Education Tweets Analysis

Chen Feng Tsai
SID: 12372316

Executive Summary

This analysis report is to analyze the patterns of tweets in education category. There are some points to be highlighted.

First, the most prolific twitterers in this category are sports related users, and after using the influential score as metrics, political entities and news oulets are two main influential source in the whole category, especially in K-12 Topic.

Second, most of the tweets come from the main city in the US, and the ranking is consistent with population ranking. For geogrpahical progression in Florida Math Book Ban, there is an increasing trend from the east coast to the west coast.

Third, there are gaps between Decemebr to March, and the peaks happen at October. Reason behind this might be school semester system and election.

Fourth, ⅕ of the tweets in Florida Math Book Ban are duplicate.

In conclusion, tweets are credible source of information since it correctly points out the author identification and where the topics emerge and progress, and imply further inference of relationship with timeline. Furthermore, most of the tweets are unique, which makes it more credible.

Methodology

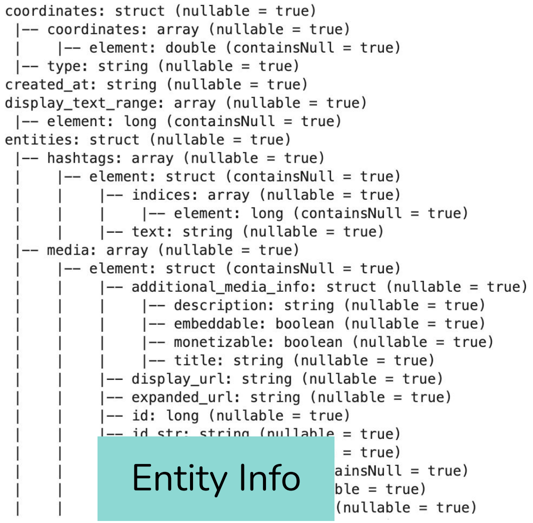
Analyzing tweet data in education to discover patterns from multiple aspects:

1. Author Identification
    a. To see who are the twitterers that post most about education and are most influential
    b. Extract information about the top prolific twitterers to cateogrize their identification
    c. Design metrics to calculate influential score for each twitterers and types of organizations
2. Location Analysis
    a. To see where are these tweets from and is their any patterns in these locations
    b. Compare the distributions from country perspectives and city perspectives
    c. Explore the geographical patterns by time using respective distribution plot
3. Timeline Analysis
    a. To see whether there are seasonality in these education tweets
    b. Compare the aggregate tweet counts by month
4. Message Uniqueness Analysis
    a. To explore whether people are just copying the same text
    b. Using Jaccard similarity to extract duplication
    c. Compare the duplication between each types of twitterers

Source Data Overview

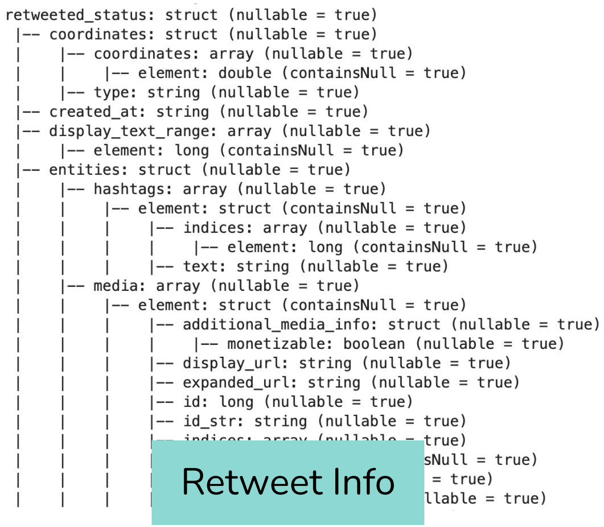
The source data is formed of million tweet objects with json format and nested structure that includes retweet and quoted information.

There are three main categories in the structure:
The first one is basic tweet information, including the user information (follower count) and extended entities (full text)
The second one is retweeted status information, including retweeted count and retweeted source
The third one is quoted status information, which is similar to retweet



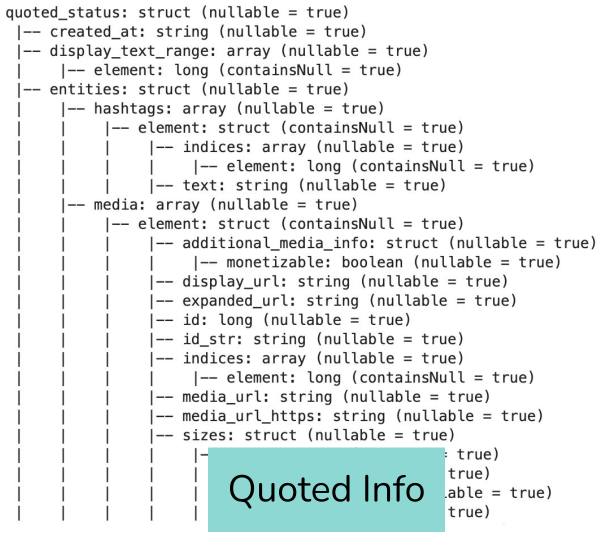
```
coordinates: struct (nullable = true)
 |-- coordinates: array (nullable = true)
 |    |-- element: double (containsNull = true)
 |-- type: string (nullable = true)
created_at: string (nullable = true)
display_text_range: array (nullable = true)
 |-- element: long (containsNull = true)
entities: struct (nullable = true)
 |-- hashtags: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- indices: array (nullable = true)
 |    |    |    |-- element: long (containsNull = true)
 |    |    |-- text: string (nullable = true)
 |-- media: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |-- description: string (nullable = true)
 |    |    |    |-- embeddable: boolean (nullable = true)
 |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |-- title: string (nullable = true)
 |    |    |-- display_url: string (nullable = true)
 |    |    |-- expanded_url: string (nullable = true)
 |    |    |-- id: long (nullable = true)
 |    |    |-- id_str: string (nullable = true)
 |    |                          = true)
 |    |                    ainsNull = true)
 |    |                    ble = true)
 |    |                    (nullable = true)
```

Entity Info



```
retweeted_status: struct (nullable = true)
 |-- coordinates: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- created_at: string (nullable = true)
 |-- display_text_range: array (nullable = true)
 |    |-- element: long (containsNull = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- text: string (nullable = true)
 |    |-- media: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |                 sNull = true)
 |    |    |    |                  = true)
 |    |    |    |                 llable = true)
```

Retweet Info



```
quoted_status: struct (nullable = true)
 |-- created_at: string (nullable = true)
 |-- display_text_range: array (nullable = true)
 |    |-- element: long (containsNull = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- text: string (nullable = true)
 |    |-- media: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |-             : true)
 |    |    |    |    |               true)
 |    |    |    |    |              able = true)
 |    |    |    |    |               true)
```

Quoted Info

Tweet Clean-up and Filtering

To filter out irrelevant tweets, I count the occurrence of the words (excluding stopwords) to find the keywords in the education category. The table below shows the top 10 words by their occurrences.

school	39735231	like	4835983
college	11044189	students	4739795
high	8461610	kids	3665067
university	8139660	professor	3507005
schools	6468330	back	2915399

After looking at the occurrences of keywords and the specific topic in education, I select 9 words to be kept, “high”, “college”, “university”, “secondary”, “primary”, “education”, “k-12”, “undergraduate”, “graduate”. I filter out data that does not contain these words.

