TWITTER ANALYSIS

- BDP FINAL PROJECT

Yue Zhang

Summary

Problem

*Whether **Twitter** can be considered a credible source of information, reflects the emergence of important trends or topics in **education**, specifically: "**Online Learning**".*

Contents

01 Overview and Filtering

- *Exploratory Data Analysis*
- *Discard irrelevant data and select the useful features*

02 Feature Engineering

- *Add new features for future analysis*

Select Topic: Online Learning

03 User Profile Analysis

- *Prolific Twitterers*
- *Location*
- *Time*
- *Message Uniqueness*

04 Conclusions

- *Insights and next steps*

Data Overview and Methodology

Data Overview and EDA

- **Twitter data** collected on the topics of education, schools, universities, learning, knowledge sharing, etc.
- Almost 100 million Tweets (~500GB) in **JSON** File
- **99994342 rows** and **39 features** in total
- **72370910** tweets from **year 2022** and **9399066** tweets from **year 2023**
- All the tweets texts are **in English only**
- **Missing value:**
  Coordinates: 97554/99992797 = **1% "null" value**
  Retweet_count: all "null" values

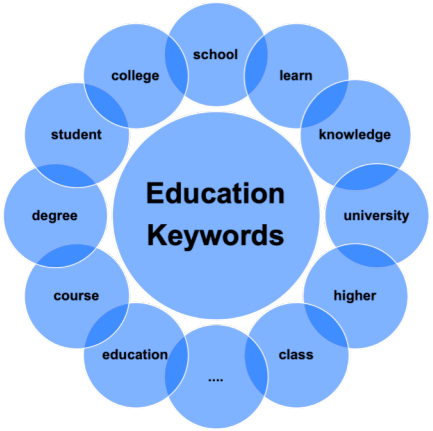
Methodology

- Spark PDD
- Spark DF: filter/withcolumn/LSH
- Pandas
- Matplotlib
- …

Data Cleaning

Filtering:

- **Discard irrelevant tweets:** Filtering for text that is related to education topic based on **education-related keywords** including “school”, “college”, “university”, ...etc.
- **Discard variables** that are **poorly populated** such as “retweet_count”.



Selecting Features:

- **Removing unhelpful columns** that contains multiple unused json files based on the variable name and its schema. (Including "tweet_text", "text"...etc)
- **Refining useful data** from the dropped data frame to get only useful information.

Feature Engineering

The following features are generated from data source and used for further analysis:

“user_name”
From: user['name']

“user_description”
From: user['description']

“user_screen_name”
From: user['screen_name']

“user_location”
From: user['location']

“place_country”
From: place['country']

“Longitude” & “Latitude”
From: coordinates.
coordinates [1],
, coordinates. coordinates [0]
(In the location analysis part)

“retweet_count”
From:
retweeted_status['retweet_
count']

“retweeted”
From:
retweeted_status['retweeted']

Identify the Most Prolific /Influential Twitterers

By message volume (original content)

user_screen_name	count
education_24x7	8040
jwindon35	4848
educationbnb	4493
techysaavy	4234
jc_james_clark	4210

User **‘education_24x7’** posted **the most** original contents.

By massage retweet (how often their massages are being retweeted)

user_screen_name	max(retweet_count)
FashionBlock08	516855
AlexaLockwood2	516850
savvh12	516795
TanjeenThat	516791
CrimsonOmbre115	516779

User **‘FashionBlock08’** has **the most** massage retweet count.

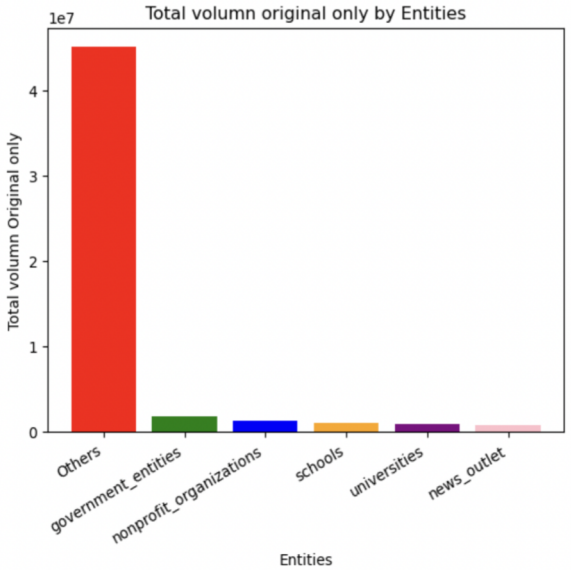
Distribution of tweet / retweet volume by Organizations

