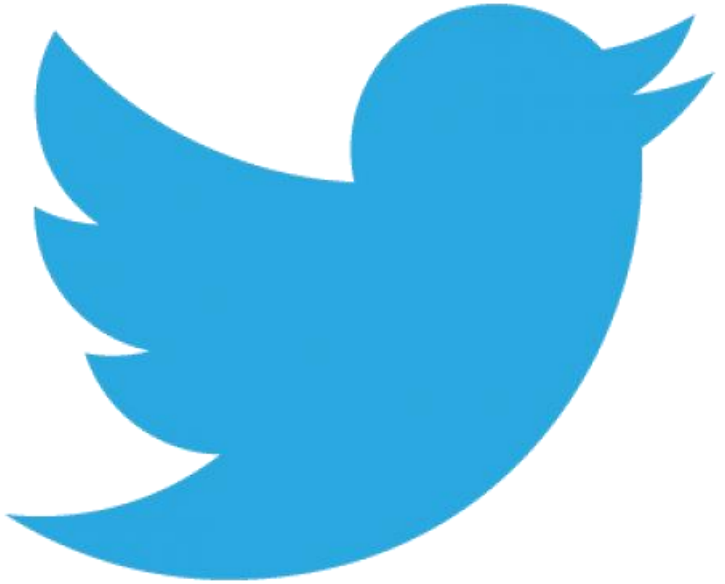
Effective Outreach Through Twitter Analysis



Jonathan Huff

Big Data Platforms, Winter 2021

Professor Nikolay Kadochnikov

Executive Summary

Twitter is a social media platform that provides a low-friction means of spreading content and ideas. To improve the effectiveness of our efforts to proliferate the awareness of our machine learning research among the academic community, we profile the Twitter users who mention related industry key terms in conjunction with the four most relevant universities for our use case: The University of Chicago, Stanford, Carnegie Mellon, and Cornell. To determine how best to allocate our resources, we have calculated several metrics broken down by university, Twitter user, and the content of the relevant tweets. As a result, we have determined which universities exhibit the most unique content and how their associated users typically engage with the platform. We conclude with actionable insight as a result of this analysis.

Methodology and Data

The dataset we will be working with is in the units of terabytes, and therefore far too large to be analyzed by conventional means. For this reason, we will leverage the power of Spark on a Hadoop cluster. Queries, data wrangling, and visualization generation will be performed in Python primarily using the Spark API, Pandas, and Seaborn. There are two distinct operations that must be performed. The initial query will filter the entire Twitter dataset spanning over three years of activity based on tweets that mention four universities: Carnegie Mellon, Cornell, Stanford, and The University of Chicago. The text will be cleaned using simple regular expression commands and further filtered for the following terms:

[*ai, artificial intelligence, big data, machine learning, data science, deep learning, analytics*]

We will then transform nested JSON data into tabular data for easier analysis. The variables we will ultimately be working with include:

Tweet_id	School	Is_reply	Favorite_count	Retweet_count	User_created_at
Tweet_date	User_favorites_count	User_followers_count	User_id	User_location	User_language
Source	User_verified	Is_quote	Is_retweet	Media	User_mentions

After analyzing these variables, we will determine the top 5 influential users per school and the top 5 originating US cities per school, among other insights. Next, we will analyze the tweet data itself to determine how unique the content is. Finally, we will distill this data into actionable insight best suited for our business case.

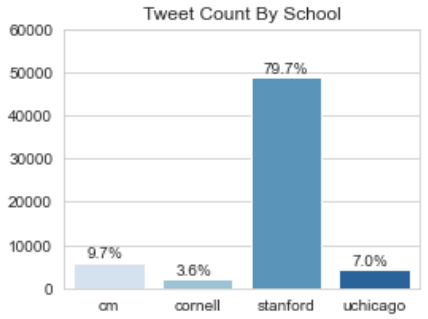
University Profiles

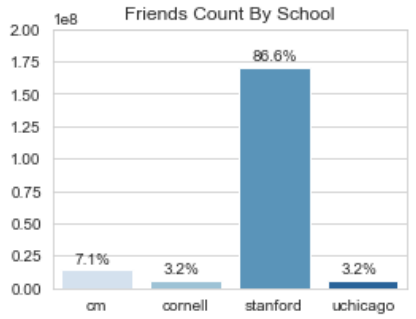
Tweet Filter Keywords:

