Similarity Analysis of Twitterers

MSCA 31013 Big Data Platforms

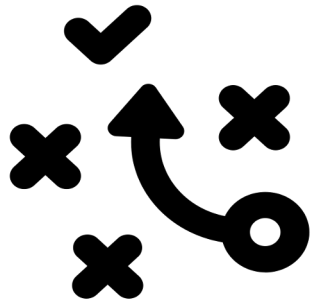
Bharadwaj Kacharla

Problem Statement

- Twitter is an American microblogging and social networking service on which users post and interact with messages known as "**tweets**".

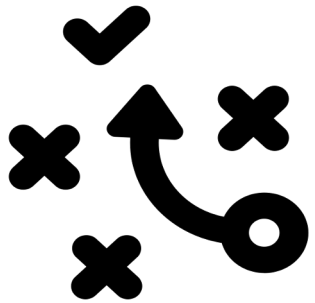
- This project deals with exploring the profiles of **Twitterers** and finding the similarity between the tweets

Executive Summary

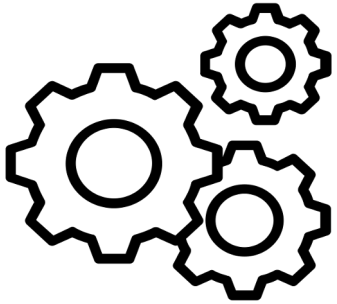


Executive Summary

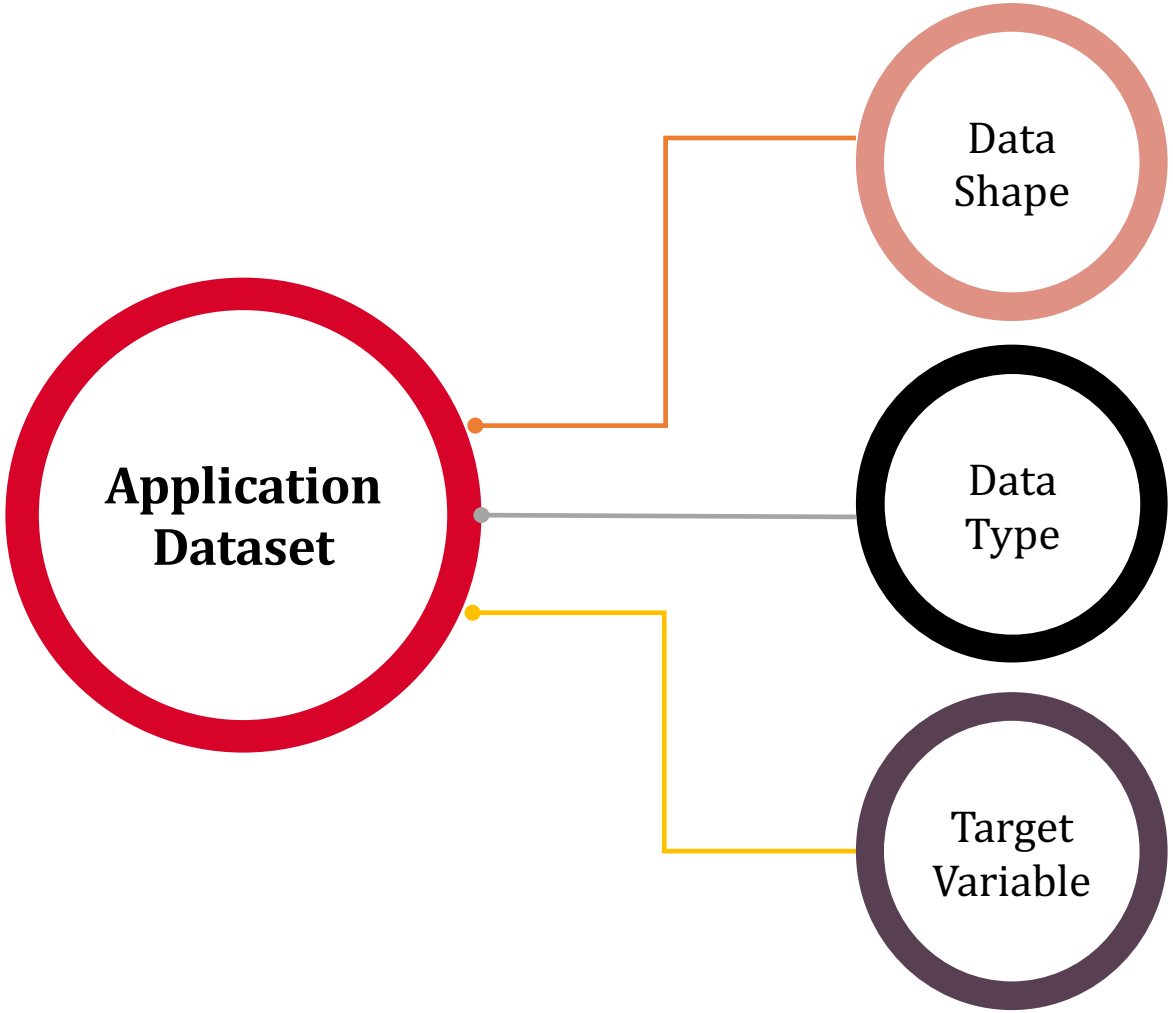
- Twitter activity associated with **The University of Chicago, Harvard University, Stanford University** and **Northwestern University** has been filtered and analyzed to understand the twitter activity
- Amongst all the universities **Harvard University** has been associated with a **major share of twitter activity**
- The accounts associated with the universities significantly engage in twitter related to university and majority of the twitter activity is from the twitterers who are in the vicinity of the university



EDA & Feature Engineering



Data Overview



- 40 struct variables
- 300M+ rows

- 12 struct variables
- 12 string variables
- 10 long variables
- 6 Boolean variables

- Each row is a tweet, retweet is recorded in the same way a tweet is recorded

Feature Engineering Process

1

Drop Columns, Rows

- Select columns that are relevant in identifying the profiles of a twitterer
- Drop columns that are irrelevant or duplicated
- Filter for tweets that are related to UChicago, Harvard, Stanford and Northwestern universities

2

Deal with Categorical/Boolean variable

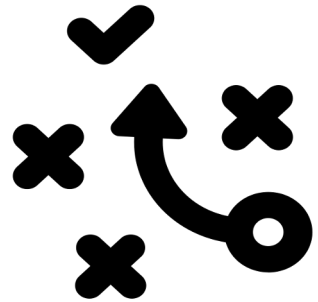
- Reduce levels in some categorical variables
- One-Hot Encoding

3

Tweets handling

- Convert the tweets to lower case to optimize the university identification process
- Label the non-university tweets to profile the twitterers

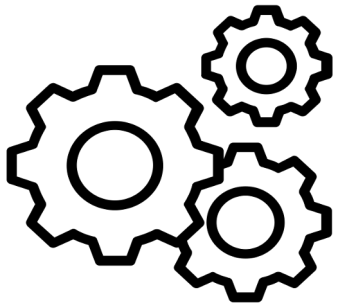
Twitterers Profiling



Data Treatment

Data Handling

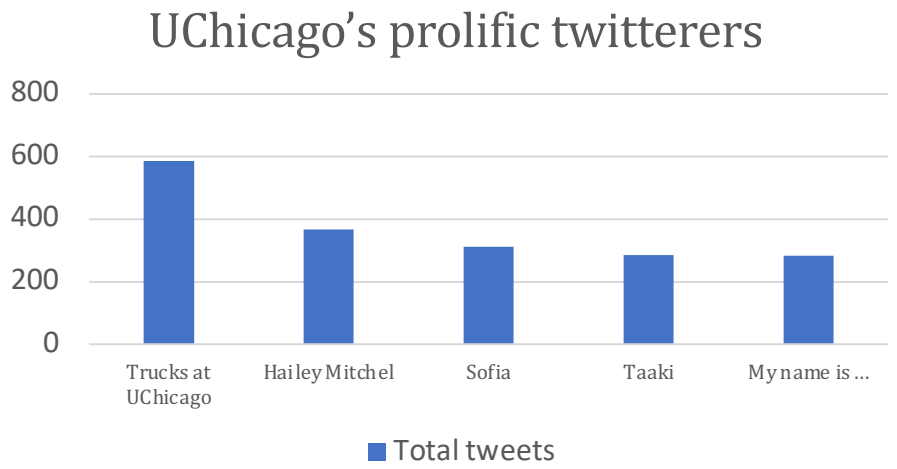
- Twitter APIs return tweets provide the data encoded un json form. JSON is based on key-value pairs, with named attributes and associated values. These attributes, and their state are used to describe objects.
- Each Tweet has an author, a message, a unique ID, a timestamp of when it was posted, and sometimes geo metadata shared by the user. Each User has a Twitter name, an ID, a number of followers, and most often an account bio
- We source the information from these columns to perform additional analysis and profile the users accordingly

