Education Tweets Analysis

Jasmeet Singh Sandhu

Agenda

1. Executive Summary with meaningful insights
2. Methodology and source data overview
3. Tweet clean-up and filtering
4. EDA
5. Author identification
6. Location analysis
7. Timeline analysis
8. Message uniqueness analysis
9. Conclusions and actionable recommendations
10. Appendix for more details

Executive Summary with meaningful insights

- Twitter is a social media website where people communicate in short text format based ‘tweets’ and is good source for projects involving text analysis, but data is expected be noisy since anyone and everyone can sign-up to the platform and start tweeting.

- The objective is to identify whether Twitter can be considered a credible source of information and profiling of twitterers to get an understanding of who and where these twitterers are located. Also, do the twitterers copy paste each other’s tweets or are they unique.

- Through the analysis I aim to seek answers to the following questions:
  - Who are the most prolific twitterers?
  - Where are these twitterers located?
  - What is the timeline of these tweets? Do we see significant peaks and valleys?
  - How unique are the messages?

Methodology and source data overview

Methodology

1. Pyspark for coding and analysis
2. Pandas for analysis on smaller datasets and for plotting
3. LSH for similarity analysis
4. seaborn, matplotlib, tqdm, geopy.geocoder, geopandas, plotly, wikipedia (pypi) to supplement analysis
5. Intermediate results stored in parquet format

Source Data Overview

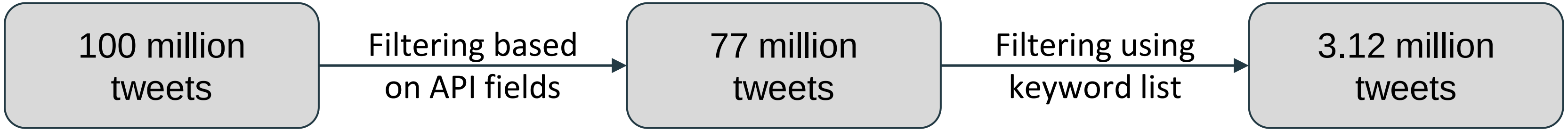
1. Stored on Google cloud
2. Total number of tweets in the source data = 100 million tweets
3. Total number of columns in the source data = 40

Tweet clean-up and filtering

1. I started by filtering the data based on twitter API doc – removing twitterers who don’t have any followers, tweets which are potentially sensitive, or which have been withheld in countries or which are truncated. Lastly, removing those tweets that are not in English.

2. Using automated script to scrape keywords and reading various articles, produced a list of words which are related to education and another list of words that should not be in tweets related to education.

3. Since, there are many tweets that contain “one” word related to education but are not related to education, I created a spark udf to only keep those tweets that have at least “two” words related to education.



100 million tweets

Filtering based on API fields

77 million tweets

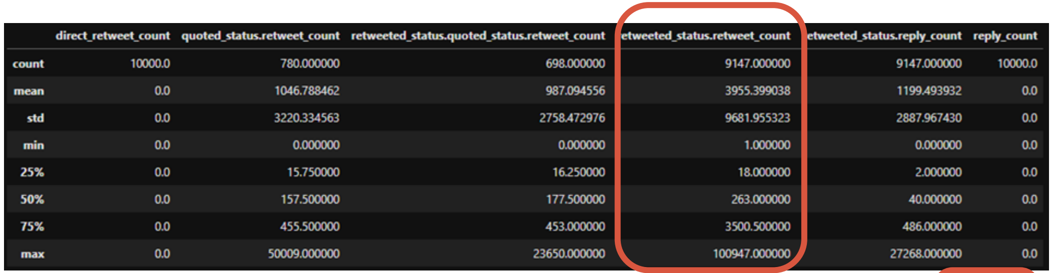
Filtering using keyword list

3.12 million tweets

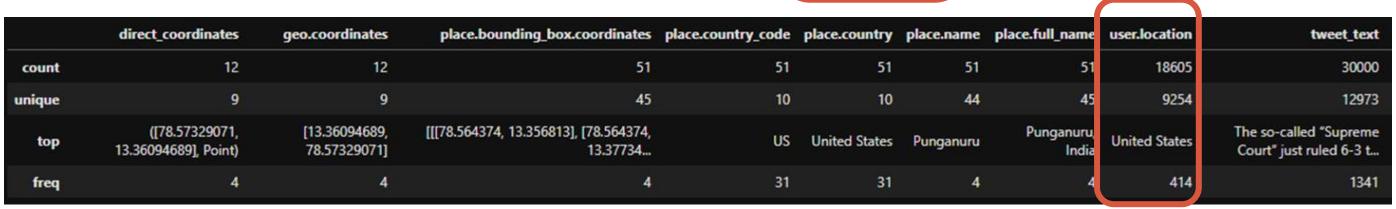
EDA – Field Analysis

To select columns to use for analysis, I selected columns which were relevant to the goal of the analysis and had fewer missing values