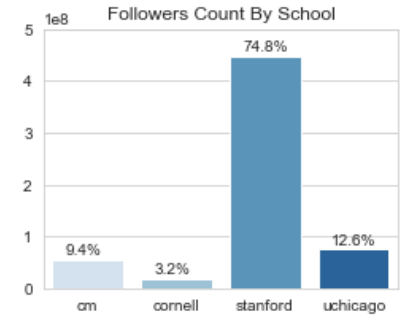
ai, artificial intelligence, big data, machine learning, data science, deep learning, analytics

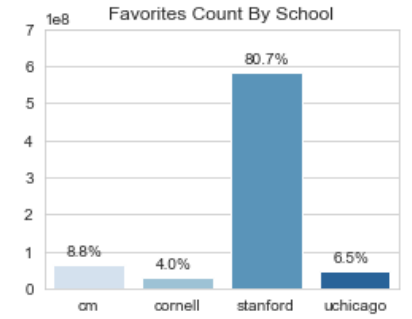
Initial investigations show that Stanford generates the vast majority of Twitter content of the four schools, filtered by the listed keywords. Cornell is on the opposite end of the spectrum, showing less platform engagement than any other school in all listed variables. Carnegie Mellon and University of Chicago are generally mixed across all categories.

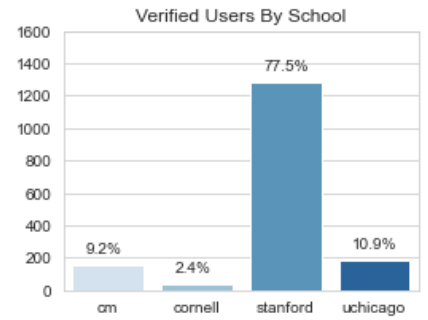
We observe the mean age of the Twitter accounts, broken down by school but there are no interesting takeaways to be had in this variable. Interestingly, due to the wide margin of Twitter activity by Stanford versus other schools, we observe the percentage of total tweets that are either retweets or quote tweets, initially expecting Stanford to be a marked leader in propagating existing content. While it is highest across the four universities, Stanford is generally in line with the others. We will later analyze the tweets themselves to check for redundant content.