EDA

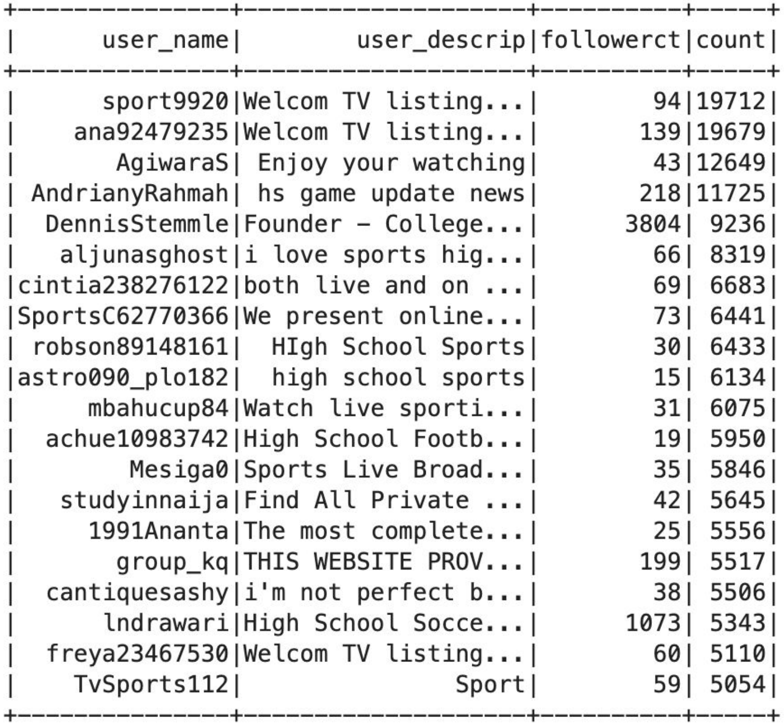
To profile twitterers, I pick

1. user_id: under the user information and unique for searching
2. user_name: under the user information and useful for identification
3. user_descrip: under the user information and useful for identification
4. user_loc: under the user information and useful in location analysis
5. user_followerct: under the user information and useful in influential score
6. created_at: under the basic information of the tweet and useful for timeline analysis
7. tweet_id: under the basic information of the tweet and unique
8. retweeted_from: under the basic information of the tweet and useful for influential score
9. retweet_ct: under the retweeted_status layer and gives the retweet count when it is being retweeted (the one under basic information is not correct)
10. retweet: under the retweeted_status layer and gives the indicators of whether it is retweeted. It can be used to categorize original tweet and retweet
11. full_text: under the extended_tweet layer to have the untruncated text (I aggregate text and full_text together to make data more comprehensive)

Author Identification - Most Prolific Twitterers

Most of the most prolific twitters are related to sports, which infers that higher tweet volumes are not reflective of the emergence of new hot topic in education.

Besides is the table for user info and their tweet count.
Next slide is the distribution for how many tweets user post.

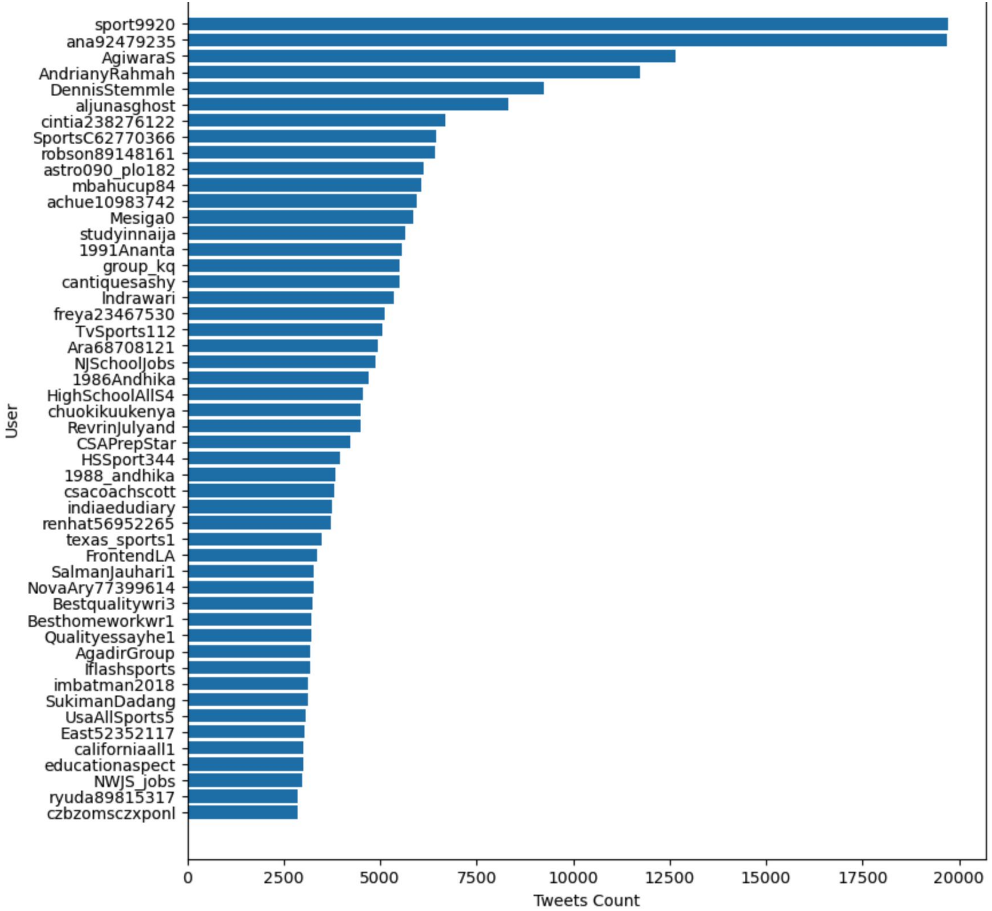


```
+---------------+--------------------+----------+-----+
|      user_name|        user_descrip|followerct|count|
+---------------+--------------------+----------+-----+
|      sport9920|Welcom TV listing...|        94|19712|
|    ana92479235|Welcom TV listing...|       139|19679|
|       AgiwaraS| Enjoy your watching|        43|12649|
| AndrianyRahmah| hs game update news|       218|11725|
|  DennisStemmle|Founder - College...|      3804| 9236|
|   aljunasghost|i love sports hig...|        66| 8319|
|cintia238276122|both live and on ...|        69| 6683|
|SportsC62770366|We present online...|        73| 6441|
| robson89148161|  HIgh School Sports|        30| 6433|
|astro090_plo182|  high school sports|        15| 6134|
|     mbahucup84|Watch live sporti...|        31| 6075|
|  achue10983742|High School Footb...|        19| 5950|
|        Mesiga0|Sports Live Broad...|        35| 5846|
|   studyinnaija|Find All Private ...|        42| 5645|
|     1991Ananta|The most complete...|        25| 5556|
|       group_kq|THIS WEBSITE PROV...|       199| 5517|
|  cantiquesashy|i'm not perfect b...|        38| 5506|
|      lndrawari|High School Socce...|      1073| 5343|
|  freya23467530|Welcom TV listing...|        60| 5110|
|    TvSports112|               Sport|        59| 5054|
+---------------+--------------------+----------+-----+
```