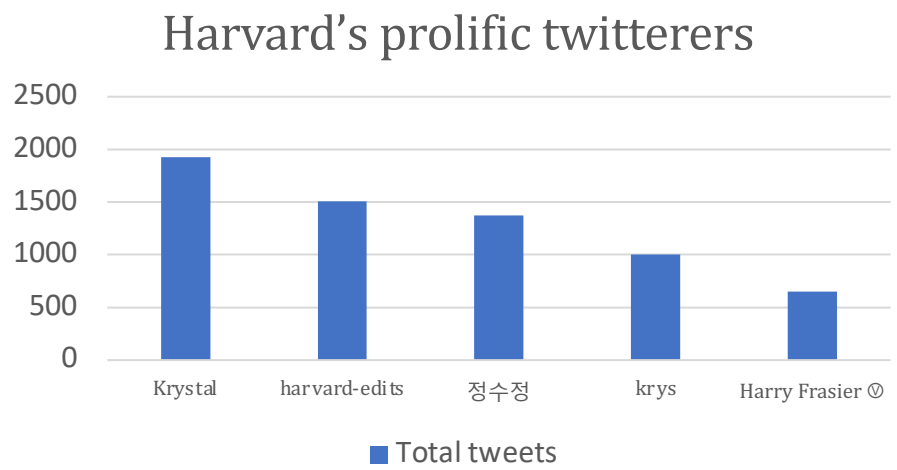


Prolific Twitterers

Following are the profiles and corresponding tweet volume of the twitterers across the universities

- Based on the initial assessment, the popular accounts belong to some twitterers and few accounts are related to some activities related to the university
- For example**, Trucks at UChicago, Hailey Mitchel, Krystal and Harvard-edits** are few accounts that are very prolific based on the number of tweets

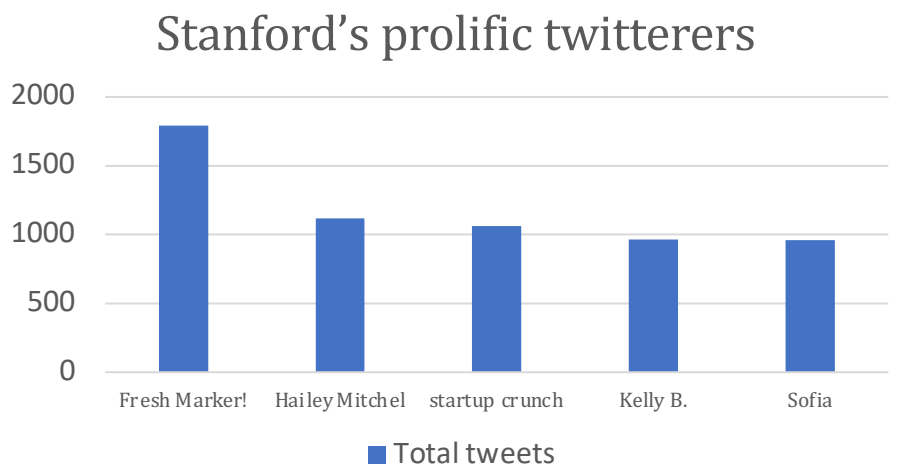


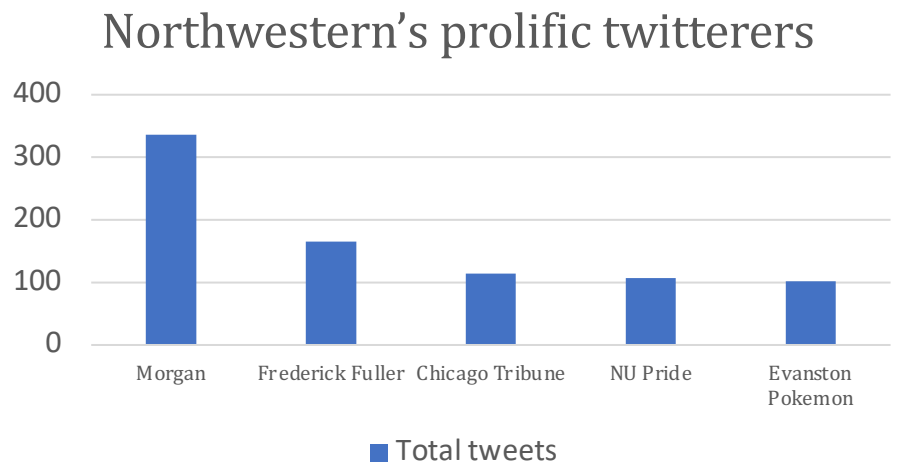


Prolific Twitterers

Following are the profiles and corresponding tweet volume of the twitterers across the universities

- Based on the initial assessment, the popular accounts belong to some twitterers and few accounts are related to some activities related to the university
- For example, **Fresh Marker!, startup crunch, Chicago Tribune, NU Pride** are few accounts that are very prolific based on the number of tweets

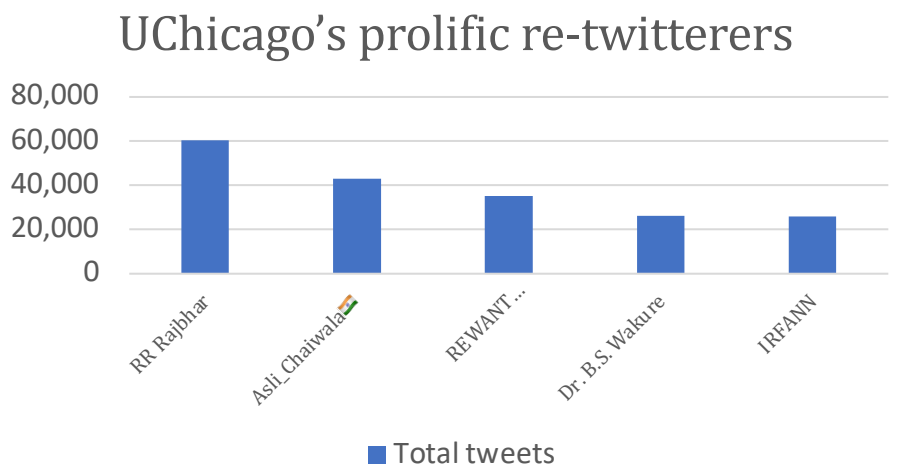


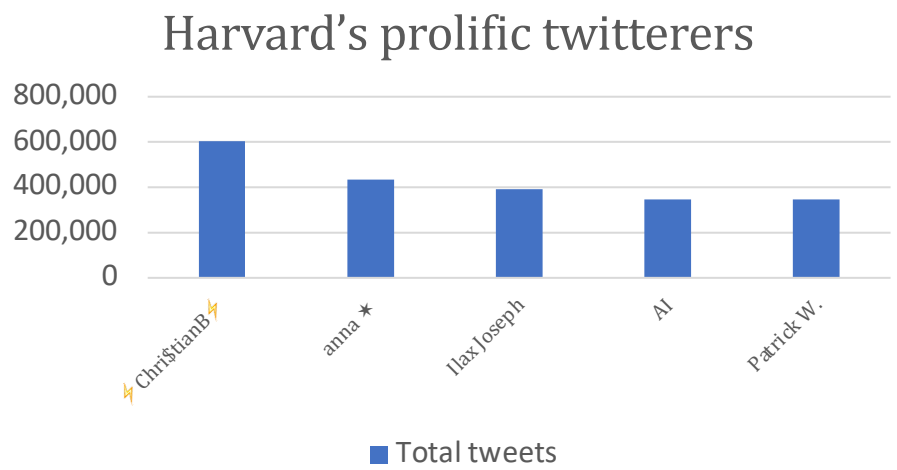


Prolific Twitterers

Following are the profiles and corresponding retweets volume of the twitterers across the universities

- Unlike the twitterers, re-twitters retweet a lot of tweets.
- Based on the retweet volume, **Harvard university** has lot of tweet activity when compared to **UChicago**
- Considering the popularity of The Harvard university, the tweet activity seemed to be in-line

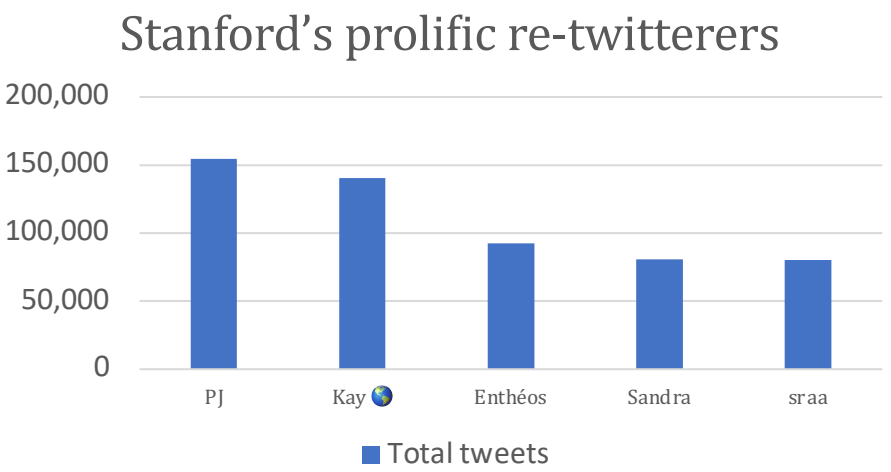


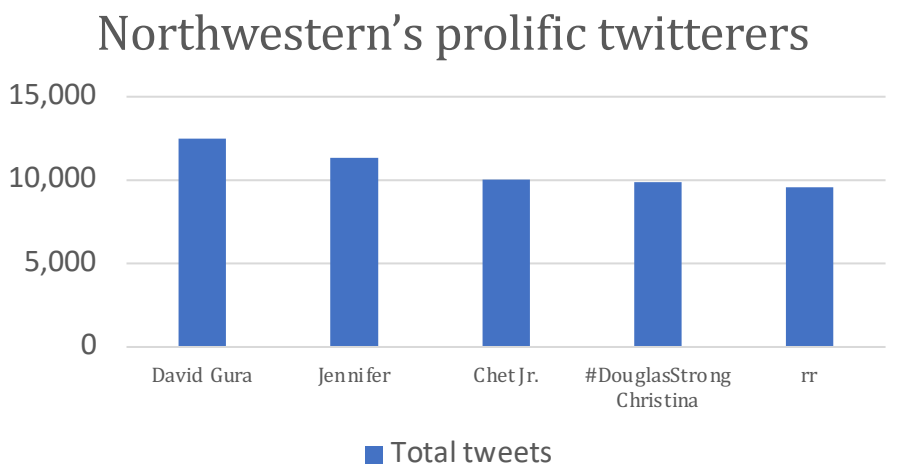


Prolific Twitterers

Following are the profiles and corresponding retweets volume of the twitterers across the universities