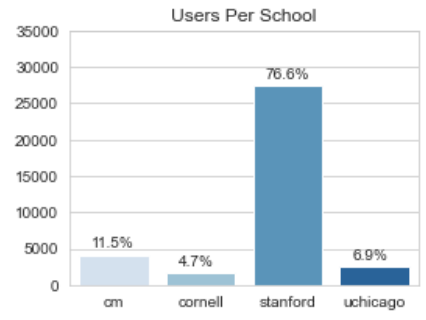


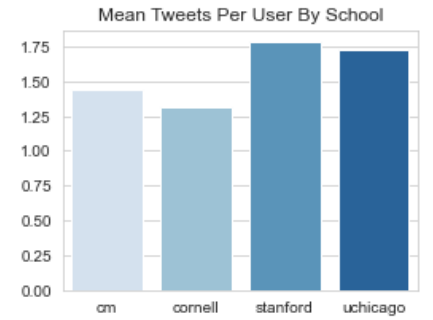


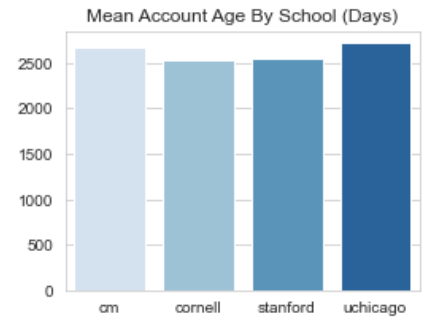


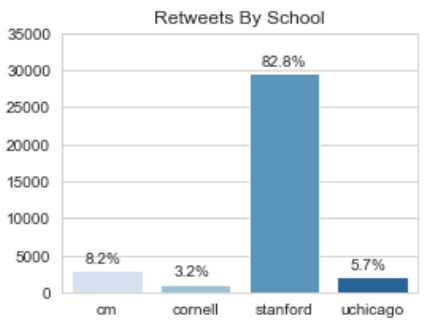


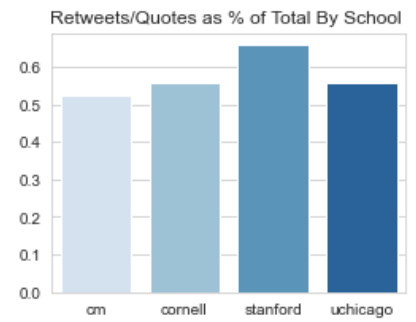


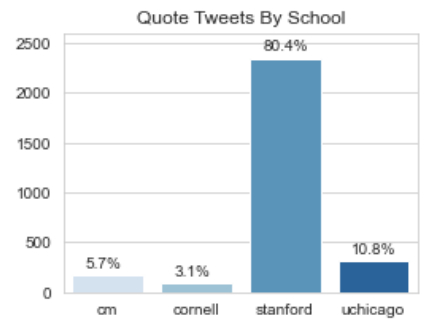












University Profiles

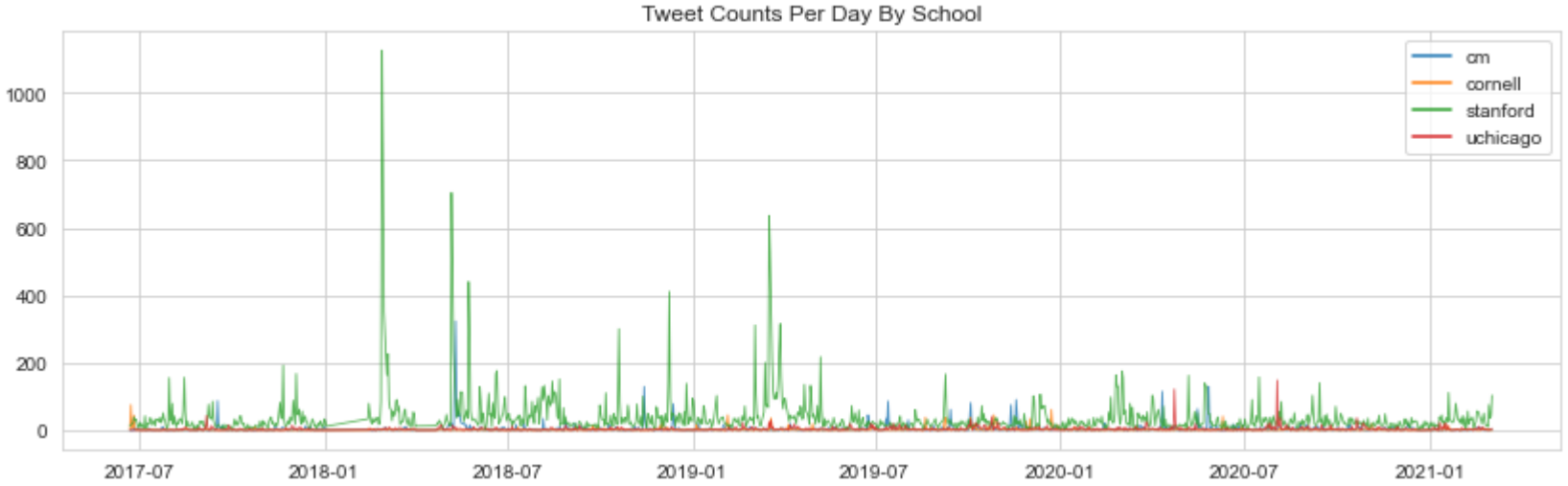
Tweet Filter Keywords:

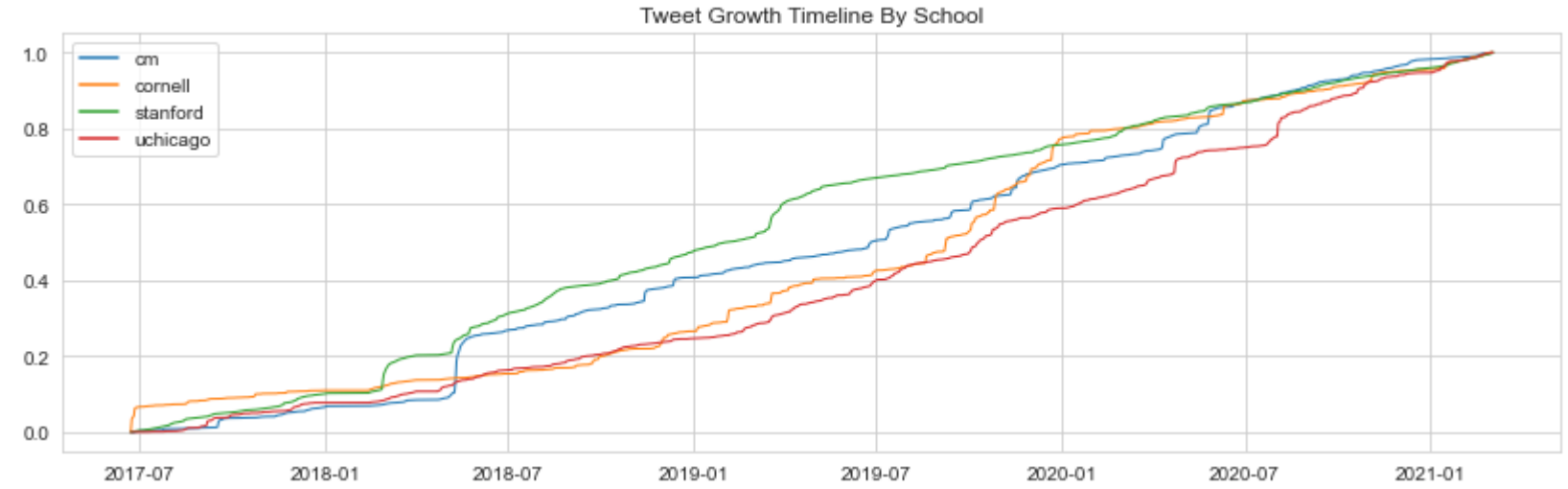
ai, artificial intelligence, big data, machine learning, data science, deep learning, analytics

Looking at the tweets per day, we see Stanford generating far more content than the other schools, as expected from the previous analysis. We observe missing data in 2018 but can assume that it does not significantly influence the dataset.

While we are curious to observe the differences in overall tweet activity by university, we also wish to explore how twitter activity is trending. The bottom plot more clearly depicts this.

Stanford contributed to the majority of its total during 2018 and 2019 but has been consistently linear at a lower growth rate from June '19 onwards. Cornell saw an exponential surge in relevant tweet activity during 2019 but peculiarly turned linear in the beginning of 2020. Carnegie Mellon witnessed a large increase in activity in 2018 but has been mostly linear since. The University of Chicago seems to be the most noteworthy, as it exhibits a slight but steady exponential growth in tweet activity. As witnessed previously, Carnegie Mellon overshadows The University of Chicago in total tweets, however this may change if the trends witnessed here continue.