From the distribution, we can see that most of them are sport related users such as sport 9920, sportsc62770366......

It might result from the fact that sport game is happening everyday so there will be post about each game everyday. Therefore, the amount of the post about sport has a big ratio in the dataset.

Top 50 Prolific Twitterers by their Tweet Volume



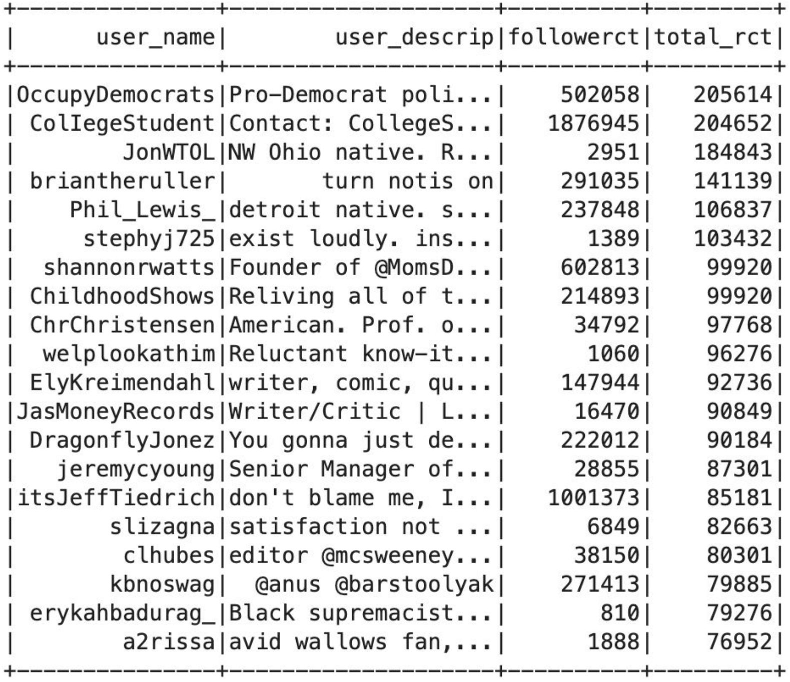
Author Identification - Most Retweeted Twitterers

Most of them are social media influencers, but retweeted_ct is not a good indicators to know how influential the users are since some of them have low follower counts.

The reason that the users have lots of retweet count is that they post a lot or the tweet is awhile ago and have certain degree of accumulation.

Besides is the table for user info and their tweet count.

Next slide is the distribution of how many retweets each user has.

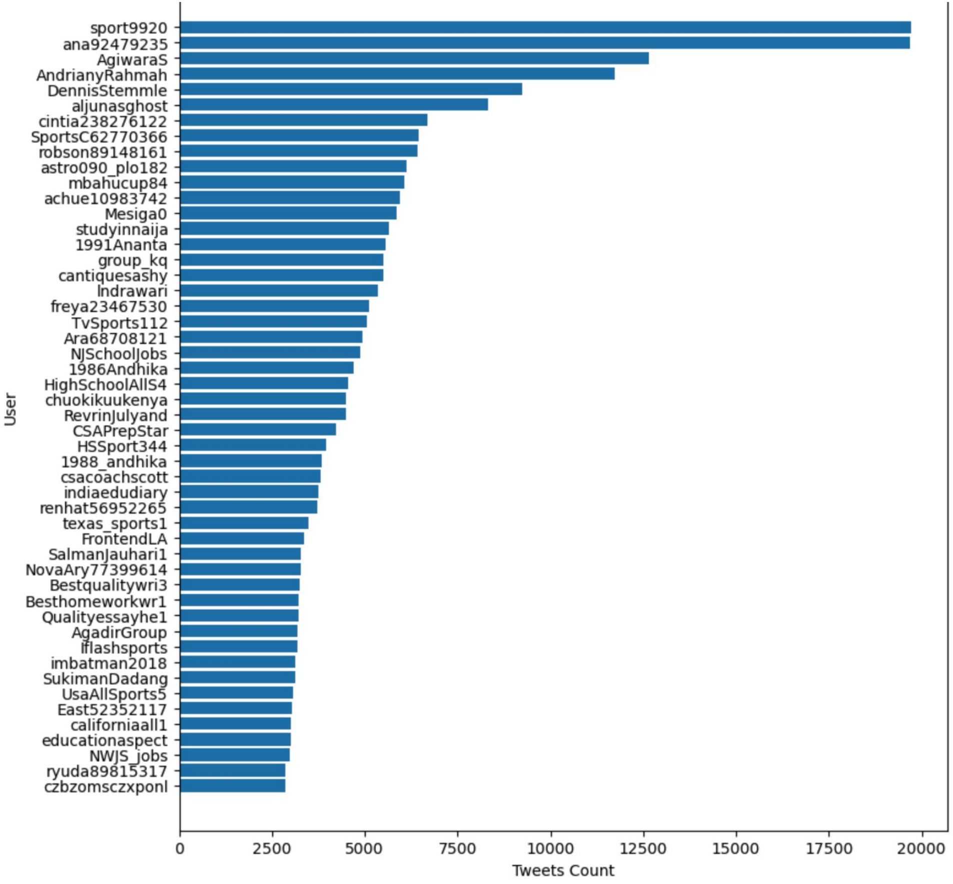


```
+---------------+--------------------+----------+---------+
|      user_name|        user_descrip|followerct|total_rct|
+---------------+--------------------+----------+---------+
|OccupyDemocrats|Pro-Democrat poli...|    502058|   205614|
| ColIegeStudent|Contact: CollegeS...|   1876945|   204652|
|        JonWTOL|NW Ohio native. R...|      2951|   184843|
| briantheruller|       turn notis on|    291035|   141139|
|    Phil_Lewis_|detroit native. s...|    237848|   106837|
|     stephyj725|exist loudly. ins...|      1389|   103432|
|  shannonrwatts|Founder of @MomsD...|    602813|    99920|
| ChildhoodShows|Reliving all of t...|    214893|    99920|
| ChrChristensen|American. Prof. o...|     34792|    97768|
|  welplookathim|Reluctant know-it...|      1060|    96276|
| ElyKreimendahl|writer, comic, qu...|    147944|    92736|
|JasMoneyRecords|Writer/Critic | L...|     16470|    90849|
| DragonflyJonez|You gonna just de...|    222012|    90184|
|   jeremycyoung|Senior Manager of...|     28855|    87301|
|itsJeffTiedrich|don't blame me, I...|   1001373|    85181|
|       slizagna|satisfaction not ...|      6849|    82663|
|        clhubes|editor @mcsweeney...|     38150|    80301|
|       kbnoswag|  @anus @barstoolyak|    271413|    79885|
| erykahbadurag_|Black supremacist...|       810|    79276|
|        a2rissa|avid wallows fan,...|      1888|    76952|
+---------------+--------------------+----------+---------+
```