- Unlike the twitterers, re-twitters retweet a lot of tweets.
- Based on the retweet volume, **Stanford university** has lot of tweet activity when compared to **Northwestern**
- Significant retweet activity can be associated with **Harvard university** followed by **Stanford, UChicago and Northwestern**

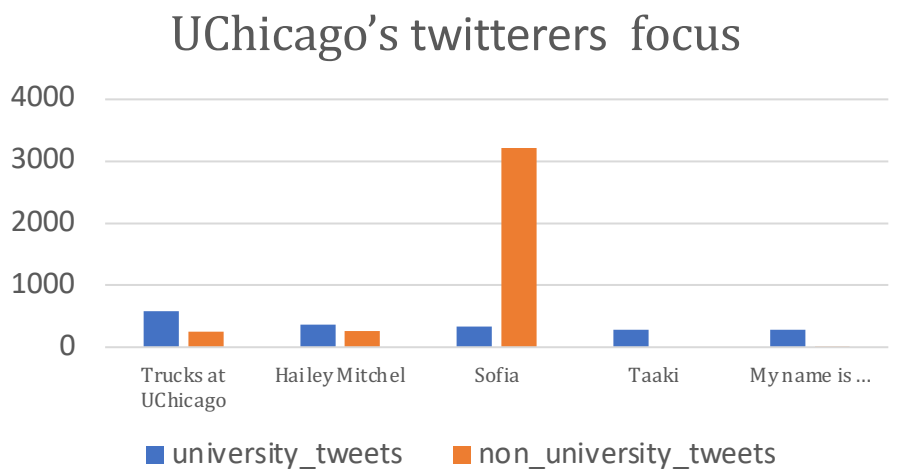


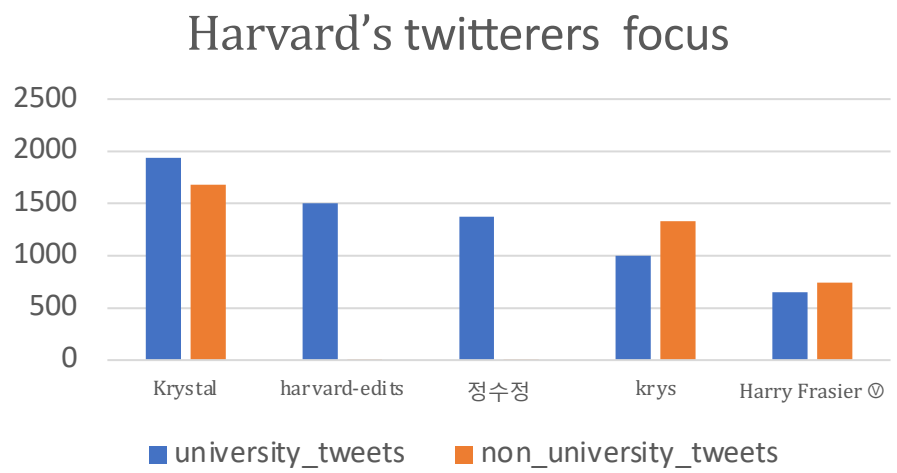


Tweet activity - Are they tweeting about other topics?

The twitterers tweet about a lot of topics, however, keeping the scope of project in consideration we have labeled the tweets into university vs non-university tweets

- The profiles associated with the universities UChicago and Harvard, namely - **Trucks at UChicago** and **harvard-edits** mostly tweet about university related topics
- Profiles associated with twitterers like **Sofia, Krystal** and **krys** tweet about both university and non-university topics
- Twitterer associated with **Sofia** profile seems to be tweet **a lot about non-university topics when compared to university related topics**

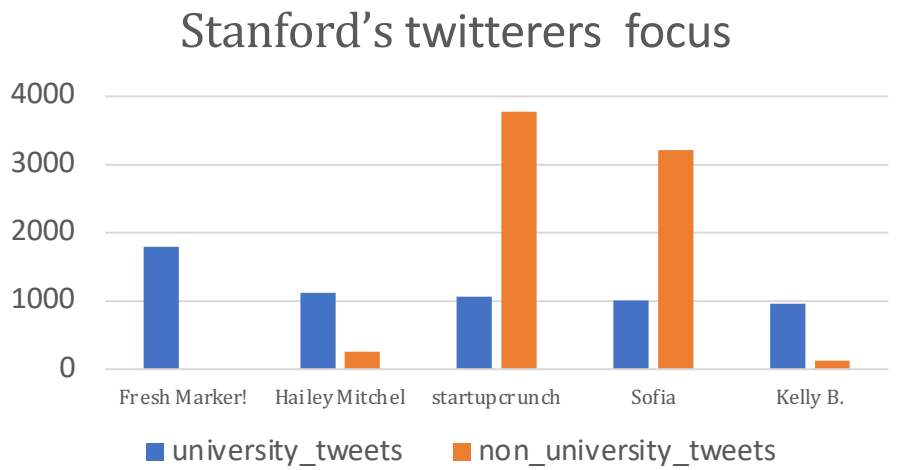


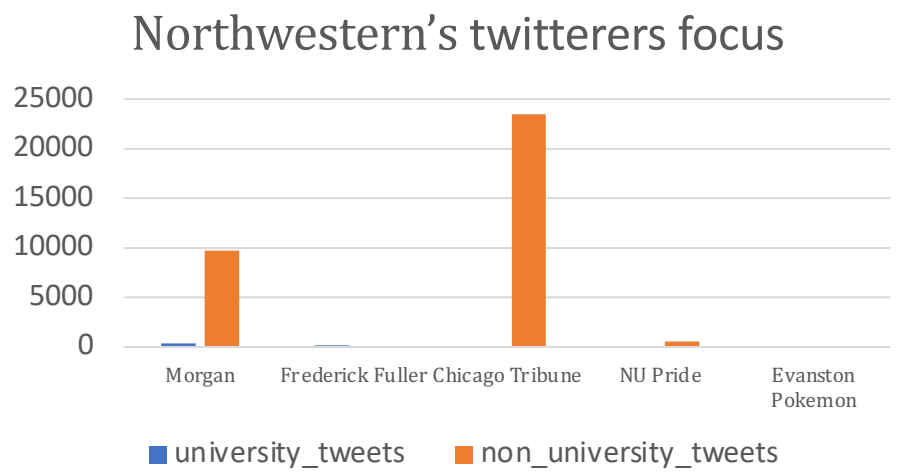


Tweet activity - Are they tweeting about other topics?

The twitterers tweet about a lot of topics, however, keeping the scope of project in consideration we have labeled the tweets into university vs non-university tweets

- The profiles associated with the universities Stanford and Northwestern , namely - **Fresh Marker!** mostly tweet about university related topics
- Few profiles like startup crunch, **Chicago Tribune** and **NU Pride** tweet significantly a lot more about non-university topics than university topics
- The profile names provides additional evidence about the higher tweet activity related to non-university topics
  - **Chicago Tribun**e - A newspaper that covers a lot of daily news, so the activity around non-university tweets is justified

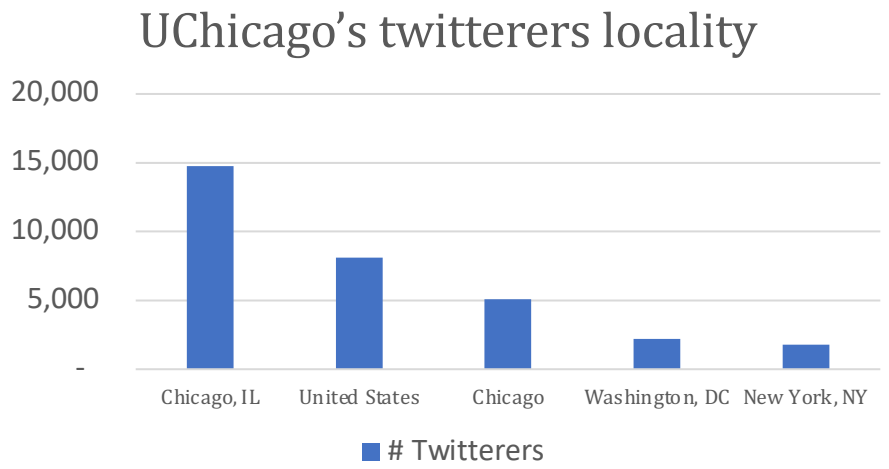


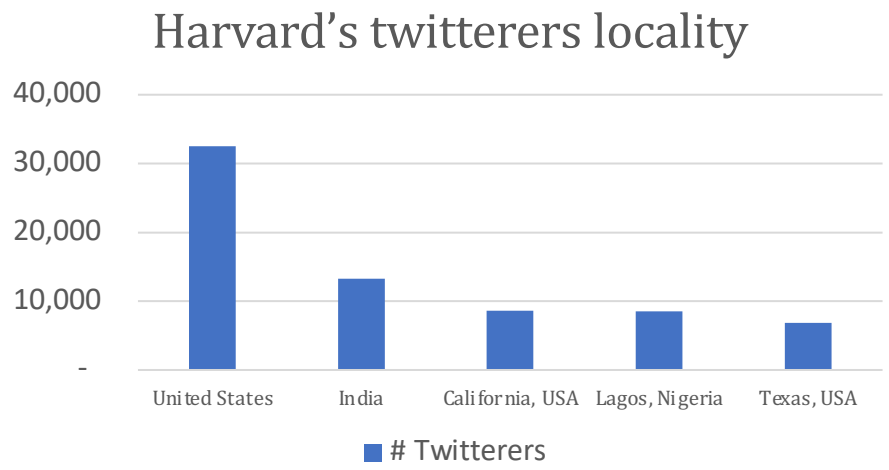


Twitterers Location - Where are they located?