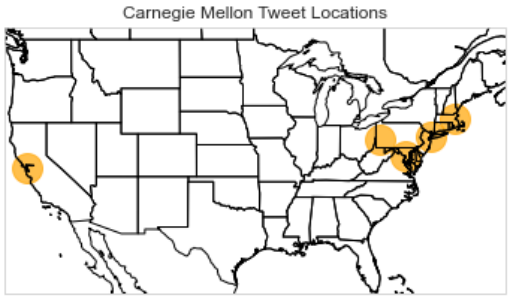


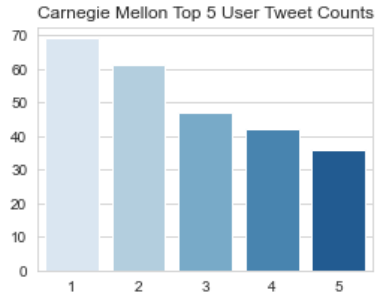
User Profiles

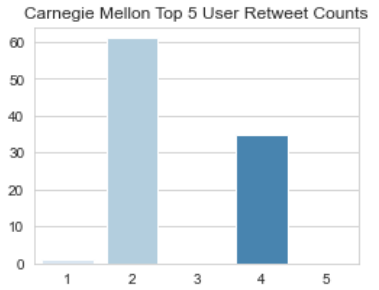
The map plots show the majority of top 5 domestic tweets originating from the east coast, primarily New York, Washington DC, and Boston. West coast cities were also popular, notably Bay Area and Los Angeles. Chicago only appeared for UChicago tweets where the vast majority of those relevant tweets originated.

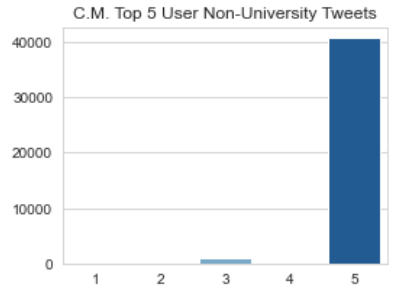
Perhaps the most striking of the data assembled here appears in the non-university related tweets. The top 5 most active twitter accounts for Stanford were relatively inactive for non-university related tweets. The other three school showed much greater non-university tweet activity, though it should be noted that Carnegie Mellon user 5 and Cornell user 1 are the same user.

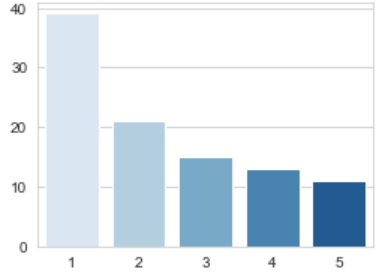
While user location data was indeed available, the majority of users did not provide locations, and so this must be considered when basing business decisions on this variable.