Categories in the Top 50 retweeted users:

Social media influencer: 22
Political Entity: 6
Reporter/ News: 5
Education Worker: 4
Sport: 4
Other: 9

Top 50 Twitterers by their Retweeted Volume

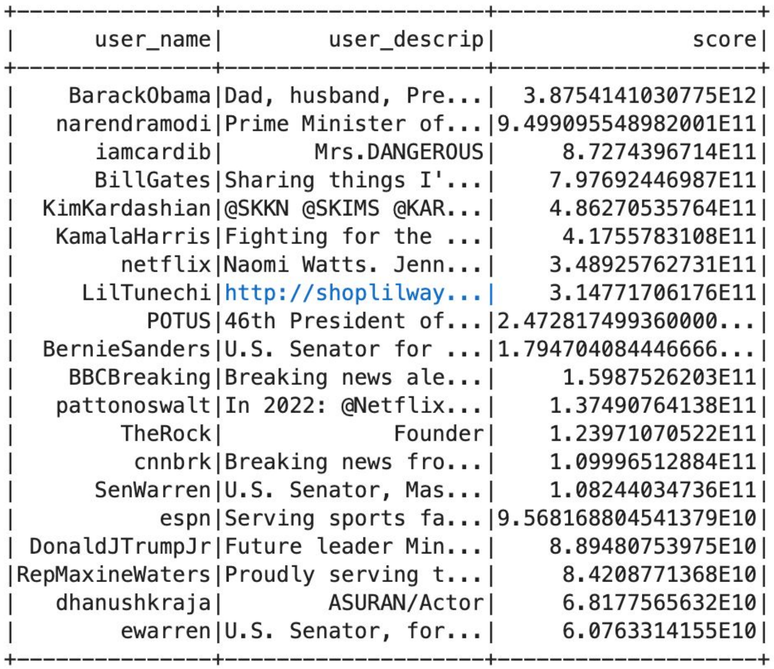


Author Identification - Most Influential Twitterers

To make a more comprehensive analysis, a new metrics is used, which is “influential score”.
Influential Score Formula = (retweet_ct)/(total_tweet_ct)*(follower_ct)

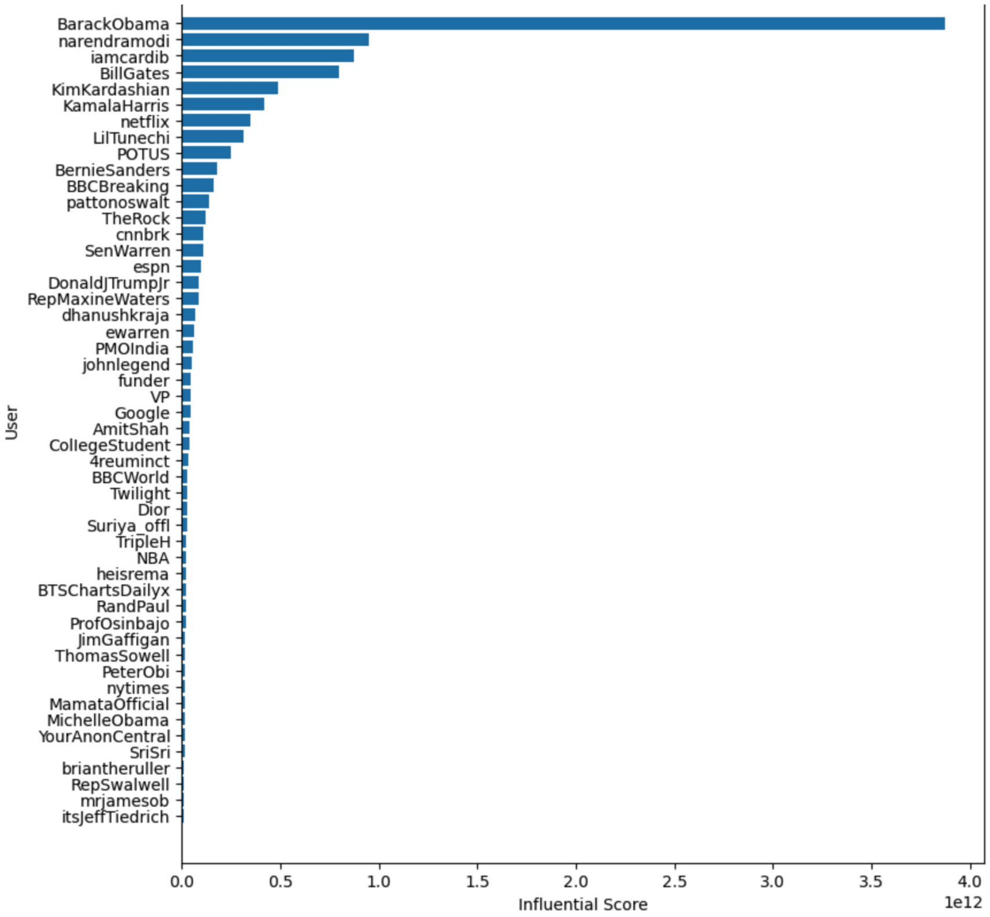
Categories in the Top 50 users:
Social media influencer: 15
Political Entity: 17
Reporter/News: 8
Education Worker: 3
Sport: 3
Other: 4

