007-03-02 14:41:42-06:00	New York City	419346	48856180
MIT	Bill Gates	2009-06-24 13:44:10-05:00	Seattle, WA	3107	47477760
UChicago	The New York Times	2007-03-02 14:41:42-06:00	New York City	403247	47064629

Daily Tweet Totals



During the initial phases of EDA, you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends.

- *Hadley Wickham*