- For retweets and geographic data, I selected retweeted_status.retweet_count and user.location since it had least percentage of nulls and based on API docs, serves our purpose for analysis of retweets and geographical data.



	direct_retweet_count	quoted_status.retweet_count	retweeted_status.quoted_status.retweet_count	etweeted_status.retweet_count	etweeted_status.reply_count	reply_count
count	10000.0	780.000000	698.000000	9147.000000	9147.000000	10000.0
mean	0.0	1046.788462	987.094556	3955.399038	1199.493932	0.0
std	0.0	3220.334563	2758.472976	9681.955323	2887.967430	0.0
min	0.0	0.000000	0.000000	1.000000	0.000000	0.0
25%	0.0	15.750000	16.250000	18.000000	2.000000	0.0
50%	0.0	157.500000	177.500000	263.000000	40.000000	0.0
75%	0.0	455.500000	453.000000	3500.500000	486.000000	0.0
max	0.0	50009.000000	23650.000000	100947.000000	27268.000000	0.0



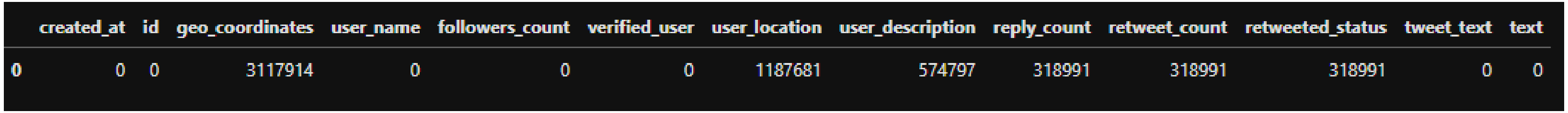
	direct_coordinates	geo.coordinates	place.bounding_box.coordinates	place.country_code	place.country	place.name	place.full_name	user.location	tweet_text
count	12	12	51	51	51	51	51	18605	30000
unique	9	9	45	10	10	44	45	9254	12973
top	([78.57329071, 13.36094689], Point)	[13.36094689, 78.57329071]	[[[78.564374, 13.356813], [78.564374, 13.37734...	US	United States	Punganuru	Punganuru, India	United States	The so-called "Supreme Court" just ruled 6-3 t...
freq	4	4	4	31	31	4	4	414	1341

- For text of the tweets, I used tweet_text because text contains retweet information but the tweet_text contains only the text of the tweet.

EDA – Checking for nulls on selected columns

After selecting columns based on the goals for analysis, I selected a list of 13 columns that are needed for the analysis, ex: tweet_text for similarity and user_description for.

Post that, I checked for the count of nulls in those selected columns to check if some more processing or imputation of missing values will be needed

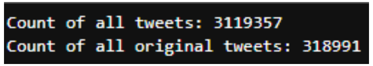


	created_at	id	geo_coordinates	user_name	followers_count	verified_user	user_location	user_description	reply_count	retweet_count	retweeted_status	tweet_text	text
0	0	0	3117914	0	0	0	1187681	574797	318991	318991	318991	0	0

1. Geographical coordinates do contain null values since many twitterers may prefer to hide their location but can still have information to share
   a. Can’t discard null geographical values
2. Same reasoning as user_description, many people can have no descriptions but can be credible source of information
   a. Can’t discard tweets with null user_descriptions
3. Missing retweet_counts, reply_counts and retweet_statuses refer to the original tweets since retweet_status is null for original tweets

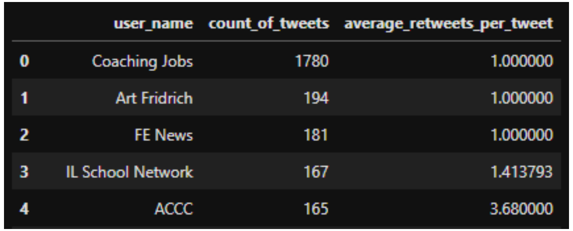
Author Identification - Prolific and Average Retweets

- Out of 3.12 million tweets, 318k were original, which shows majority of twitterers just retweet/quote each other.

