BIG DATA PLATFORMS, SPRING 2021

TWITTER ACCOUNT PROFILE ANALYSIS (500M TWEETS)

Student: Nic Carlson
Instructor: Nick Kadochnikov

EXECUTIVE SUMMARY

Project Goals

Goal 1: Identify the profiles of Twitterers who are tweeting about University of Chicago and compare them to the profiles of Twitterers who are tweeting about other universities

Goal 2: Make actionable business recommendations to help University improve the social media outreach programs.

Data Processing

- Use keywords to select uChicago, Harvard, Northwestern, and Yale tweets
- 1% of 500M tweets were 'relevant'
- Twitter API features very sparsely populated
- Several key features selected for analysis

Message Similarity

- Leveraged Jaccard Distance to determine Tweet similarity
- Selected .7 as distance threshold given marginal change from .3, .5, .7
- 20% of messages are similar, which makes sense given high level of 'bot' activity and given Retweeting is large part of Twitter culture

Twitterer Profiling

- Definition of 'Influencers' user depends heavily on metric used (quality vs quantity). uChicago has most active university accounts, with least active top 'influencers'
- uChicago 'localized' Twitterer base most local vs competing colleges
- uChicago 'Retweet influencers' less popular than competing universities, but they tweet much more frequently about university
- Twitterer activity peaks in PM, with variation between university

Recommendations

- Develop method for tagging bots and university accounts to enable marketing penetration analysis and true 'Twitterer' analysis
- Leverage sum of retweets as metric for measuring 'influence'
- Consider ways to engage new geographies to compete with other major universities (Specifically: CA, TX, India)
- Focus advertising spend on peak Twitterer activity hours (2-10pm)
- Consider ways to engage and encourage influencers (scholarships?)
- Leverage LSH to speed up similarity analysis. Then segment analysis by school, location, and user type. Filter out similar tweets created after date of first 'tweet_created_at' time.

DATA FILTERING & SELECTION

Methodology and source data overview	**Data consisted of 500 million Tweets**, so it was necessary to use cloud computing via a Google Cloud Services auto-scaling cluster, accessed via Spark **Leveraged Spark for big data processing** (EDA, data filtering, and feature selection), exported data to CSV, and then leveraged Python for final analysis once features and observations were both significantly narrowed
Tweet clean-up and filtering	Wholistic keyword list developed to filter for uChicago text, including: 1) school names (e.g. Booth, Graham), 2) school environment (Maroons, Rockefeller Chapel, etc.), 3) staff (e.g. Robert J. Zimmer, Robert Zimmer) Tweets also selected for Harvard, Northwestern, and Yale **Output was ~6M million tweets, roughly 1% of full data set**
Exploratory Analysis and Variable Selection/Engineering	**Selection process**: 1) Twitter documentation helped identify ~30 initially relevant fields, 2) Leveraged Spark to calculate percent of null values in each column, 3) Removed non-essential columns that were >20% null. Most essential include: user_screen_name, tweet_text, retweet_count **Engineering and cleaning**: Exported filtered data to desktop and leveraged Alteryx for data cleaning and feature engineering. Engineered features for text length and school.

USER ANALYSIS REVEALS 4 IMPORTANT TRENDS

1 Definition of 'influencer' is important to define

Method drastically changes list of influencers

1. Number of tweets tags bots and university accounts as influencers. May be helpful to get better understanding of University Twitterer Footprint

2. Number of retweets tags 'organic' influencers that other users retweet. Suggest using this for analysis of true user analysis

3. Hybrid: Weighted number of tweets: Retweets worth 2x value of Tweet. Balances two methods above by assigning 'score' of 1 to Tweets and a bonus 'point' ('score' of 2 total) to offset Twitterers that might have had one or two Tweets that went viral and to give credit for volume of univeresity related Tweets

See table of top Twitterers on next slide

2 Depending on university, Twitterer location has varying relationship

uChicago and **Northwestern** twitter activity are both relatively contained to the Chicago area, with spikes in their respective towns.