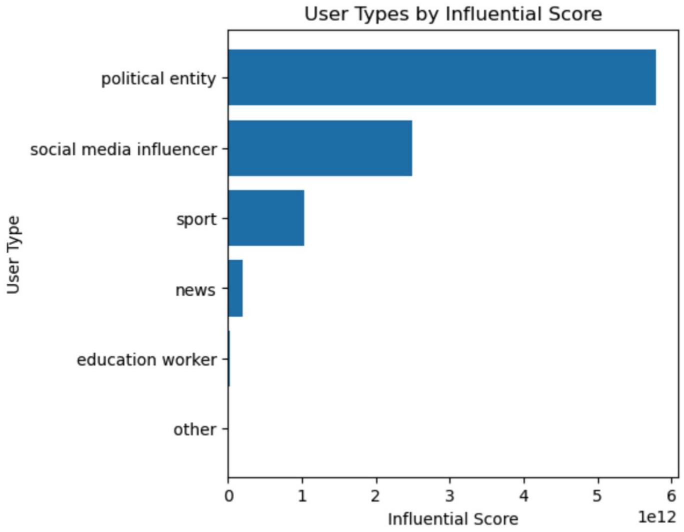
Political entities are the biggest group that tweet about education with big influence, then the social media influencer.



```
+---------------+--------------------+--------------------+
|      user_name|        user_descrip|               score|
+---------------+--------------------+--------------------+
|    BarackObama|Dad, husband, Pre...|  3.8754141030775E12|
|   narendramodi|Prime Minister of...|9.499095548982001E11|
|      iamcardib|       Mrs.DANGEROUS|     8.7274396714E11|
|      BillGates|Sharing things I'...|    7.97692446987E11|
|  KimKardashian|@SKKN @SKIMS @KAR...|    4.86270535764E11|
|   KamalaHarris|Fighting for the ...|     4.1755783108E11|
|        netflix|Naomi Watts. Jenn...|    3.48925762731E11|
|     LilTunechi|http://shoplilway...|    3.14771706176E11|
|          POTUS|46th President of...|2.47281749936000...|
|  BernieSanders|U.S. Senator for ...|1.794704084446666...|
|    BBCBreaking|Breaking news ale...|     1.5987526203E11|
|   pattonoswalt|In 2022: @Netflix...|    1.37490764138E11|
|        TheRock|             Founder|    1.23971070522E11|
|         cnnbrk|Breaking news fro...|    1.09996512884E11|
|      SenWarren|U.S. Senator, Mas...|    1.08244034736E11|
|           espn|Serving sports fa...|9.568168804541379E10|
| DonaldJTrumpJr|Future leader Min...|    8.89480753975E10|
|RepMaxineWaters|Proudly serving t...|     8.4208771368E10|
|   dhanushkraja|        ASURAN/Actor|     6.8177565632E10|
|        ewarren|U.S. Senator, for...|     6.0763314155E10|
+---------------+--------------------+--------------------+
```

Top 50 Twitterers by Influential Score





We can see that after using the influential score, the ratio of political entites increase, which are reasonable to infer that Senate and government orginization have large influential power in the education topic.

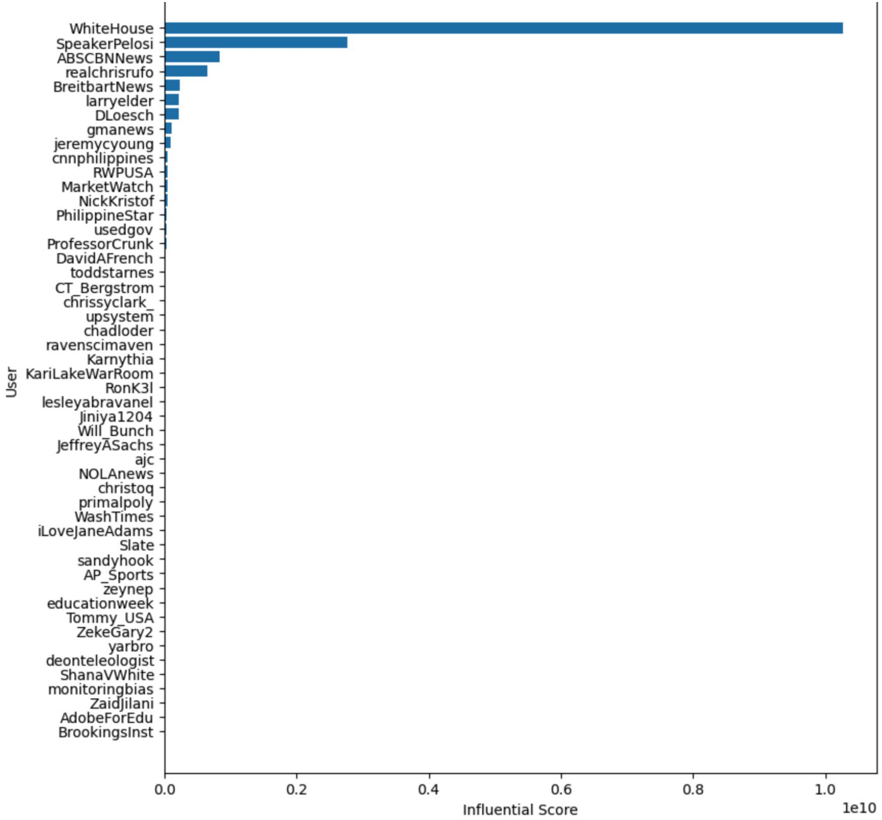
Author Identification in Topic K-12

Categories in the Top 50 users in topic K-12 :

Social media influencer: 13
Political Entity: 12
Reporter/News: 17
Education Worker: 5
Sport: 1
Other: 4

Most of the influential twitterers that tweet about K-12 are News, journalists, and political entites.

They occupied ⅗ of the list.