Who are these Twitters:

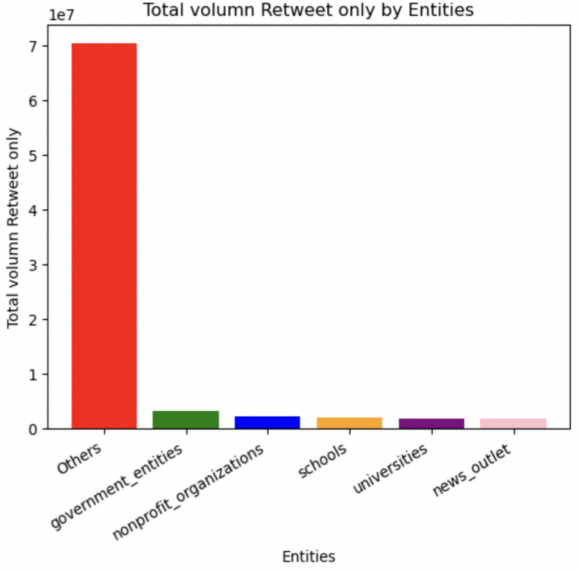
	entities	count
0	Others	70376518
1	government_entities	3216683
2	nonprofit_organizations	2200748
3	schools	2146956
4	universities	1970089
5	news_outlet	1858982

Except the general user, government entities contribute **most** to the twitters about Education.

Distribution of tweet volume by organizations

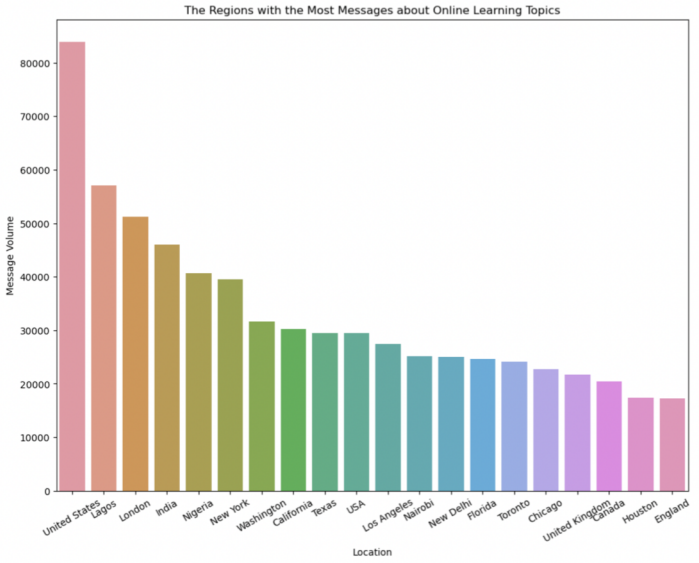


Distribution of retweet volume by organizations



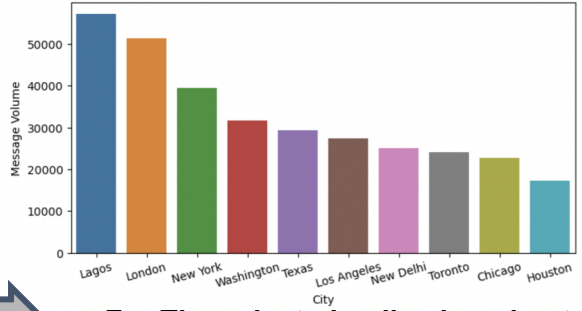
Location Analysis

EDA: Regions with the largest message publication volume on the selected topic

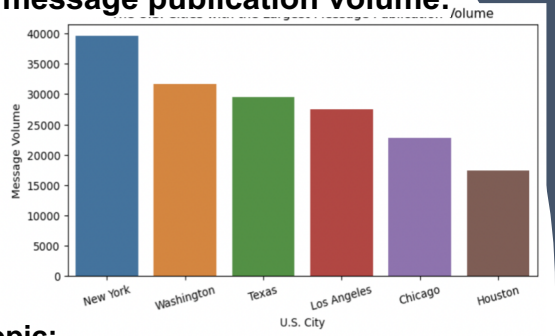


The name of the location is **messy**. Therefore, some **typical cities** are selected for further analysis.