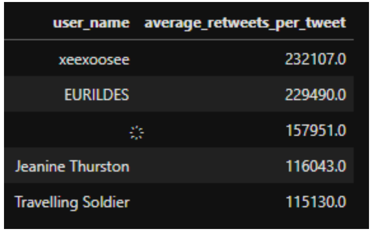


```
Count of all tweets: 3119357
Count of all original tweets: 318991
```

- To identify the authors of tweets vs people who just retweet others, I analysed only those tweets who had the ‘retweeted_status’ as null, since they’re the original tweets.
    - After filtering, using group by user_name, I identified the twitterers who had highest number of original tweets (only showing top 5) and on average how many retweets they get (which is surprisingly low)

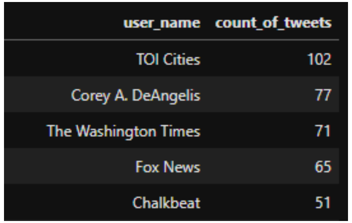


	user_name	count_of_tweets	average_retweets_per_tweet
0	Coaching Jobs	1780	1.000000
1	Art Fridrich	194	1.000000
2	FE News	181	1.000000
3	IL School Network	167	1.413793
4	ACCC	165	3.680000



user_name	average_retweets_per_tweet
xeexoosee	232107.0
EURILDES	229490.0
҉	157951.0
Jeanine Thurston	116043.0
Travelling Soldier	115130.0

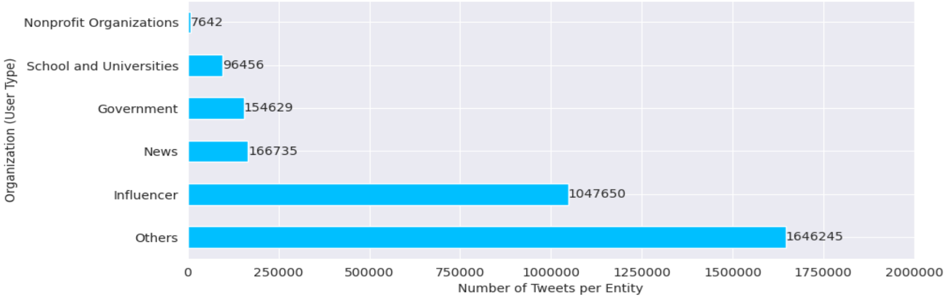
- However, verified twitterers, tweet less as compared to non-verified twitterers about education.



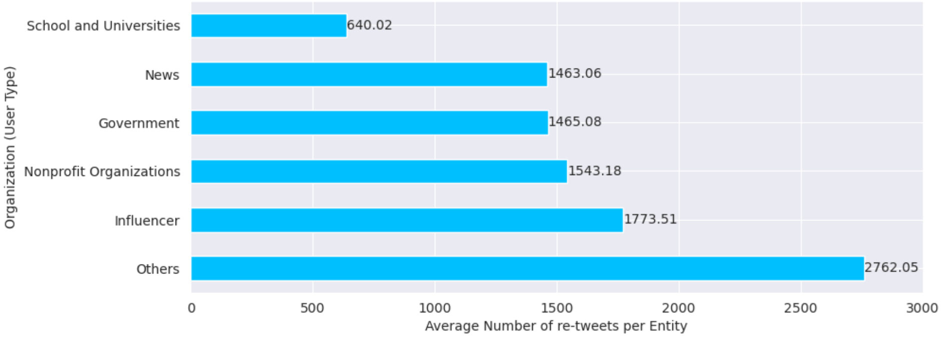
user_name	count_of_tweets
TOI Cities	102
Corey A. DeAngelis	77
The Washington Times	71
Fox News	65
Chalkbeat	51

Author Identification - Entity Analysis

- To assign each user to either government, or influencers, I used a series of words and based on if these 'filter' words exist in user_description or user_name, assigned them to entities



From this graph we can see that the majority of the tweets are random twitterers tweeting about education, with non-profit organization tweeting the least about education

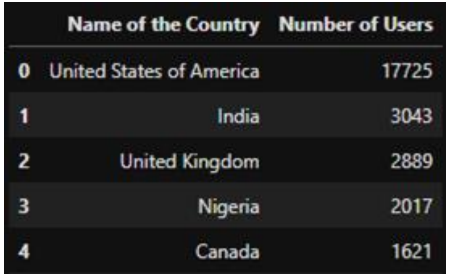


As "Others" can include people who are not in other categories, it's possible to have some viral tweets that skewed the average number of tweets to be maximum for the "Others" group.