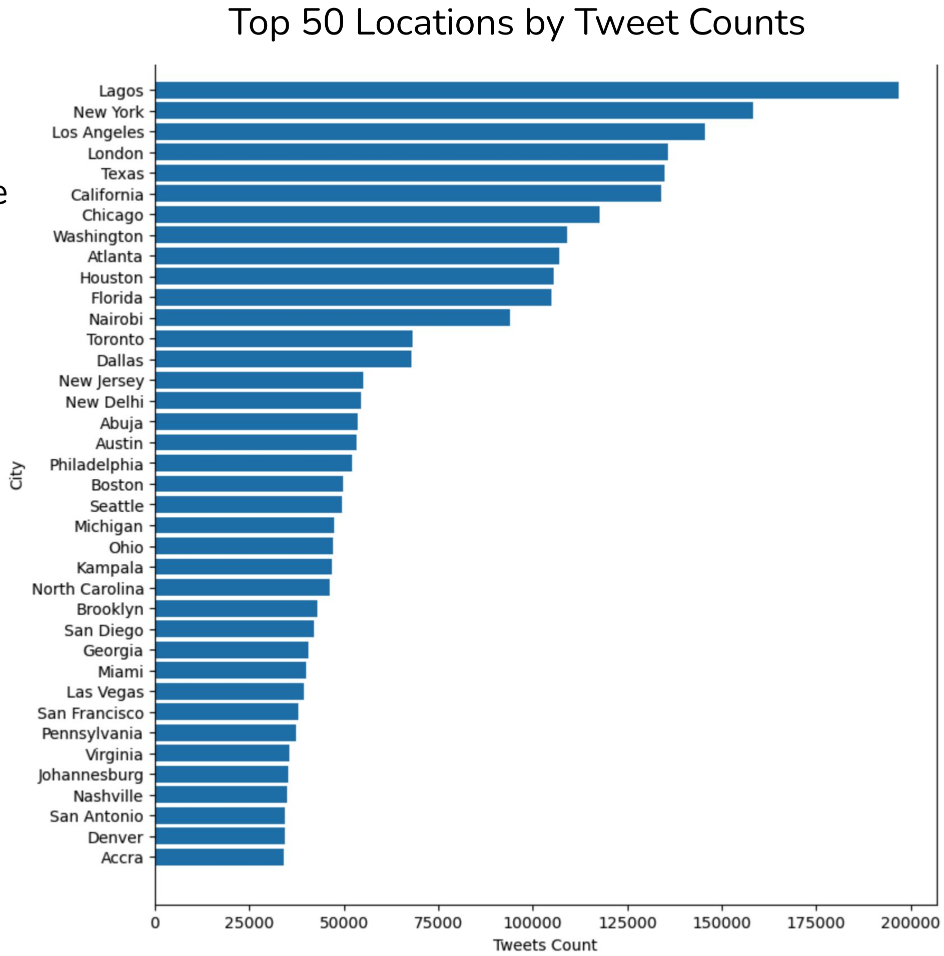
Top 50 Twitterers by Influential Score in K12

Location Analysis

I filter out data that only has country informaiton since I will like to analysis to be in a city or state scale.
Besides is the distribution of tweets count in each city.

I realize that it has the same distribution as population ranking with NY the most then LA, Chicago and Houston, Dellas, Miami. Most of them on the list are from main city, and they occupied the top list of this ranking. Then the second tier are some cities closed by the main city, such as Michigan, Boston, San Diego, Seattle.

Some of the active twitter users are from UK, Africa(Nigeria, South Africa), India


Top 50 Locations by Tweet Counts

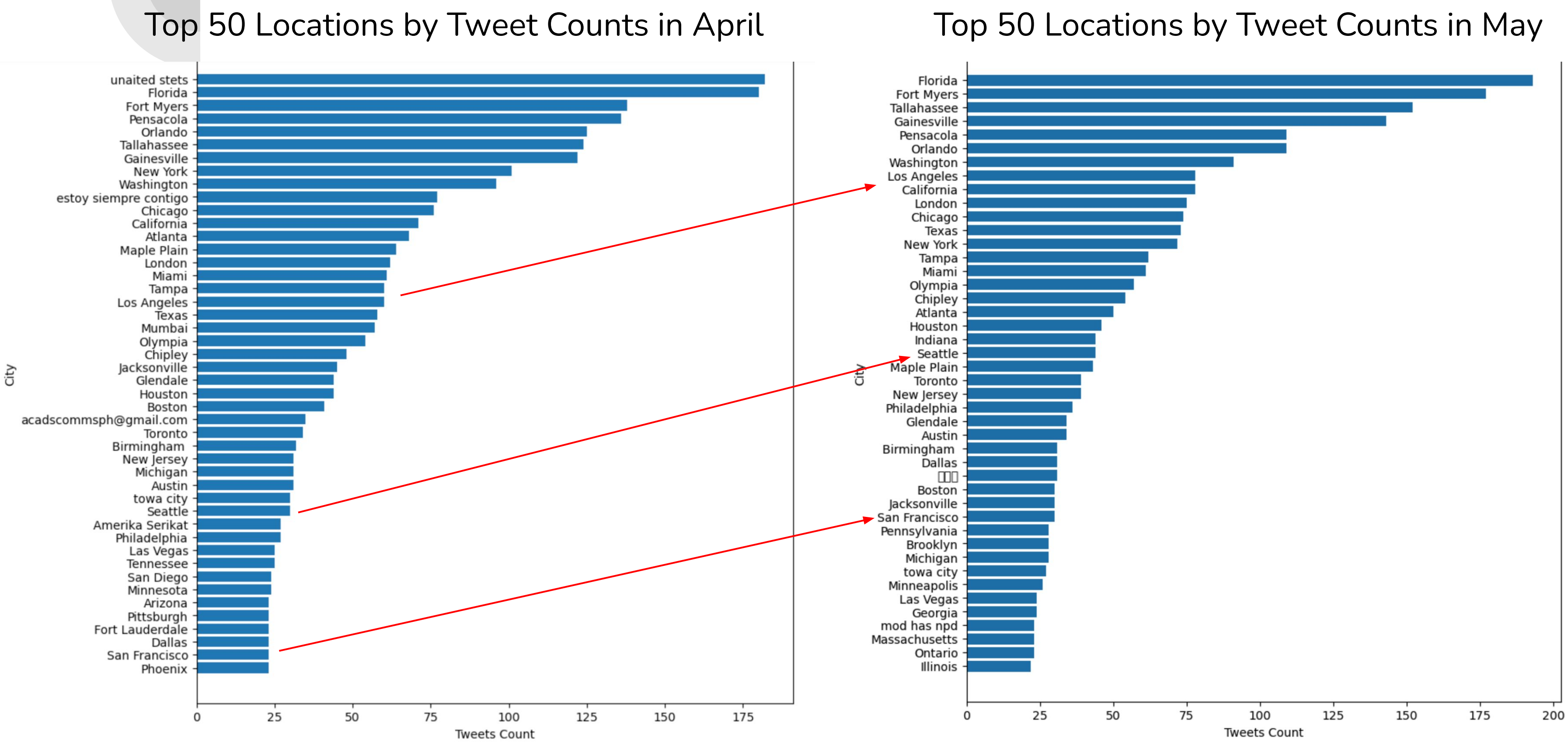
Topic Geo Progression Analysis

To discover how the topic is spread, we need to compare how different the amount of tweets in each location changed. Extracting tweets about “Florida Math Book Ban”, I visualize two geographical distributions in April and May in the next slide to see how it spreads. I choose thes months since most of the news are in these months.