Cities with the largest message publication volume:



U.S. cities with the largest message publication volume:

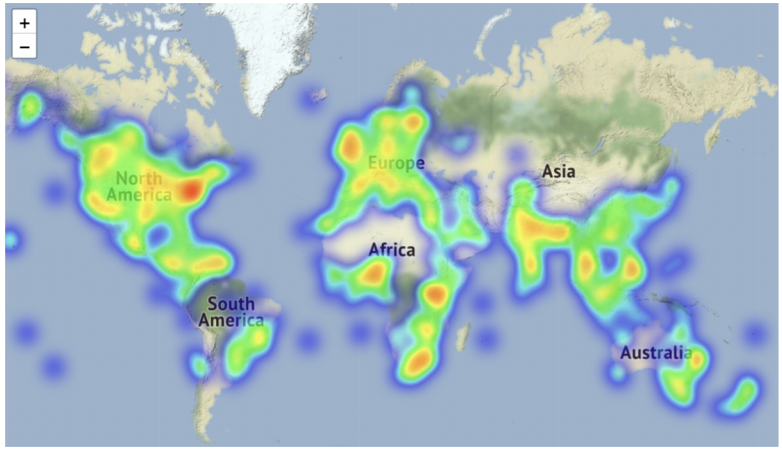


For The selected online learning topic:

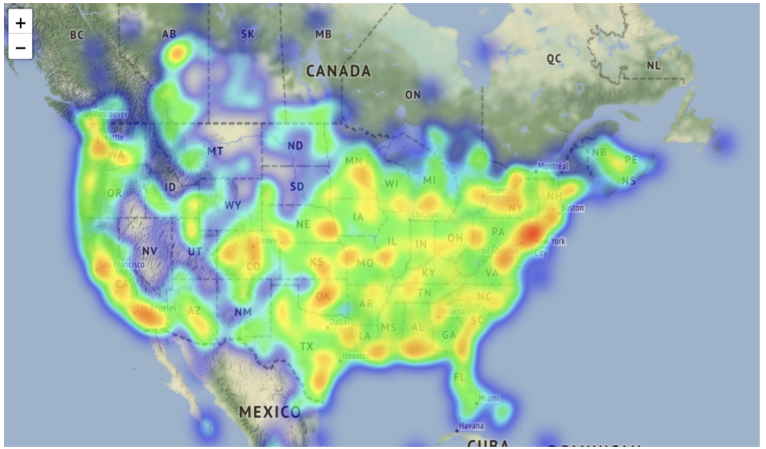
- From these three plots, **United States** is the country having **the most** tweets about Education and the online learning topic, which may result from the advanced technology level and large population.
- **Lagos** is the city having **the most** tweets related to the online learning topic. The reason for this may be that: 1) Lagos may have a high concentration of educational institutions that offer online courses; 2) Lagos may have been particularly affected by the pandemic, leading to a higher number of tweets about online learning.
- In United States, **New York** is the city having **the most** tweets related to the online learning topic. Perhaps this is because that it has both high concentration of educational institutions and large population, and is also known for its thriving tech industry.