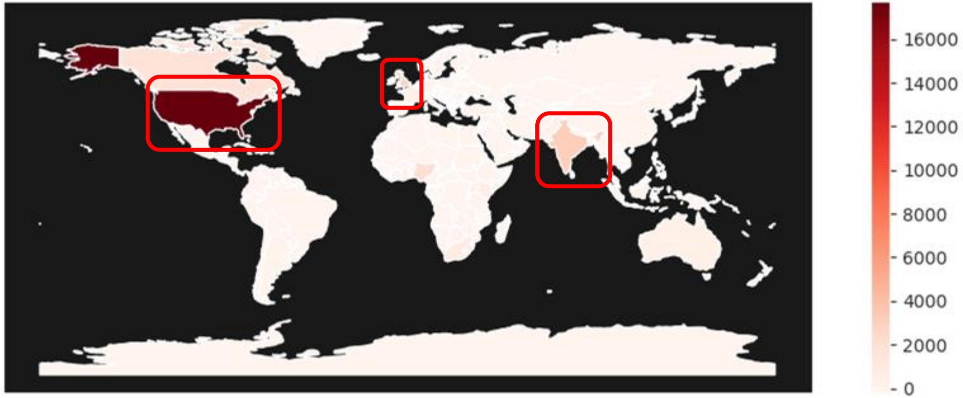
Apart from "Others", the "influencers" get the maximum amount of retweets on average

Location analysis - User location

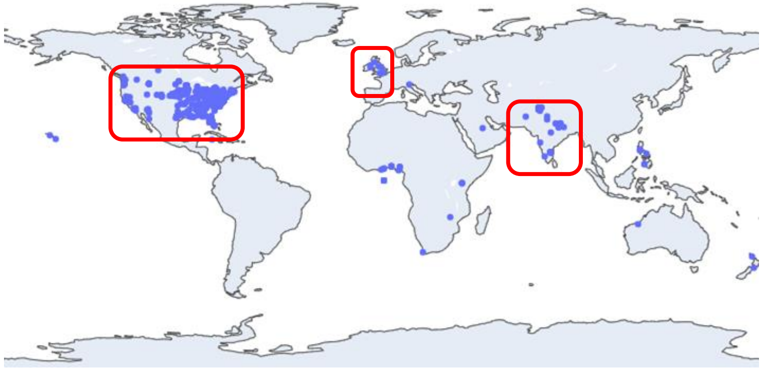
- To identify where the twitterers are located, I did analysis on
  - Coordinates
  - Geo-decoding the user_location using ‘Nominatim’ from the geopy module
    - I used a geo-decoder since we have more data in user_location and thus can get more accurate distribution
    - The place.country can also give us country but it has only 6557 non-null values in entire dataset (vs 43k transformed from user_location sample)



	Name of the Country	Number of Users
0	United States of America	17725
1	India	3043
2	United Kingdom	2889
3	Nigeria	2017
4	Canada	1621



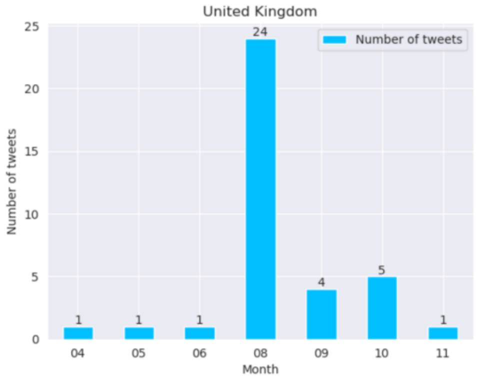
Number of twitter twitterers per country

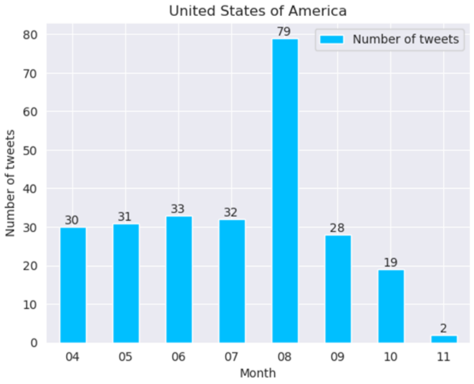


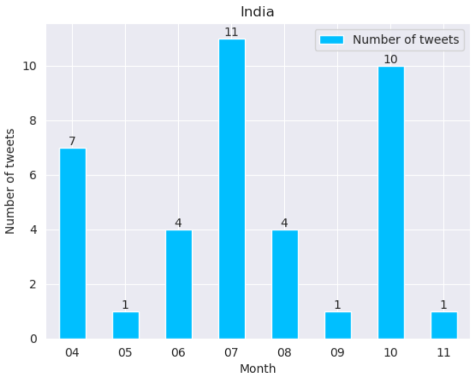
Twitter user’s location using coordinates data