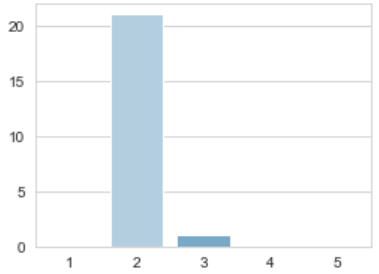


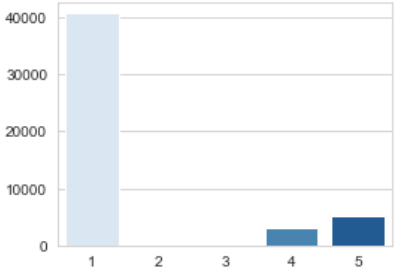


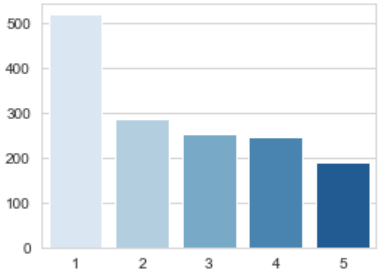


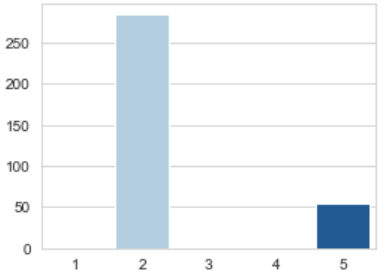


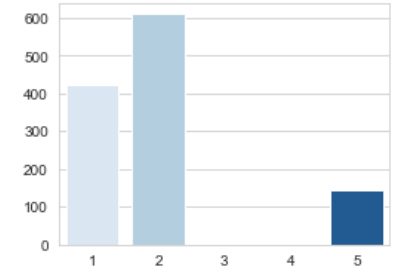


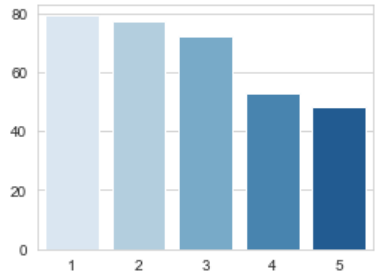


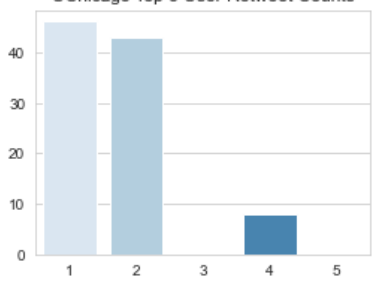


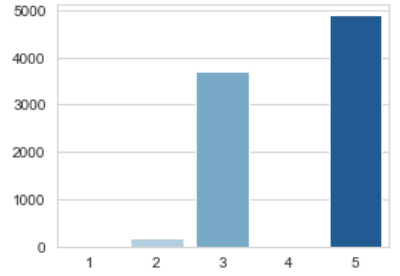










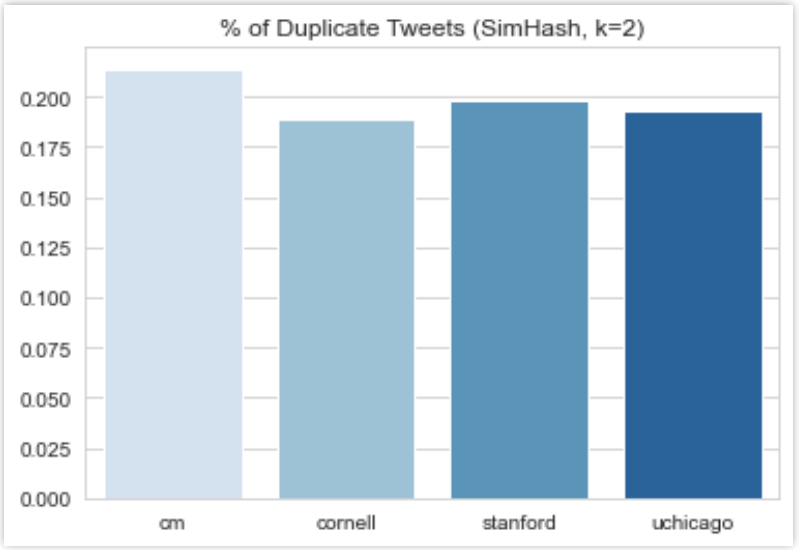


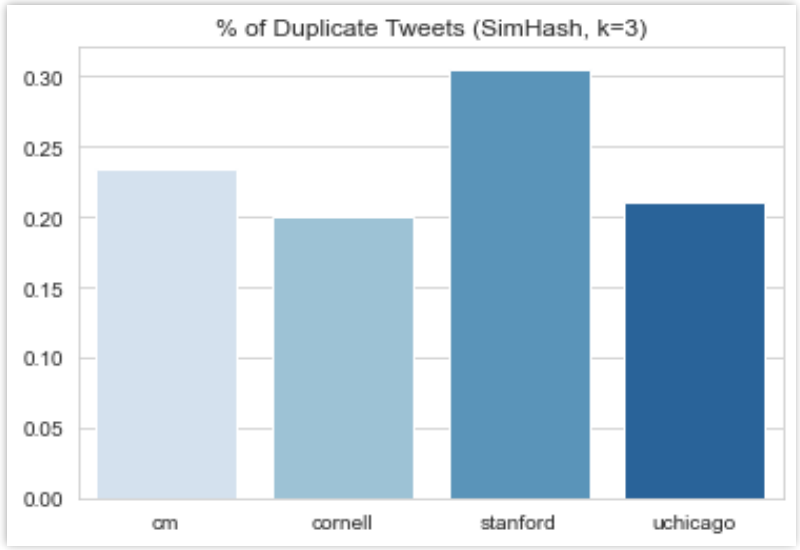
Content Analysis

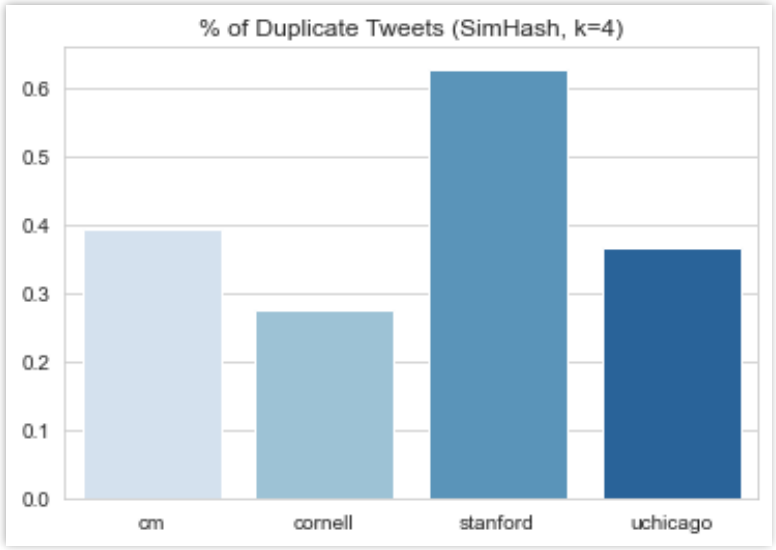
Tweet Filter Keywords:

ai, artificial intelligence, big data, machine learning, data science, deep learning, analytics

Below we can visualize the approximated percentage of total tweets containing near-identical content, broken down by school. As this is an approximation, we have included three different versions of the results returned by the algorithm for the sake of thoroughness. The takeaway here is that tweets containing “Cornell” as well as terms such as “machine learning”, “big data”, “ai”, etc. tend to be the most unique content, while those mentioning “Stanford” typically generate a much greater percentage of duplicated content.







Recommendation

After examining the Twitter data filtered for our chosen universities and keywords, we have a better understanding of how to best target our audience. Stanford's platform engagement, while exhibiting the highest level of duplicate or near-duplicate content, far exceeds every other school by a very wide margin. The redundant content is not so pervasive that it merits dismissing the significance of the tweet impressions, especially given the user base. For this reason, we can focus marketing campaigns on targeting the top Stanford users, while allocating less resources to Carnegie Mellon and University of Chicago. At this point in time, given Cornell's meager Twitter engagement, we do not see added value in targeting those users.