Location Analysis
- Geographical Distribution

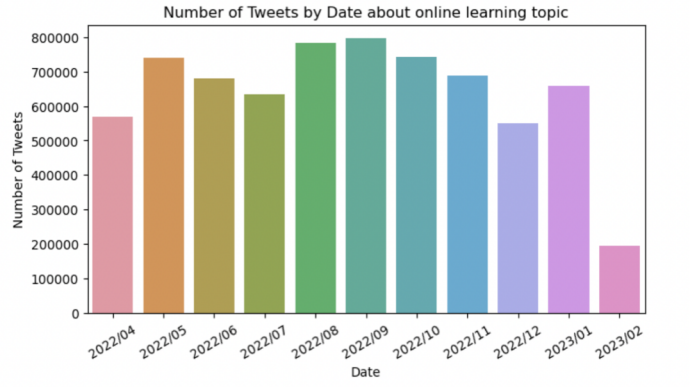
Global

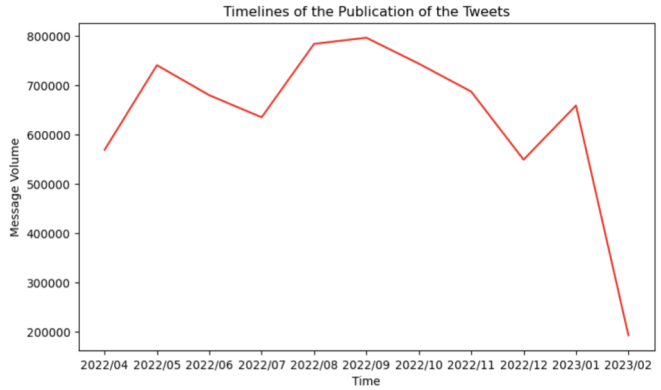


United States



Timeline Analysis

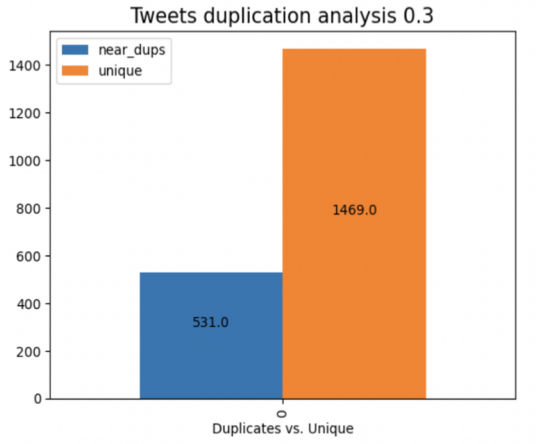


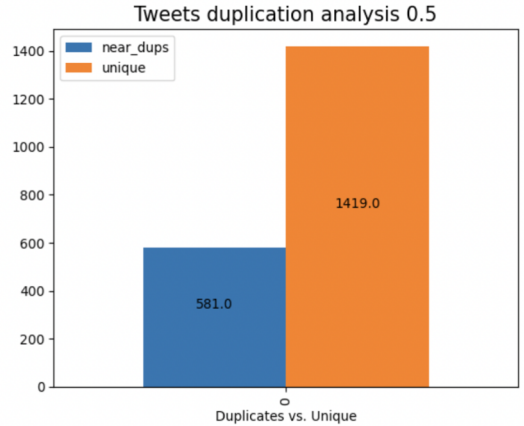


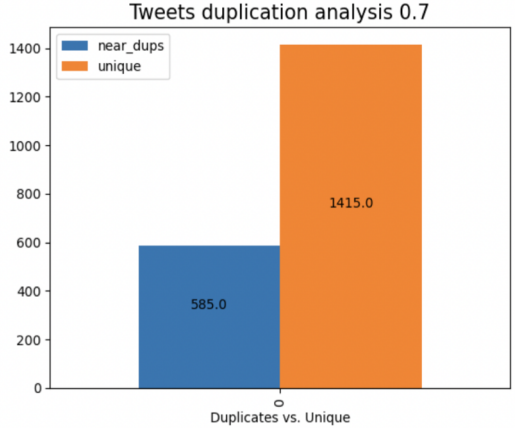
For The selected online learning topic:

- There are **obvious gaps** between different dates on the amount of tweets
- **A significant peaks** appear on **August and September, 2022**. The reason for this may be that schools or universities usually start at this time.
- **From Oct to Dec in 2022**, the amount of tweets **keeps decreasing**, which may reflect people's **decreased attention on online learning** as the COVID situation improves and people gradually get used to online learning.
- And there is **a significant decrease in Feb**, 2023 and **this may due to the lack of data** (not yet get updated since this is the most recent month)
- In general, the amount of tweets about online learning **fluctuated along with the time**.