Location analysis - Topic Progression

- Since the location data had few non null values and due to API performance issues, I was able to do the progression analysis with country on 43000 tweets
- I looked at progression on tweets containing ‘policy’.
  - Since, the tweets are already filtered - We can see progression of number of tweets related to education policy for the top three countries - USA, India, UK



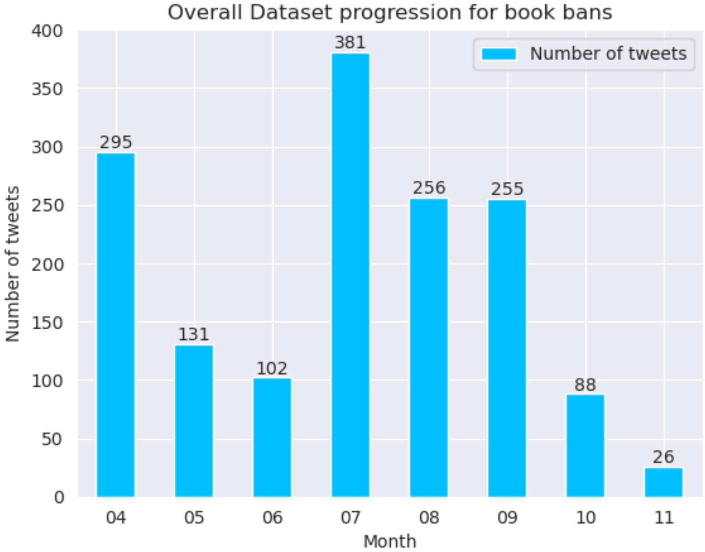




- Since, the number of tweets is not high, I’m not confident in calling it a trend.

Location analysis - Topic Progression

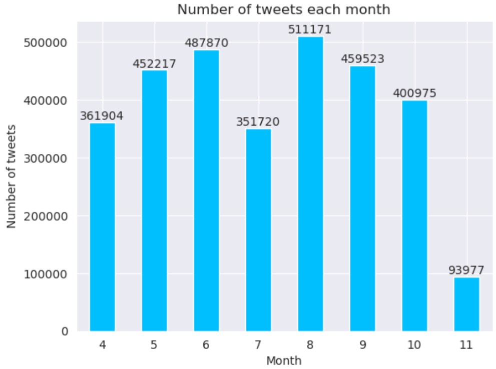
- However, to complete the analysis, I ran the analysis for ‘book bans’ on the entire dataset irrespective of country.



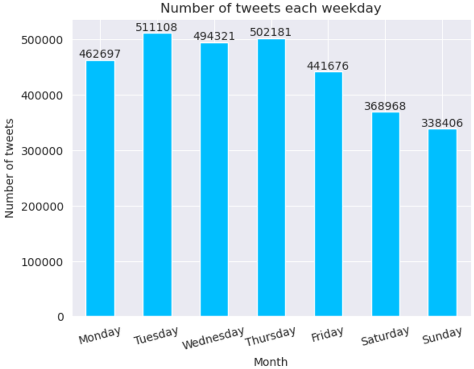
- To summarize, we see peaks during the start of school year (July - September) and before the summer break (April) and a drop in summer months and after start of school year.

Timeline analysis

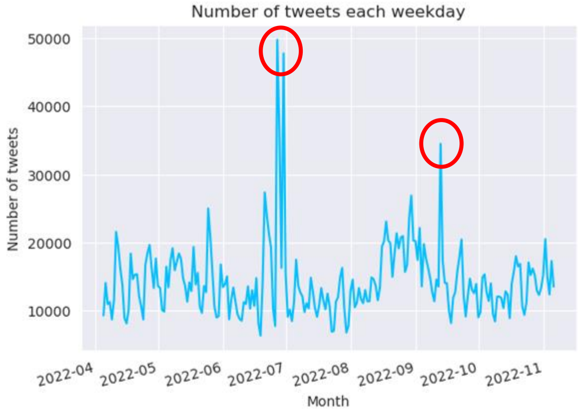
- Since, we only have data starting April 2022, there's a data gap for first three months of 2022 and the month of December 2022.