Based on the location variable, majority share of the twitterer's are located in the same locality as the university is located

- For **UChicago**, significant number of users are based in Chicago followed by Washington DC and New York
- For **Harvard**, significant users are based in United States, followed by India and California, USA
- There are few discrepancies in the data collection of the location variable. In few instances the locality of the user isn't collected
    - For example, Harvard university has significant twitterers located in United States. However, the location should be a locality instead of whole United States

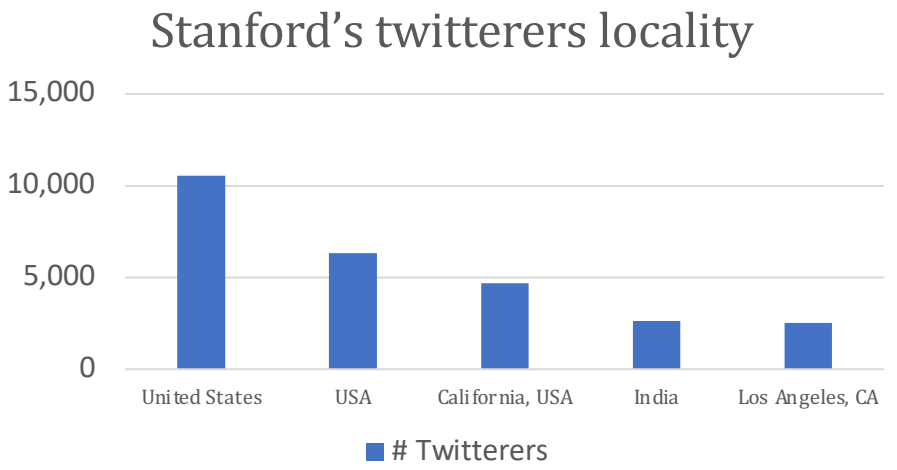


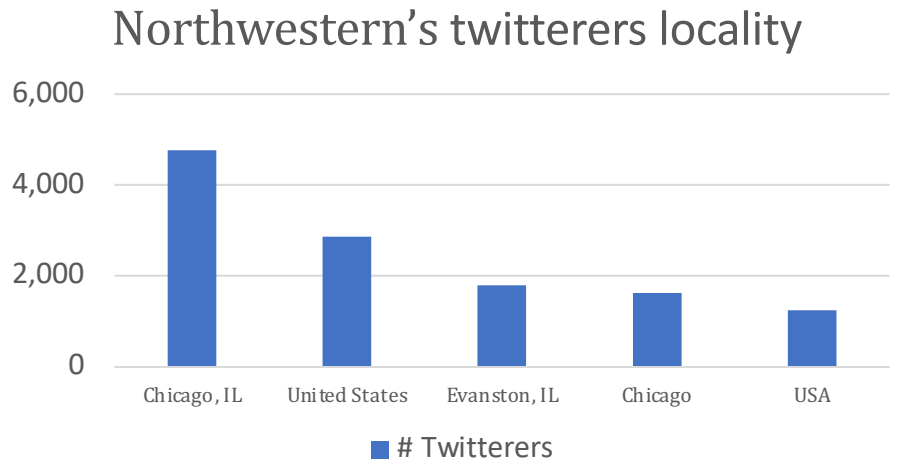


Twitterers Location - Where are they located?

Based on the location variable, majority share of the twitterer's are located in the same locality as the university is located

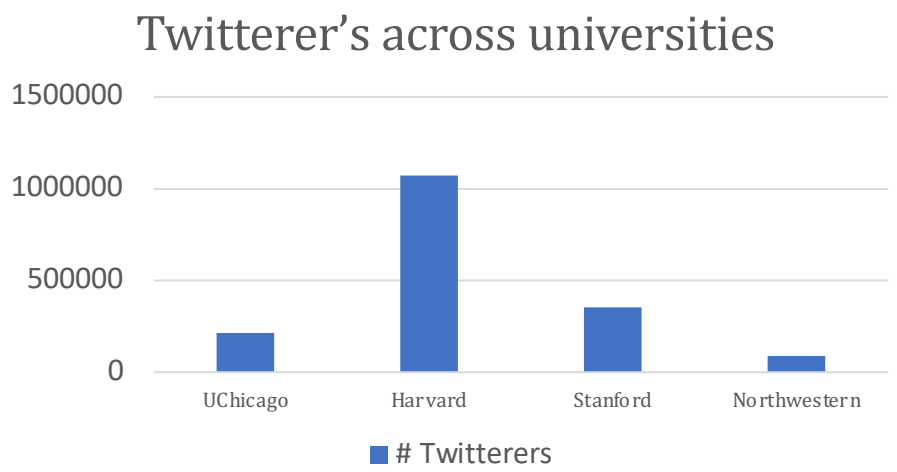
- For **Stanford**, significant users are based in United States, followed by California, USA and India
- For **Northwestern**, significant users are based in Chicago, IL followed by Evanston
- There are few discrepancies in the data collection of the location variable. In few instances the locality of the user isn't collected
    - For example, Stanford university has significant twitterers located in United States. However, the location should be a locality instead of whole United States

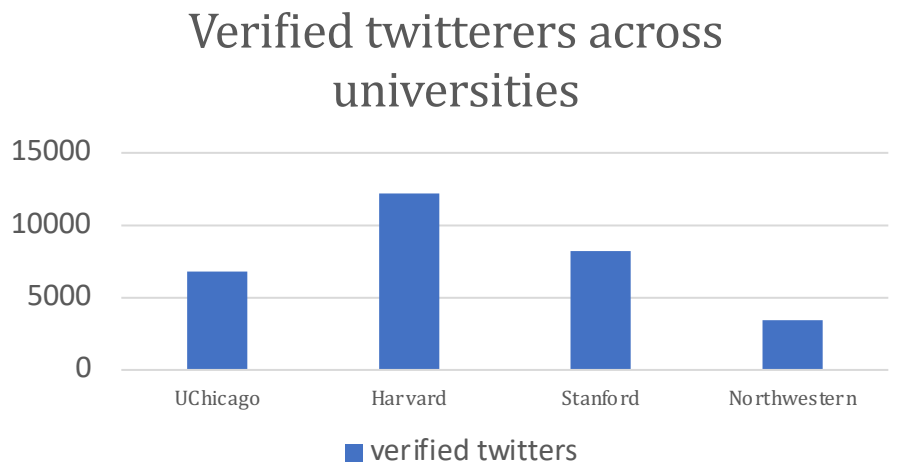




University Twitterers - Who are they?

- **Harvard has significantly higher number of twitterer's** followed by UChicago, Stanford and Northwestern
- A similar trend is seen across the verified twitterers. **Harvard has a higher verified twitterers** followed by Stanford, UChicago and Northwestern
- The twitter activity reinforces the popularity of the universities. **Harvard** being one of the most popular universities in the world

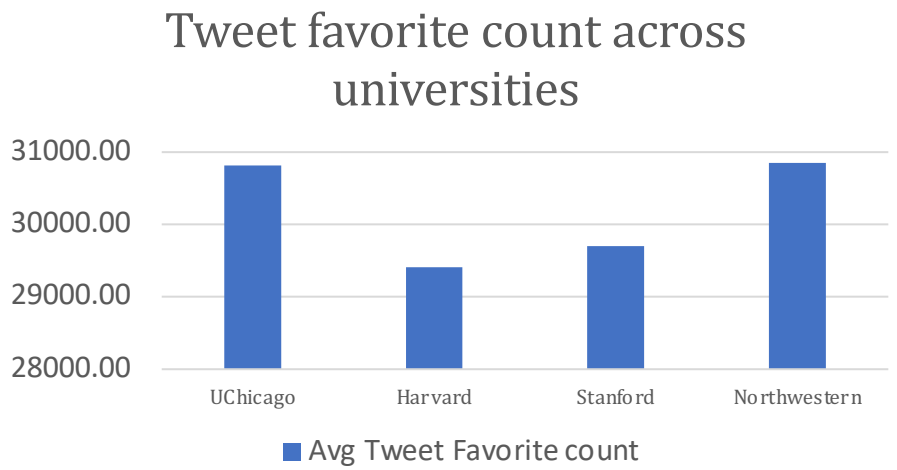


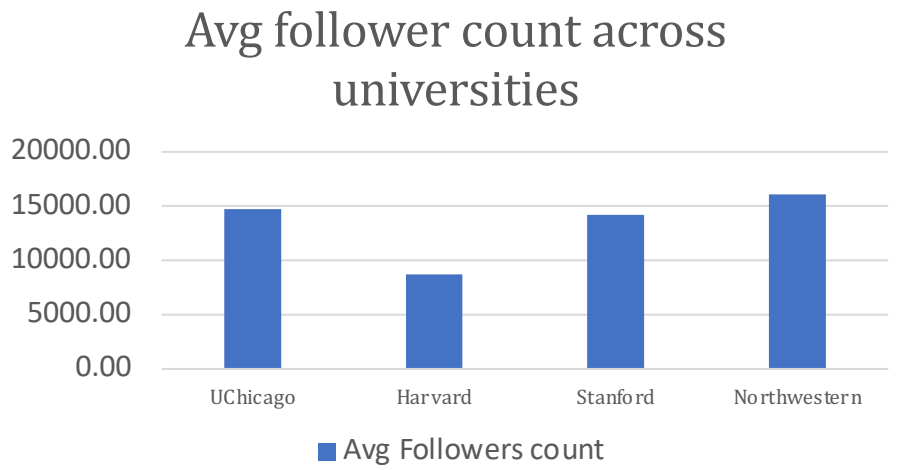


University Twitterers - Who are they?