From the distribution in April, most of the high rank position is occupied by cities in south east US such as Miami, Fort Myers. After compare it with distribution in May, we can see that some of the cities in the west has higher rank such as Los Angelas, Seattle, San Francisco.

Therefore, I infer that there is a topic progression from the east to the west in the US.

Distribution Comparison in Florida Math Book Ban

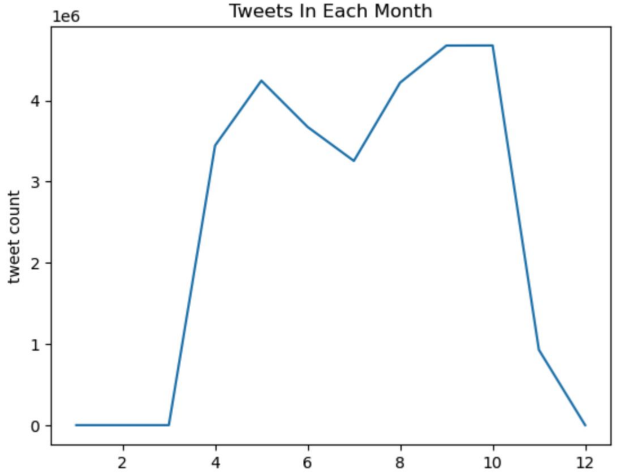


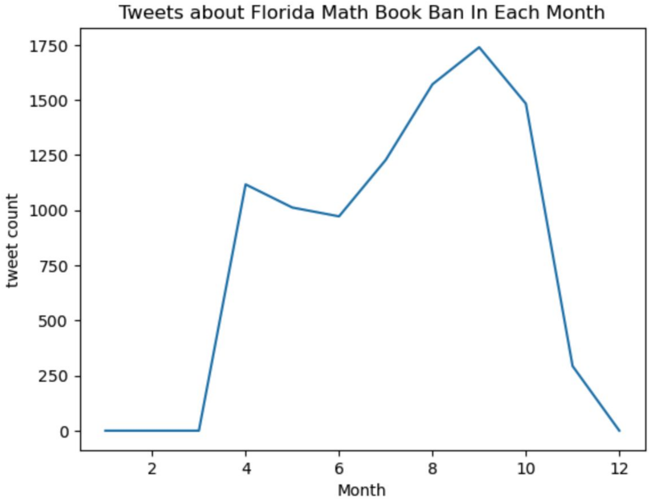
Top 50 Locations by Tweet Counts in April

Top 50 Locations by Tweet Counts in May

Timeline Analysis

There is a valley between December to March, and relatively low value in November and July.
There are two peaks in Spring and Autumn.
The reason that there exist peaks and valleys might because lots of schools are semester system and only have classes during Spring and Autumn.
Most of the school activities happen in these two seasons, which increases the discussions on twitter.
Another possible reason that October is the highest might be that the election takes place at November.



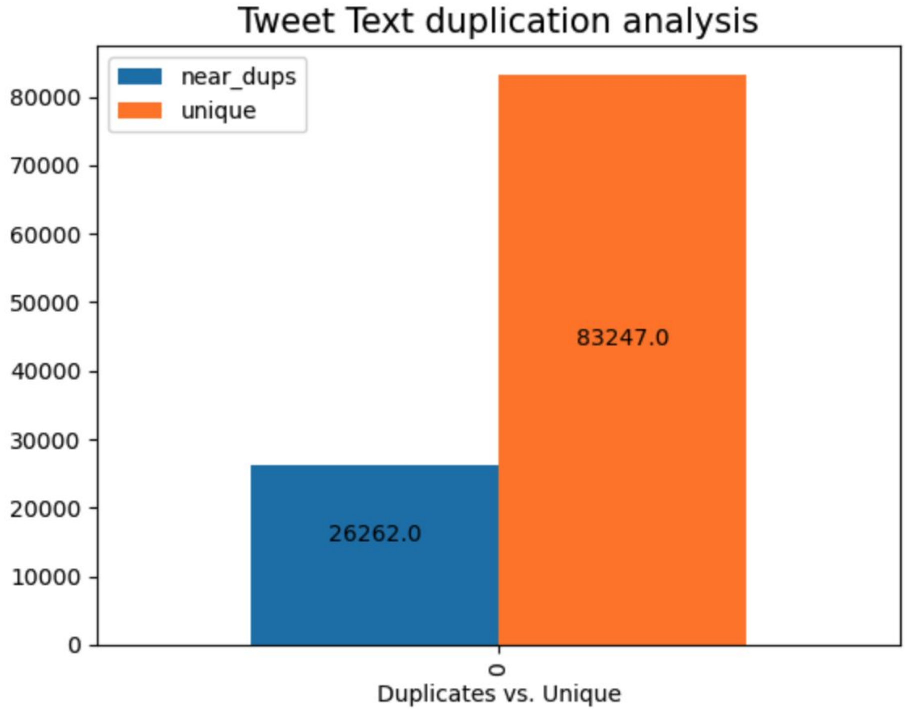


Message Uniqueness Analysis

Florida Math Book Ban

Using Jaccard similarity analysis, we can see that there are ⅕ of tweets that are almost the same with 0.2 Jaccard distance threshold. Therefore, we can tell that many people are copy-pasting the same text.

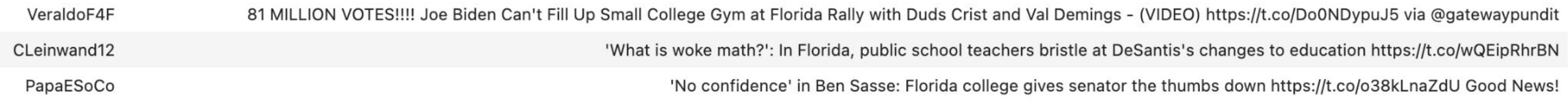
(I pick 0.2 as the threshold since I want to extract text that are highly similar, and after using 0.5 threshold, there are still diversity in the tweets although the method claims that they are the duplicate.)