Graph1



Graph2



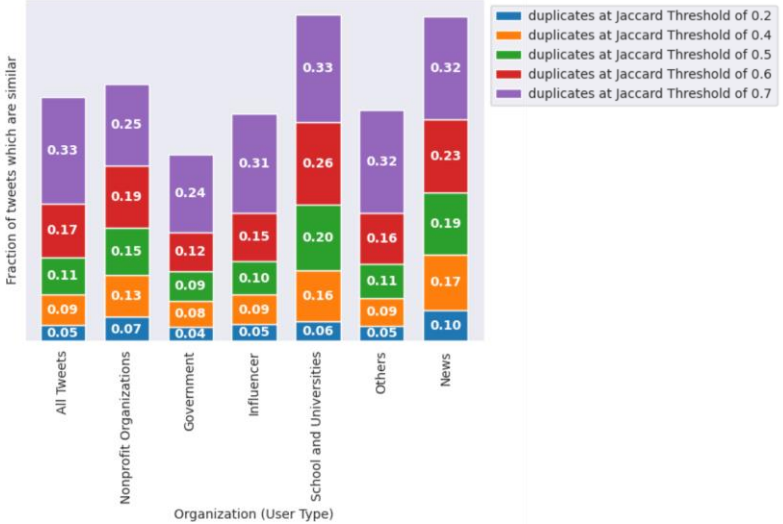
Graph3

- From the Graph3, we can see the maximum number of tweets during last few days of June.
- From the Graph1, we can see the number of tweets per month peak in August and June, i.e., start of academic year and during summer.
- From the Graph2, we can see that maximum number of tweets are on Tuesdays and Thursdays.
- To summarize, we see peaks during the start of school year (July - September) and during the summer break (June)

Similarity analysis

- I checked the similarity for the following
  - On a sample of original texts (Total 54890 out of 318k tweets) across all twitterres
  - For each entity
    - All original tweets from Non-profit organizations, Government, Schools and Universities
    - Influencers - 55% sample (41k tweets)
    - Other twitterres - 25% sample (45k tweets)
    - News - 80% sample (25k tweets)

- From the graph we can see that most of tweets are unique even at the Jaccard threshold of 0.7
  - Most of the duplicates can be seen in the tweets from the ‘News’ and ‘Schools and Universities’ at 23% and 26% duplicates at a threshold of 0.6.
  - The least number of duplicates are for Government, 12% at a threshold of 0.6.
  - ‘Others’ and ‘Influencers’ have similar percentage of duplicates across all thresholds.



Conclusions

- After starting with a dataset of 100 million I was able to bring it down to 3.12 million which I used for my analysis.
- Most prolific twitterers
  - By message volume, the most prolific twitterers are ‘Others’ followed by ‘Influencers’, ‘News’, ‘Government’, ‘Schools and Universities’ and ‘Non-profit organizations’.
  - Ironically, most prolific twitterers are not the ones who get retweeted the most on average!
- Location of these twitterers
  - The top 5 countries where these twitterers are located: United States, United Kingdom, India, Nigeria and Canada
  - Since, the location data was limited in the dataset after filtering, I was not able to find any significant trends in the data for policies in top three countries.
  - However, when I saw the trend of ‘book ban’ in the dataset for all locations, I saw a peak in the months of April and July with a decline from July to November.
- Timeline of tweets
  - After plotting tweet volumes per month, weekday and daily for the data set, I saw the following patterns:
    - The most number of tweets daily happened in the last few months of June and mid-September
    - The monthly volumes were highest for the month of August, June, and September respectively
    - Tuesdays and Thursdays have the highest number of tweet volumes
    - I see peaks during the start of school year (August - September) and during the summer break (June)
  - Since the data starts from April and is only till November, we can see data collection gaps for January, February, March and December of 2022
- 90% of twitter volume after filtering consisted of retweeted tweets, but the other 10% ‘original’ tweets had very few duplicates (10% duplicates at a Jaccard threshold of 0.5)

Actionable Recommendations

- Since there are only few verified twitterers in the dataset after filtering, and majority of tweets being from non-verified twitterers, I would not consider twitter as a good source for topics related to education
    - In order to make twitter as a good source, we should only use tweets from verified sources, or tweets which have links to other news sources or research articles.
    - We should be mindful of who’s tweeting, and if are they knowledgeable on the topic they’re tweeting about.