Even though Harvard has the highest number of twitterers and verified twitterers, same trend isn't observed in **tweet activity**

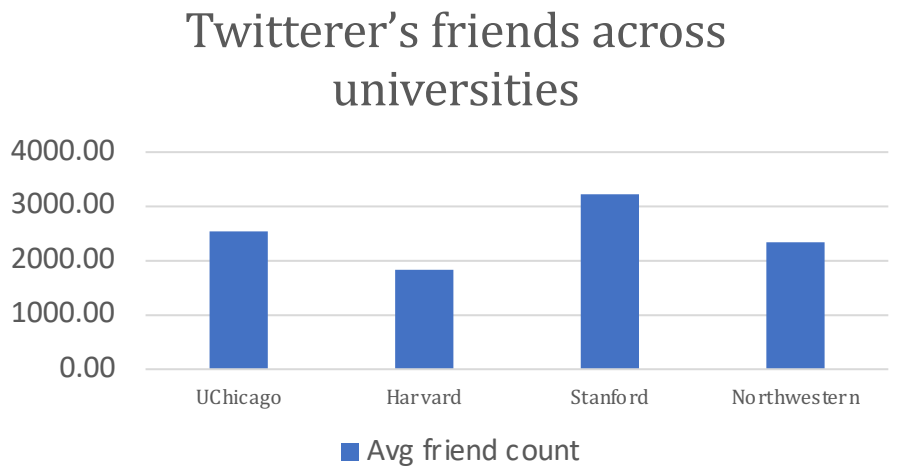
- **UChicago** has a higher twitter engagement, followed by **Northwestern, Stanford and Harvard**
- A similar trend is seen across the **avg no. of follower's** associated with **Northwestern** university twitterers, followed by **UChicago, Stanford and Harvard**

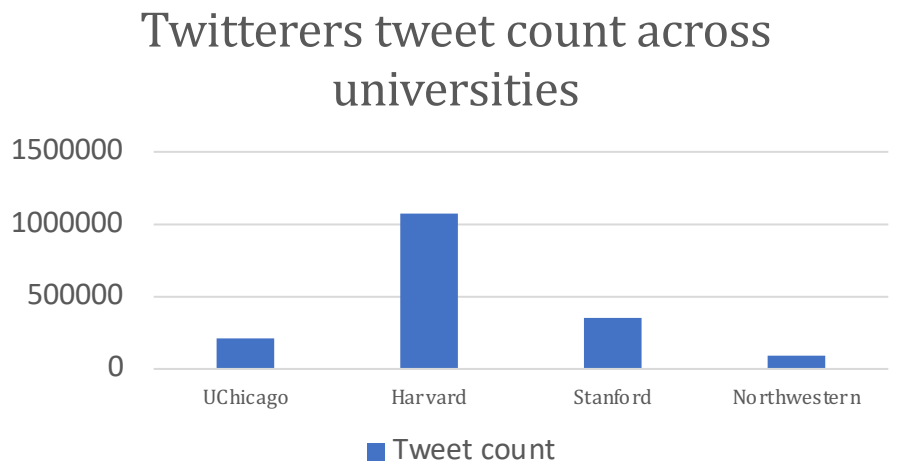




University Twitterers - Who are they?

- When it comes to **avg number of friends** the twitterers have, **Stanford tops the list, followed by UChicago, Northwestern and Harvard**
- Finally, when it comes to **tweets**. **Harvard has significantly higher number of tweets** associated with the university followed by **Stanford, UChicago and Northwestern**

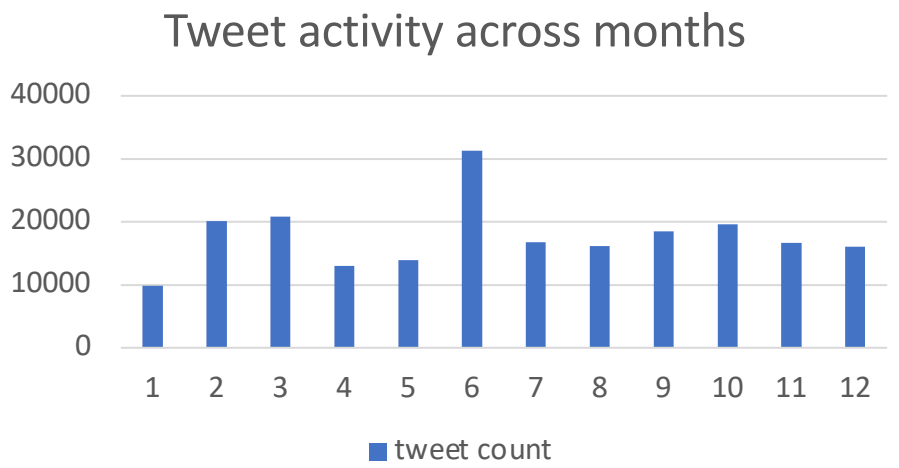


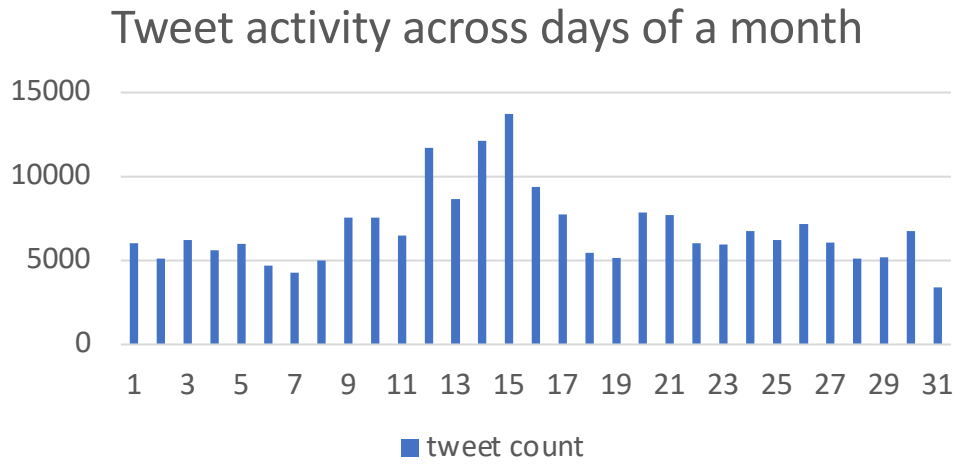


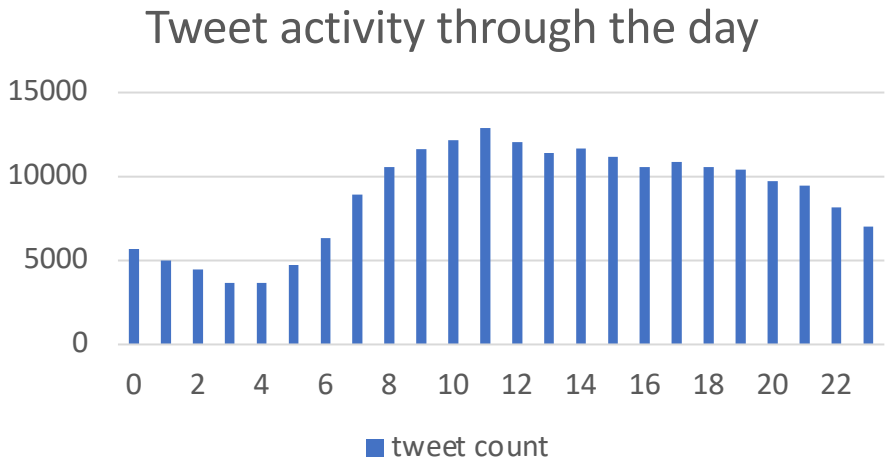
When do they tweet ? - UChicago

Twitter activity associated with **UChicago** is as follows:

- Twitter activity is on the rise as the month progresses. **January** has the **least twitter activity** and **June** has the **highest twitter activity**
- For any given month, the twitter activity is **a bit low in the beginning and the end compared to other days**
- For any given day, the twitter activity seem to be on the rise as the day progresses. **Maximum twitter activity is observed in the noon**