Massage Uniqueness Analysis

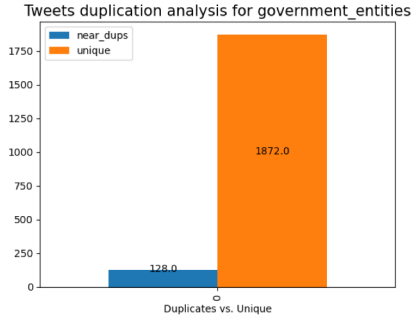


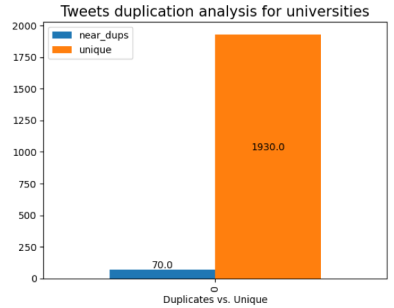


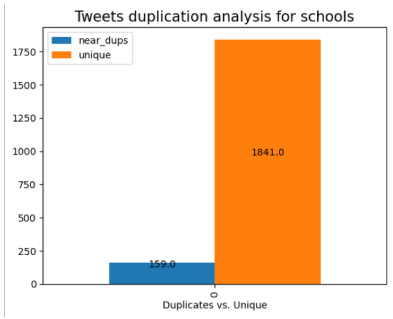


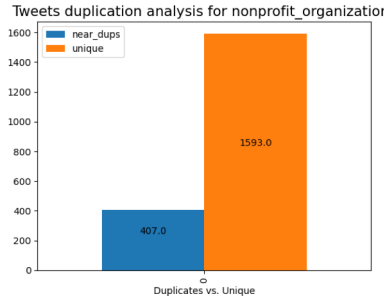
- Use **Jaccard Similarity** to find out whether the messages are mostly unique
- Choose **2000 random samples**
- **The threshold is set as 50 for maximum accuracy (Jaccard distance = 0.5)** based on the duplication text analysis
- From the result, **most Tweets are unique**

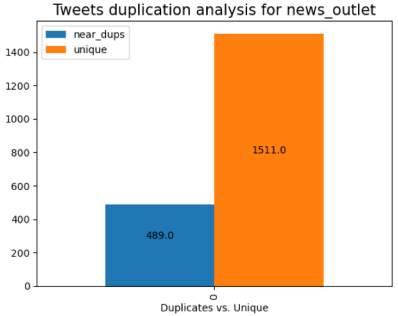
Massage Uniqueness Analysis - Visualization for Each Group

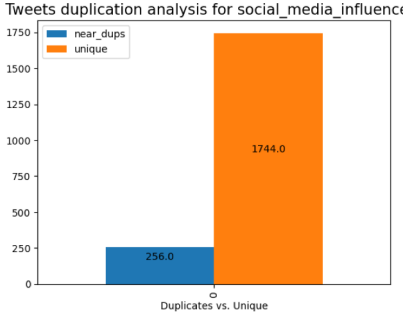


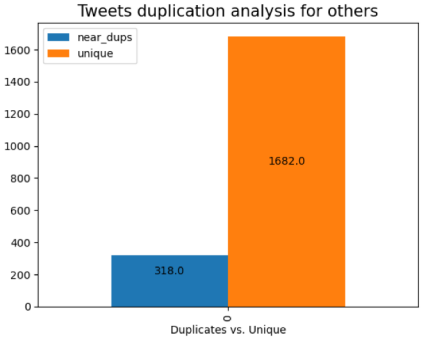












	group	duplicate/unique
1	universities	0.036269
0	government_entities	0.068376
2	schools	0.086366
6	Others	0.140482
5	social_media_influencers	0.189061
3	nonprofit_organizations	0.255493
4	news_outlet	0.323627

Almost all groups of Twitterers (2000 samples each group) **tend to post more unique tweets** than the ordinary users. **News outlet** has **the highest rate of retweet count** (duplicate/unique rate = 0.324) compared to the rest of the entities while **universities** has **the least** (duplicate/unique rate = 0.036) .

Conclusions and Next Steps

Conclusions

- In this project, **online learning** is chosen as the “important topic in education”. And analysis on **author identification, location, timeline, message uniqueness** is conducted to see whether Twitter can be considered a credible source of information.
- **From the result,** Twitter **could be considered an important source of information** that can reflect the emergence of important trends or topics in education. And it also could be useful for understanding the public’s opinion on a certain topic or trend in education.
- **However,** based on the current analysis, it should **not be considered a creditable enough source** to obtain knowledge for the topic in general **until further tweet analysis.** Since many tweets are original content created by social media influencers other than authority agencies such as governments, schools, news, and non-profit organizations.

Next Steps

- In addition to the current analysis, the analysis of the credibility of some **original tweets** created by **social media influencers** could be the aim for the next step: *“How credible are the original tweets could be taken for topic knowledge gaining from non-authority entities?”*
- We could also divide the Twitterers into **more groups** and conduct some further analysis.
- In this project, the key objective is to analyze the profiles of Twitters, rather than their actual text messages. In the future, we could **also analyze the contents of all these tweeters** to see if there exists any interesting finding, using texting mining and NLP methods.