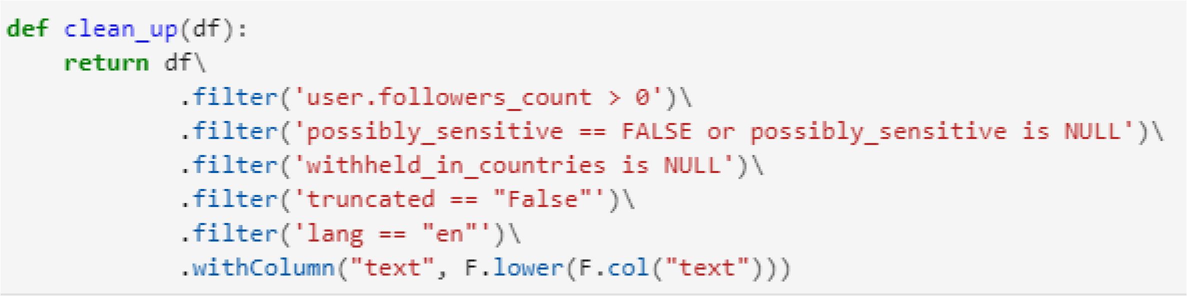
- Though not covered during this analysis, we should also be on the lookout for very sentimentally charged tweets, i.e., tweets which are either very positive or very negative and how these tweets can skew public opinion.

- Since most of the twitterers originate from the United States, to get a more global outlook regarding education, we’ll need to supplement our analysis with data from other sources.

- In my opinion, for topics such as education, it would be more beneficial to analyse news websites to conduct the analysis since twitter data is very noisy.

Appendix

API based filtering



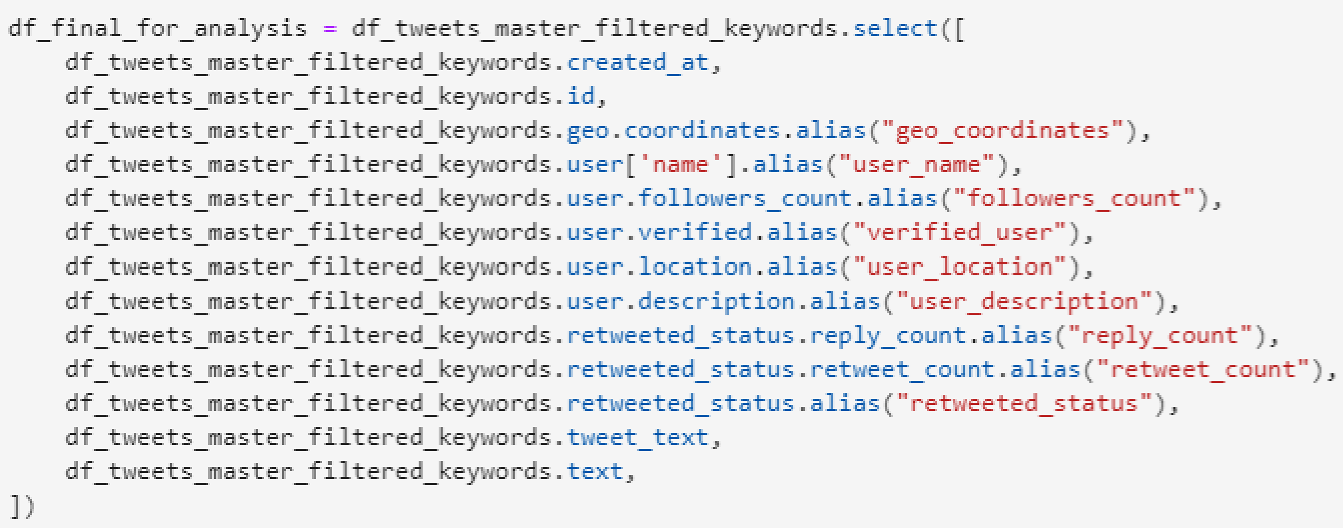
```
def clean_up(df):
    return df\
            .filter('user.followers_count > 0')\
            .filter('possibly_sensitive == FALSE or possibly_sensitive is NULL')\
            .filter('withheld_in_countries is NULL')\
            .filter('truncated == "False"')\
            .filter('lang == "en"')\
            .withColumn("text", F.lower(F.col("text")))
```

Before going to the text analysis, I started by going over the API docs to eliminate tweets based on what the columns represent.

1. I removed twitterers who had followers count zero, since they may be bot accounts.
2. I removed tweets that had been marked as sensitive content, since tweets related to education are not sensitive.
3. Similarly, I removed tweets that are withheld in some countries.
4. I also removed tweets which didn’t have the full text
5. Next, I removed the tweets that were not in english.