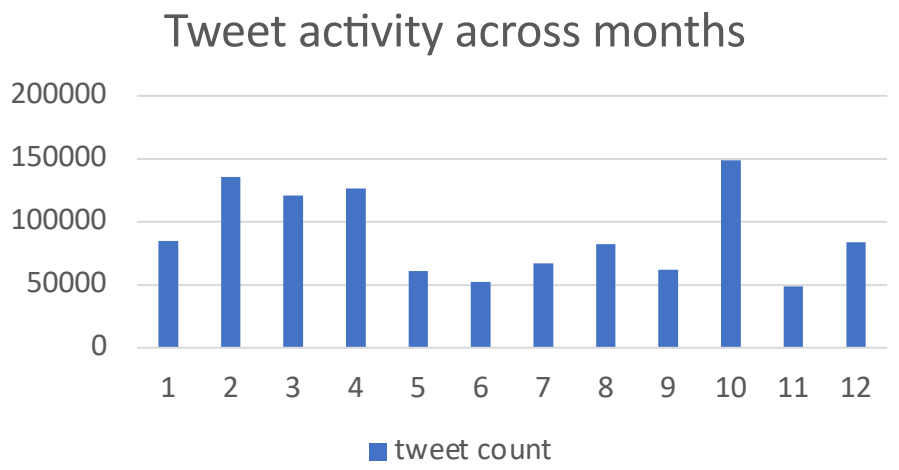


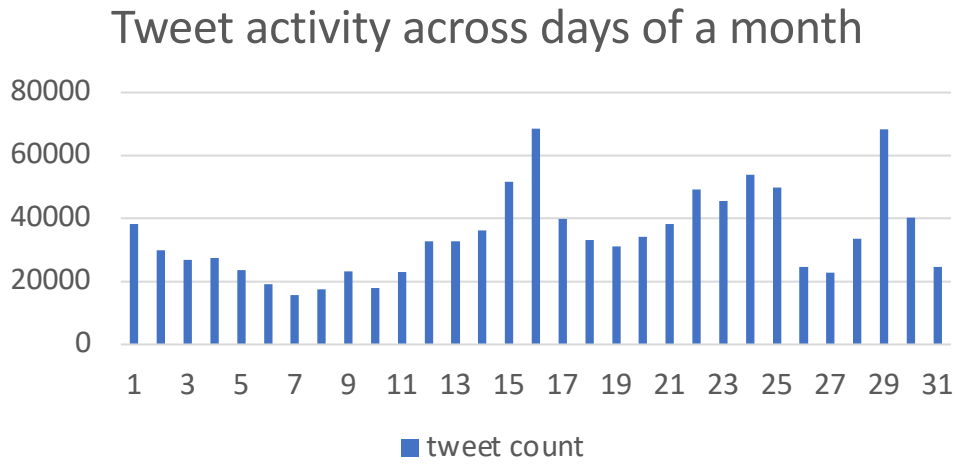


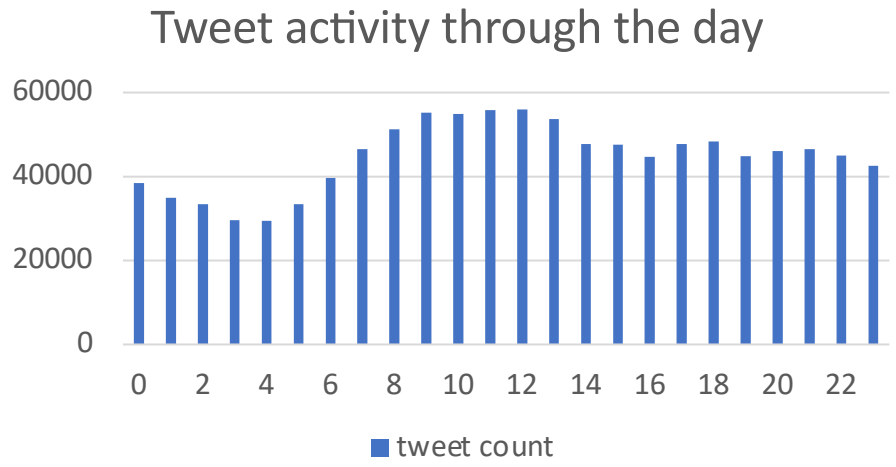
When do they tweet ? - Harvard

Twitter activity associated with **Harvard** is as follows:

- Twitter activity decreases as the month progress and there's increased activity in the month of October. **October** has the **largest twitter activity** and **November** has the **least twitter activity**
- For any given month, the twitter activity is **a bit low in the beginning and the end compared to other days**
- For any given day, the twitter activity seem to be on the rise as the day progresses. **Maximum twitter activity is observed in the noon and the same trend is maintained through out the day**



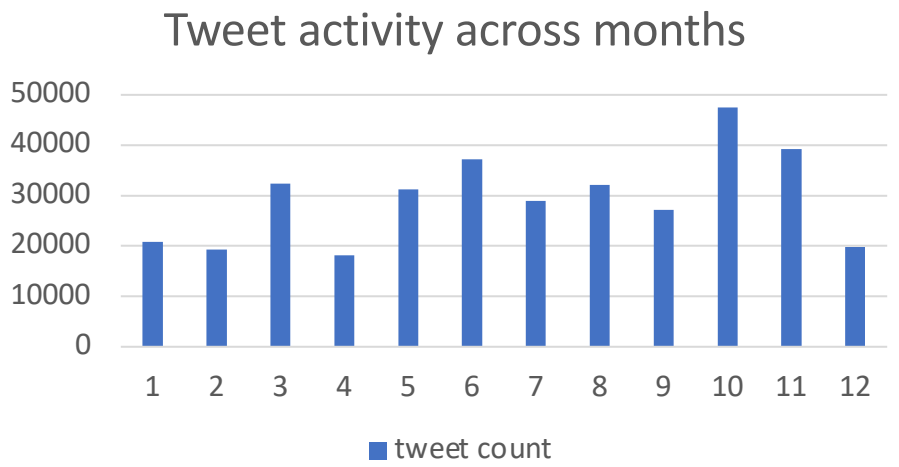


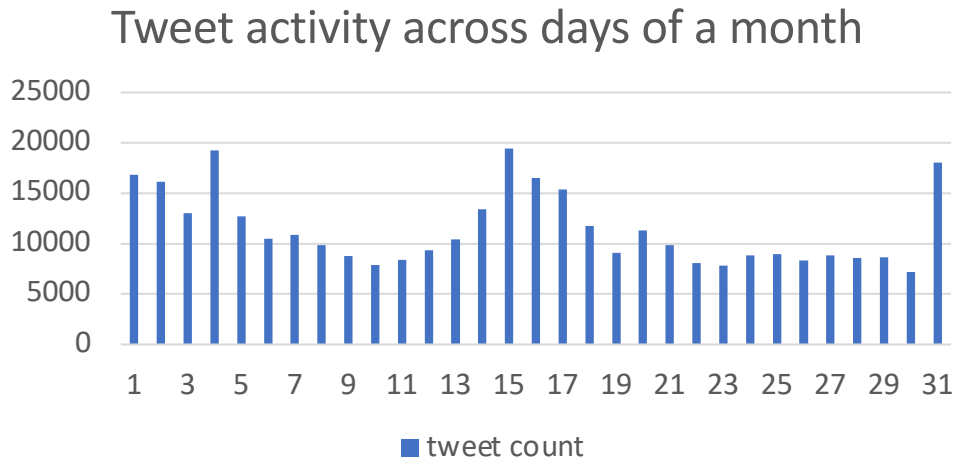


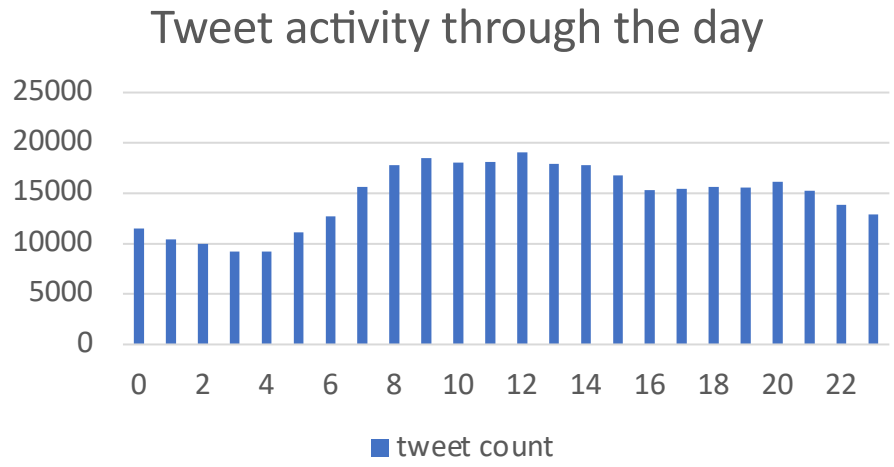
When do they tweet ? - Stanford

Twitter activity associated with **Stanford** is as follows:

- Twitter activity is on the rise as the month progresses. **October has** the **highest twitter activity** and **April** has the **least twitter activity**
- For any given month, the twitter activity is **a bit low as the month progresses, there's a spike in the middle of the month followed by decrease in the twitter activity**
- For any given day, the twitter activity seem to be on the rise as the day progresses. **Maximum twitter activity is observed in the noon and the same trend is maintained through out the day**