Harvard and **Yale** have a further reaching geographic tweet influence

See backup chart in two slides

3 Twitterer activity peaks in evening, with differences between University

Overall, peak was at 10pm and off-peak at 9am, but uChicago and Yale had less hourly seasonality compared to Harvard and Northwestern.

See backup chart in three slides

TOP TWITTERS VARY GREATLY BASED ON CHOSEN 'INFLUENCER' METRIC

Type of Twitterers vary by school
- In all cases, 'Retweet' metric unearths 'organic' users
- **uChicago** appears to have most active University sponsored accounts, smallest organic influencer base
- **Northwestern** only potential University account appears to be mba_buddy
- **Harvard** has larger sports presence + two bot Twitterers (AmericaAlerts is weather + newstarsbot)
- **Yale** 'influencers' represented by mostly 'organic' users, with exception of Finance_graduat

Value legend:
Tweets = Count of Tweets: "Noisy due to University accounts and Bots"
Retweets = Sum of user Retweets: "Organic Influencers" <- Suggested metric
Hyrbid = 1 point per Tweet with no Retweets, 2 points per Tweet with Retweets: "Retweet weighted influencers" still includes lots of University accounts

Of top 5 Retweet Influencer Tweets

11% of 2.5k Tweets were uChicago related

1% of 11.5k Tweets were NW related

uChicago

Tweet		Retweet		Hybrid	
chicagoboothrev	639	CyanFour	25502	UChicagoMedJobs	2153
uchicagogsu	457	yiizo	24480	chicagoboothrev	1424
AlexBender7	392	the_gwelle	24437	ChicagoMaroons	1234
ruthpaget	359	Katja_Thieme	24183	uchicagogsu	1128
n8taki	291	YamraaMcNeil	24064	AlexBender7	785

Northwestern

Tweet		Retweet		Hybrid	
mba_buddy	810	idekchie	688240	HarvardStudyBot	2272
grhluna24	803	_nnejib	632280	kryssew	1915
BourbonStreet2	753	Christian 856	603869	cecesoojung	1849
FrasierHerry	550	olanaaaae	589366	mba_buddy	1620
WaterP0100714	206	Rosenchild	544043	grhluna24	1606

Harvard

Tweet		Retweet		Hybrid	
NSUDemonsFans	2500	wmzafravelasco	89722	AmericaAlerts	6313
Sportsuoumal	1019	raybae68g	74209	NSUDemonsFans	5620
Andrew weather	933	jared_secret	44027	WXHAZAROSKS	4643
cbhhughes21	853	JacobG88	39369	newstarsbot	3642
sportsKT	732	bz80qrlXkWpoBcT	38263	insidenu	3285

Yale

Tweet		Retweet		Hybrid	
kinibottommintd	1661	meghanaae	219025	kinibottommintd	6994
honischgalerie	691	raybae68g	186120	honischgalerie	1801
danteren010	230	iambrittanie	180870	esquireattire	1349
wendchain	224	EXFUR181STA	120301	Finance_graduat	562
n8taki	223	kaylaiko	114448	danteren010	460