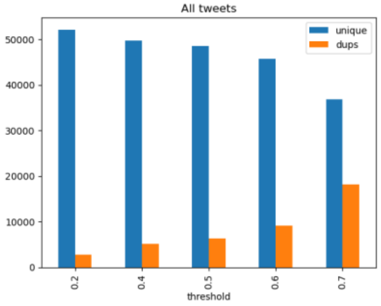
As a result of this filtering, I was left with ~77 million tweets out of 100 million tweets

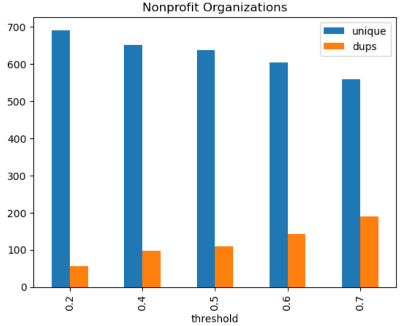
List of variables used for analysis

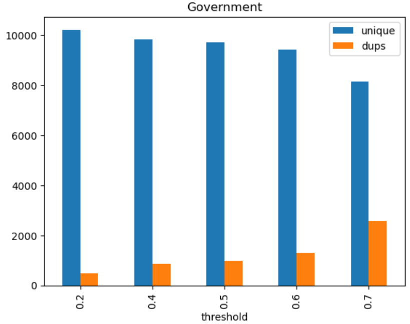


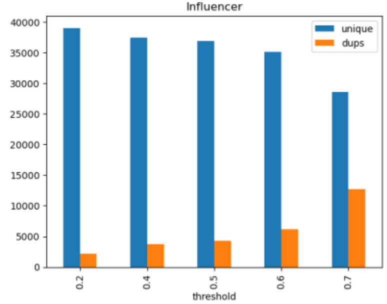
```
df_final_for_analysis = df_tweets_master_filtered_keywords.select([
    df_tweets_master_filtered_keywords.created_at,
    df_tweets_master_filtered_keywords.id,
    df_tweets_master_filtered_keywords.geo.coordinates.alias("geo_coordinates"),
    df_tweets_master_filtered_keywords.user['name'].alias("user_name"),
    df_tweets_master_filtered_keywords.user.followers_count.alias("followers_count"),
    df_tweets_master_filtered_keywords.user.verified.alias("verified_user"),
    df_tweets_master_filtered_keywords.user.location.alias("user_location"),
    df_tweets_master_filtered_keywords.user.description.alias("user_description"),
    df_tweets_master_filtered_keywords.retweeted_status.reply_count.alias("reply_count"),
    df_tweets_master_filtered_keywords.retweeted_status.retweet_count.alias("retweet_count"),
    df_tweets_master_filtered_keywords.retweeted_status.alias("retweeted_status"),
    df_tweets_master_filtered_keywords.tweet_text,
    df_tweets_master_filtered_keywords.text,
])
```

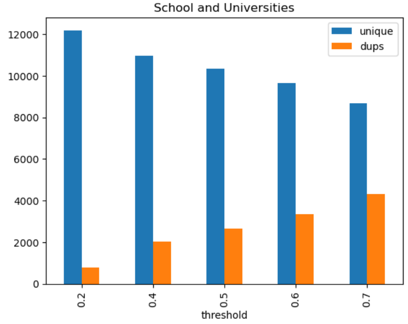
Similarity analysis for all entities for various Jaccard thresholds

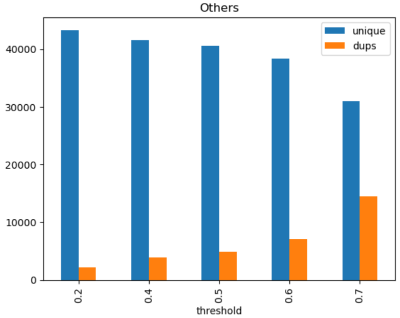

All tweets
unique
dups
threshold

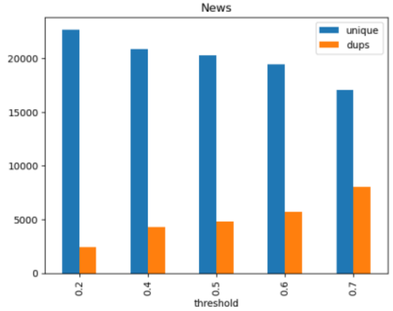

Nonprofit Organizations
unique
dups
threshold


Government
unique
dups
threshold


Influencer
unique
dups
threshold


School and Universities
unique
dups
threshold


Others
unique
dups
threshold


News
unique
dups
threshold