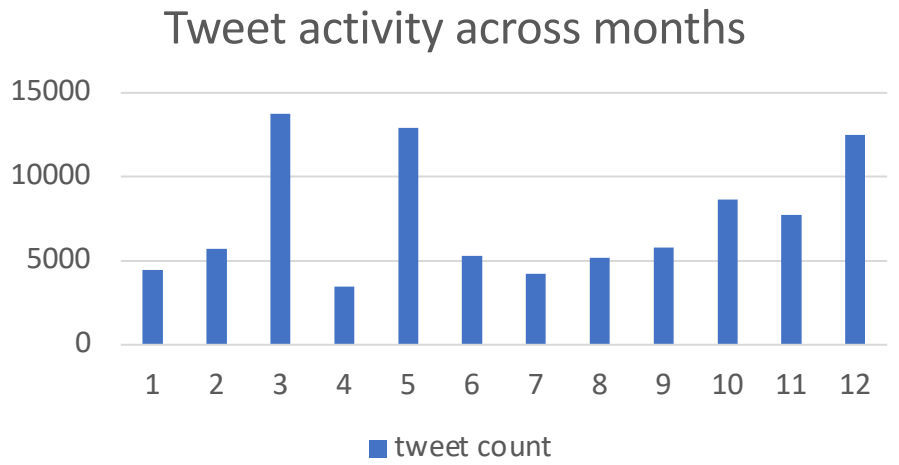


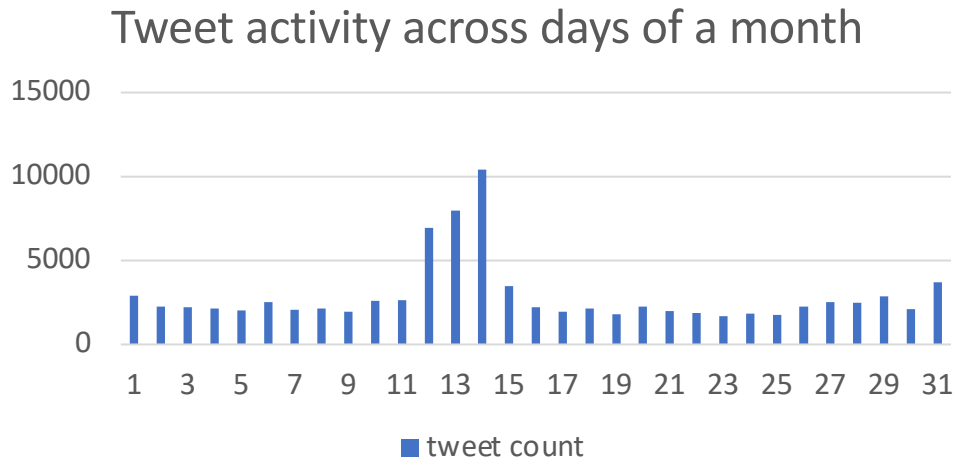


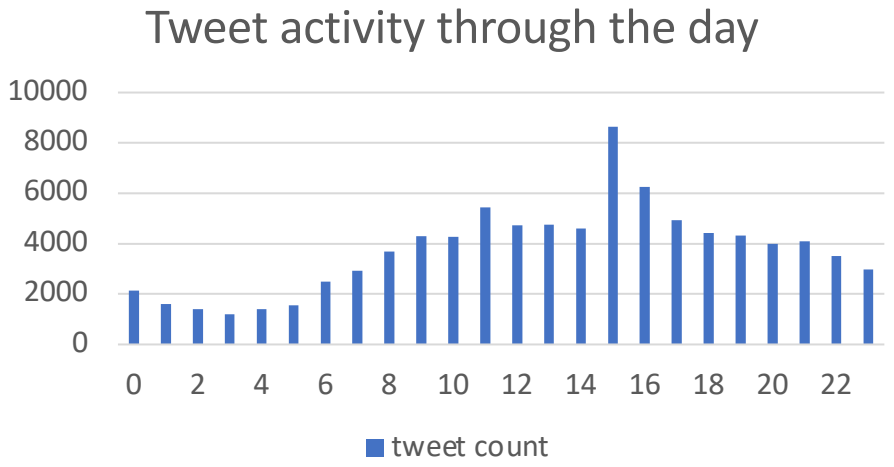
When do they tweet ? - Northwestern

Twitter activity associated with **Northwestern** is as follows:

- There isn't a definite pattern in the twitter activity. **March and May months** has the **highest twitter activity** and **April** has the **least twitter activity**
- For any given month, the twitter activity is **a bit low in the beginning and the end compared to other days with maximum activity in the middle of the month**
- For any given day, the twitter activity seem to be on the rise as the day progresses. **Maximum twitter activity is observed in the noon and the same trend is maintained through out the day**





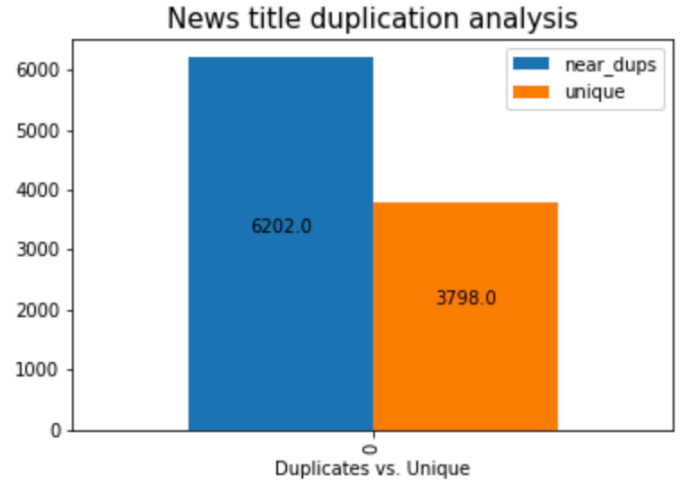


Are the tweets similar ? - UChicago

The following is a random representation of the tweets associated with **UChicago**

- Based on the tweets comparison we can conclude that **Jaccard distance of 0.3** identifies the similar text better when compared to Jaccard distance of 0.5 and 0.7
- **On further inspection of the tweets, we can observe that there's further scope to treat the data to better identify the similar tweets**
- Further, **majority of the tweets are near duplicates**

	text_A	text_B	threshold_30	threshold_50	threshold_70
10	(#law school lectures (audio) - the university of chicago law... #law https://t.co/l7izjjnsod,)	(rt @bioviniciux: arendt, hannah. the human condition. chicago: university of chicago press, 195...	Non-Dup	Duplicate	Duplicate
11	(rt @vallbb: law school lectures (audio) - the university of chicago law... https://t.co/ccgb3vq...	(rt @anthonyocampo: the university of chicago—whose endowment is $8.5 billion—won't be paying th...	Non-Dup	Duplicate	Duplicate
12	(rt @natalieymoore: the university of chicago adult level1 trauma center has been open 4 weeks. ...	(rt @markmobility: university of chicago historian @kathleen_belew: "this is an action carried o...	Non-Dup	Duplicate	Duplicate
13	(rt @jordannoying: ladies i just got accepted to the university of chicago with a scholarship an...	(nurses not allowed to return to university of chicago medical center after strike https://t.co/...	Non-Dup	Duplicate	Duplicate
14	(rt @anthonyocampo: the university of chicago—whose endowment is $8.5 billion—won't be paying th...	(rt @taniel: appalling. \n\nand fully predictable from the university of chicago, an institution...	Non-Dup	Duplicate	Duplicate
15	(rt @thecrimson: kremer, who will join both the economics and harris school of public policy fac...	(rt @uchicagoreg: lighthouse at little sable point, michigan. photo part of the university of ch...	Non-Dup	Duplicate	Duplicate
16	(rt @abc7chicago: his streak may have ended monday, but what a streak it was. he was bested by u...	(i love how the university of chicago's divinity school just got a new dean and he's a jew https...	Non-Dup	Duplicate	Duplicate
17	(rt @silentmoviegifs: sessue hayakawa originally came to america to study political economics at...	(obama foundation scholars program 2019/2020 for emerging leaders to study at the university of ...	Non-Dup	Duplicate	Duplicate
18	(does voting by mail increase the risk of voter fraud? | university of chicago news https://t.co...	(rt @nbcnews: the university of chicago bookworm, who actually wrote a master's paper featuring ...	Non-Dup	Duplicate	Duplicate
19	(rt @abc7chicago: his streak may have ended monday, but what a streak it was. he was bested by u...	(james holzhauer's jeopardy winning streak ended by university of chicago librarian emma boettch...	Non-Dup	Duplicate	Duplicate

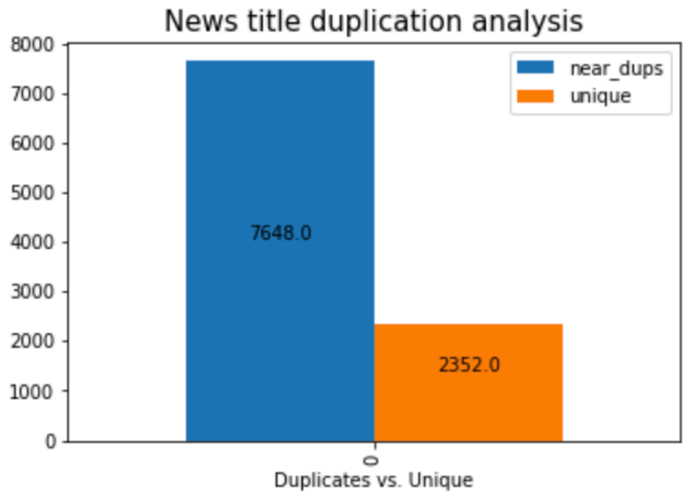


Are the tweets similar ? - Harvard

The following is a random representation of the tweets associated with **Harvard**