Geographically, focusing in-person events on the west coast is prudent given Stanford's user base. Alternatively, Boston was present in the top 5 locations across all four universities, presumably due to institutions such as Harvard and MIT. For this reason, it makes sense to focus on events here as well, but it is imperative to track future Twitter activity as The University of Chicago may continue to trend higher, and therefore shifting resources to those users would be justified. Future analysis will increase the breadth of the universities considered.

Appendix

- Due to space constraint, the top 5 user IDs were omitted. Below are the corresponding IDs for labels 1-5 on the charts on page 6:

Label	Carnegie Mellon User_ID
1	60258204
2	1257170815881228288
3	13584132
4	756040164149977088
5	3149729430

Label	Cornell User_ID
1	3149729430
2	1257170815881228288
3	1001229541585686528
4	3243551873
5	614974325

Label	Stanford User_ID
1	4074372137
2	841437061
3	844456269222232064
4	844458707178139648
5	718175074537177088

Label	UChicago User_ID
1	969234723619966976
2	772170180
3	24698230
4	741636827178799104
5	929740814657441792

- Below are the count values for the top 5 cities per school

Carnegie Mellon City	Count
Pittsburgh	333
New York	145
San Francisco	104
Washington, DC	76
Boston	49

Cornell City	Count
New York	82
Ithaca	48
Boston	17
Los Angeles	22
San Francisco	15

Stanford City	Count
Stanford	1238
San Francisco	1263
Palo Alto	596
New York	1070
Boston	480

UChicago City	Count
Chicago	1025
Los Angeles	65
New York	82
Washington	90
Boston	42