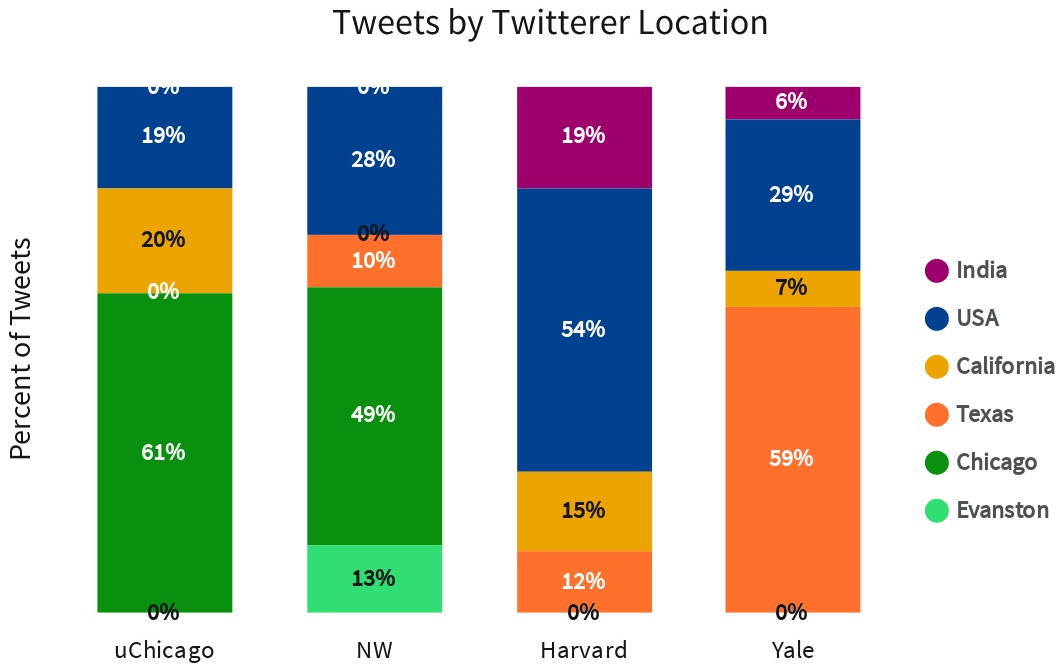
Of top 5 Retweet Influencer Tweets

0.5% of 13.5k Tweets were Harvard related

1% of 6.3k Tweets were Harvard related

DETAIL SLIDE: PERCENT OF EACH SCHOOLS' TWEETS BY LOCATION

uChicago has opportunity to grow non-Chicago twitter user base, potentially raising school reputation



Internationally, Harvard and Yale have far higher percent of their Tweets from India, potentially increasing school attraction outside US

Domestically, uChicago has gap of users in Texas compared to other universities. **Despite being in same city, TX makes up 10% of Northwestern tweets**

Note: Due to sparse tweet location data, this data only represents ~225k tweets

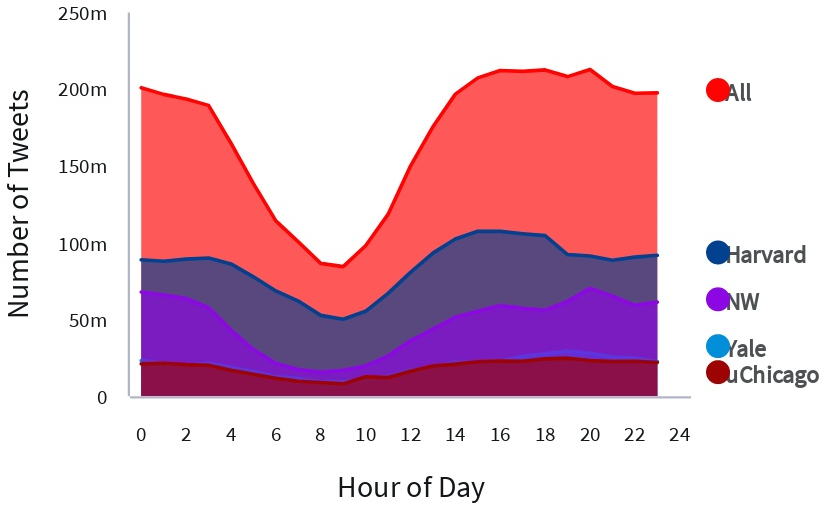
Note: Similar locations 'collapsed' based on proximity to universities in scope (e.g. 'Los Angeles' and 'California' combined whereas 'Chicago' and 'Evanston' kept separate)

DETAIL SLIDE: TWEET TRAFFIC TENDS TO BE CYCLICAL, WITH

1. Overall, daily seasonality definitely exists in Twitter patterns
2. uChicago and Yale's Twitterers have much less daily seasonality, suggesting some differences in users
3. Harvard's larger number of 'late-night' Tweets may be due to a more global Twitterer base with different cultural sleeping patterns

Used Jaccard Distance as measure for uniqueness of messages
Overall threshold had little impact on results
Determined that 0.7 threshold was most appropriate by comparing output of each threshold

Tweet Volume Over Time



Unique Tweets Using Jaccard Distance