Message Duplication Visualization

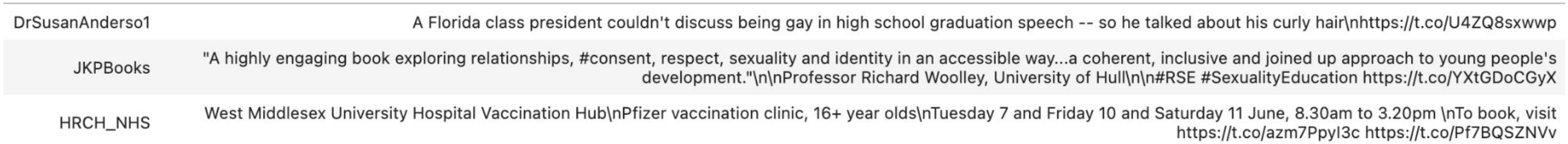
Topics in Florida

Political Entities



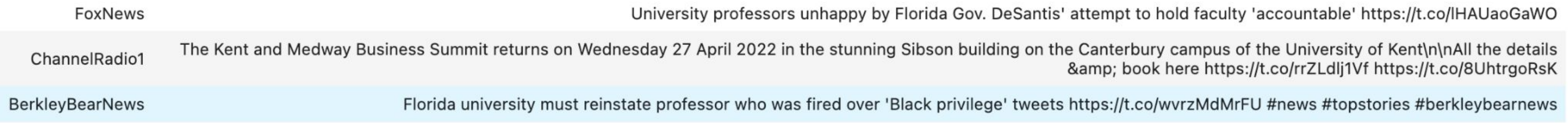
VeraldoF4F	81 MILLION VOTES!!!! Joe Biden Can't Fill Up Small College Gym at Florida Rally with Duds Crist and Val Demings - (VIDEO) https://t.co/Do0NDypuJ5 via @gatewaypundit
CLeinwand12	'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/wQEipRhrBN
PapaESoCo	'No confidence' in Ben Sasse: Florida college gives senator the thumbs down https://t.co/o38kLnaZdU Good News!

Health Related Users



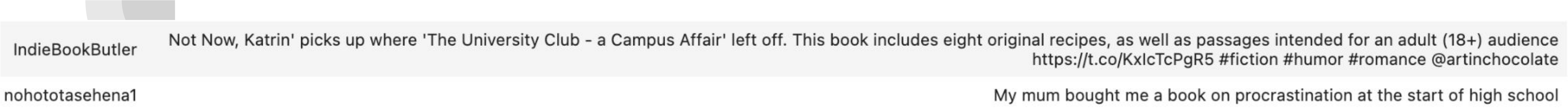
DrSusanAnderso1	A Florida class president couldn't discuss being gay in high school graduation speech -- so he talked about his curly hair\nhttps://t.co/U4ZQ8sxwwp
JKPBooks	"A highly engaging book exploring relationships, #consent, respect, sexuality and identity in an accessible way...a coherent, inclusive and joined up approach to young people's development."\n\nProfessor Richard Woolley, University of Hull\n\n#RSE #SexualityEducation https://t.co/YXtGDoCGyX
HRCH_NHS	West Middlesex University Hospital Vaccination Hub\nPfizer vaccination clinic, 16+ year olds\nTuesday 7 and Friday 10 and Saturday 11 June, 8.30am to 3.20pm \nTo book, visit https://t.co/azm7PpyI3c https://t.co/Pf7BQSZNVv

News Outlets



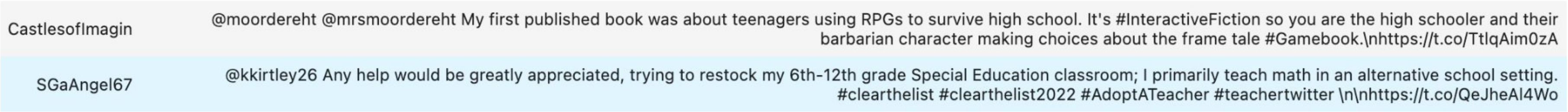
FoxNews	University professors unhappy by Florida Gov. DeSantis' attempt to hold faculty 'accountable' https://t.co/lHAUaoGaWO
ChannelRadio1	The Kent and Medway Business Summit returns on Wednesday 27 April 2022 in the stunning Sibson building on the Canterbury campus of the University of Kent\n\nAll the details &amp; book here https://t.co/rrZLdlj1Vf https://t.co/8UhtrgoRsK
BerkleyBearNews	Florida university must reinstate professor who was fired over 'Black privilege' tweets https://t.co/wvrzMdMrFU #news #topstories #berkleybearnews

Social Media Influencers



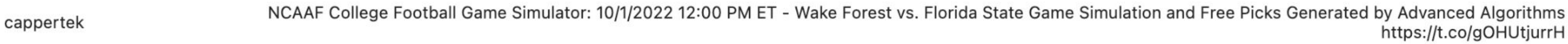
IndieBookButler	Not Now, Katrin' picks up where 'The University Club - a Campus Affair' left off. This book includes eight original recipes, as well as passages intended for an adult (18+) audience https://t.co/KxlcTcPgR5 #fiction #humor #romance @artinchocolate
nohototasehena1	My mum bought me a book on procrastination at the start of high school

Education Workers



CastlesofImagin	@moordereht @mrsmoordereht My first published book was about teenagers using RPGs to survive high school. It's #InteractiveFiction so you are the high schooler and their barbarian character making choices about the frame tale #Gamebook.\nhttps://t.co/TtlqAim0zA
SGaAngel67	@kkirtley26 Any help would be greatly appreciated, trying to restock my 6th-12th grade Special Education classroom; I primarily teach math in an alternative school setting. #clearthelist #clearthelist2022 #AdoptATeacher #teachertwitter \n\nhttps://t.co/QeJheAl4Wo

Sports Platforms



cappertek	NCAAF College Football Game Simulator: 10/1/2022 12:00 PM ET - Wake Forest vs. Florida State Game Simulation and Free Picks Generated by Advanced Algorithms https://t.co/gOHUtjurrH

We can see that only political entities have duplicate tweets in topic of Florida Math Book Ban. The other categories have their respectives duplicate in different topics in Florida.
This might be a good sign for using tweets as a credible information source since most of the important tweets in certain topics are not duplicate.

Conclusion

- Author Identification:
  - Most of the prolific twitterers are related to sports
  - Twitterers with most influential effect are political entities, then social influencer
  - Those that tweet about K-12 are news and political entities
- Location Analysis
  - Most of the tweets aggregate in main city like New York, Chicago, Los Angelas
  - Topic like Florida Math Book Ban spreads from East to the West
- Timeline Analysis
  - There is a gap from December to March which does not have any tweets
  - There are peaks in Spring and Autumn
  - Reason might be semester system and election
- Message Uniqueness Analysis
  - 26262 duplicate out of 109509 tweets in Florida Math Book Ban
  - Most of the duplicate tweets are not about the topic except political entities

Actionable Recommendations

1. To improve the analysis result, I think a better text mining method will work, such as not only considering one word token search, but using ngram words to filter topics.

2. Finding better topic could better analyze the geographical spread in tweets, which can make the analysis more clear and more logical.

3. Many tweets have the same user_id but different profile, which could be problem. This time I use max function to extract one user name for each user id. Next time, I should use create time to find themost recent identity.

4. User location variable I pick might not be a good indicator to find where exactly the post is made since the users might move; however, using place information of the tweet does not work either since there are lots of missing value. Finding a better way to profile their location could be a huge improvement in location analysis.