- Based on the tweets comparison we can conclude that **Jaccard distance of 0.3** identifies the similar text better when compared to Jaccard distance of 0.5 and 0.7
- **On further inspection of the tweets, we can observe that there's further scope to treat the data to better identify the similar tweets**
- Further, **majority of the tweets are near duplicates**

	text_A	text_B	threshold_30	threshold_50	threshold_70
10	(tribute from harvard university for professor...	(rt @wakaflocka: metrobooming was excepted to ...	Non-Dup	Duplicate	Duplicate
11	(rt @oluwaloninyo: so harvard university used ...	(rt @krauselabkcl: 3/3 4:30pm: 5 flash talks: ...	Non-Dup	Duplicate	Duplicate
12	(rt @timmythick: good evening, "timmythick" wa...	(rt @the_hindu: india-born gita gopinath, prof...	Non-Dup	Duplicate	Duplicate
13	(rt @milesgeorge8: my brother and i were just ...	(rt @theroddick: @jessekellydc @yale let's kee...	Non-Dup	Duplicate	Duplicate
14	(rt @2morrowknight: harvard university is crea...	(rt @venky_pspk55: pawan kalyan is the first t...	Non-Dup	Duplicate	Duplicate
15	(i liked a @youtube video https://t.co/a5qtpe2...	(rt @harvardesc: @harvard university police de...	Non-Dup	Duplicate	Duplicate
16	(rt @timmythick: good evening, "timmythick" wa...	(rt @brockloclegend: so excited to announce th...	Non-Dup	Duplicate	Duplicate
17	(rt @2morrowknight: harvard university is crea...	(rt @brockloclegend: so excited to announce th...	Non-Dup	Duplicate	Duplicate
18	(rt @brookebarreira: scientists at harvard uni...	(rt @adventistchurch: dr david williams, harva...	Non-Dup	Duplicate	Duplicate
19	(teen who grew up in homeless shelters earns f...	(@maua_11 harvard university,)	Non-Dup	Duplicate	Duplicate

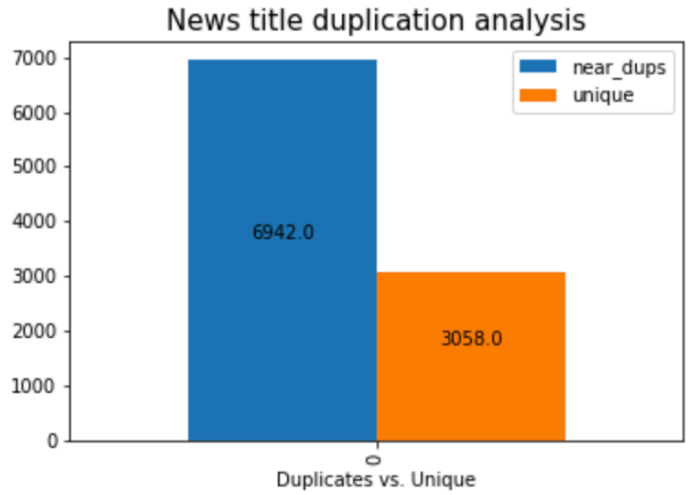


Are the tweets similar ? - Stanford

The following is a random representation of the tweets associated with **Stanford**

- Based on the tweets comparison we can conclude that **Jaccard distance of 0.3** identifies the similar text better when compared to Jaccard distance of 0.5 and 0.7
- **On further inspection of the tweets, we can observe that there's further scope to treat the data to better identify the similar tweets**
- Further, **majority of the tweets are near duplicates**

	text_A	text_B	threshold_30	threshold_50	threshold_70
10	(rt @cnnmsnbce: @josh_reif same sht in 2016 on...	(rt @natgeo: "i hope no one is shocked that bi...	Non-Dup	Duplicate	Duplicate
11	(stanford #music - stanford university | music...	(mark nicolls, md, (stanford university) provi...	Non-Dup	Duplicate	Duplicate
12	(stanford university finds that ai is outpacin...	(report: stanford university official urged fr...	Non-Dup	Duplicate	Duplicate
13	(rt @msnbc: breaking: two current stanford uni...	(rt @futbolbible: according to a stanford univ...	Non-Dup	Duplicate	Duplicate
14	(rt @craigrsawyer: when i met general "mad dog...	(#cobrakai @ stanford university https://t.c...	Non-Dup	Duplicate	Duplicate
15	(@scotgov @nicolasturgeon nobel laureate prof ...	(rt @tsinclair_: incredibly honored to receive...	Non-Dup	Duplicate	Duplicate
16	(the arrest of the teen's parents was the insp...	(rt @football__tweet: according to a stanford ...	Non-Dup	Duplicate	Duplicate
17	(rt @espnuk: according to a stanford universit...	(stanford university has one of the largest ca...	Non-Dup	Duplicate	Duplicate
18	(rt @thenewana1: neurology jobs: division chie...	(rt @ajinkyashinde18: .@dilipkpandey and @ipat...	Non-Dup	Duplicate	Duplicate
19	("this is going to save us sooo much money!" -...	(rt @setiinstitute: a new study, published in ...	Non-Dup	Duplicate	Duplicate

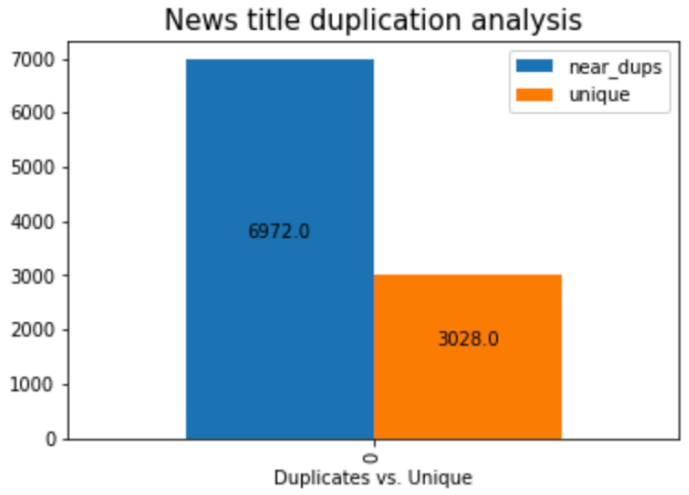


Are the tweets similar ? - Northwestern

The following is a random representation of the tweets associated with **Northwestern**

- Based on the tweets comparison we can conclude that **Jaccard distance of 0.3** identifies the similar text better when compared to Jaccard distance of 0.5 and 0.7
- **On further inspection of the tweets, we can observe that there's further scope to treat the data to better identify the similar tweets**
- Further, **majority of the tweets are near duplicates**

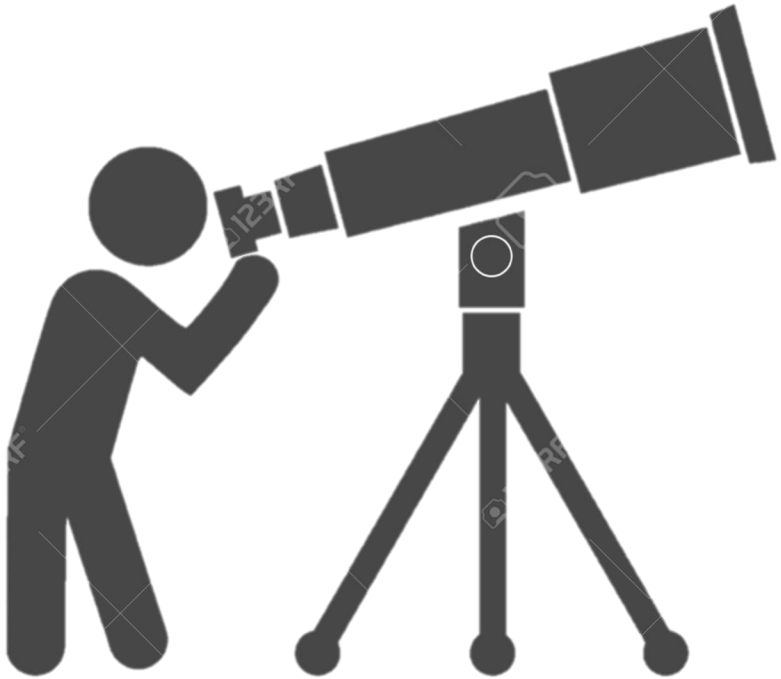
	text_A	text_B	threshold_30	threshold_50	threshold_70
10	(rt @vafootball14: extremely blessed to receiv...	(head coach of northwestern university sailing...	Duplicate	Non-Dup	Duplicate
11	(rt @michaelmagausn: a weather report \n\...	(scientists at argonne, the university of chic...	Duplicate	Non-Dup	Duplicate
12	(rt @vafootball14: extremely blessed to receiv...	(rt @wbez: northwestern university's president...	Duplicate	Non-Dup	Duplicate
13	(rt @realsaavedra: daryl morey's educational a...	(rt @proseb4bros: beyond ecstatic to announce ...	Duplicate	Non-Dup	Duplicate
14	(rt @keanukoht: blessed and honored to have re...	(rt @ebonymag: lawrence crosby, the northweste...	Duplicate	Non-Dup	Duplicate
15	(rt @realsaavedra: daryl morey's educational a...	(rt @northwesternup: books by northwestern uni...	Duplicate	Non-Dup	Duplicate
16	(he went to northwestern university on a footb...	(today at 1 p.m., dr. robert murphy of northwe...	Duplicate	Non-Dup	Duplicate
17	(rt @johnjdevine: salinas high quarterback car...	(rt @cmclymer: i completely missed the ass-who...	Non-Dup	Duplicate	Duplicate
18	(@alexofatlanta i graduated from northwestern ...	(rt @bjsorrell_: all glory goes to god, with t...	Non-Dup	Duplicate	Duplicate
19	(rt @ivan_ahernandez: apply before sunday marc...	(rt @lostblackboy: joseph epstein's wsj op-ed ...	Non-Dup	Duplicate	Duplicate



Conclusion & Future Work

Based on data and results there are couple of additional areas that can be focused

- We can perform text similarity analysis on the location variable to identify near similar locations. For example, there are couple of instances where the locations are listed as Chicago, IL and Chicago
- The mentioned places are the same, however, these are considered to be different in data collection process. Text similarity analysis can be used to identify these cases and rectify
- Also, sentiment analysis can be further employed to understand the sentiment of the tweets



Thank You