# Data Science Mini Project 3

Aarohi Garg

**Link to Colab Notebook:**
https://colab.research.google.com/drive/1eRBw_yQ900dwdyTw6akzhvq9_5xFzg4J?usp=sharing

**Problem Statement**

Explore the tweets posted by any two international universities. Analyze the data in tweets and answer some common questions that international students applying to the university might have, like:

a) Which words are mostly used in the tweets?
b) Which hashtags does the university use mostly?
c) Which university is more active during summer?

Perform some analyzes that helps the university, like:

a) What day of the week do the shared tweets attract more attention?
b) What time of day do the shared tweets attract maximum likes?

The scope is not limited to answering these questions. The range of analysis can widen depending upon time constraints.

**Problem Importance**

This project helps the international students applying to universities gain more insight into what the university promotes, student life etc. It can further help them compare two universities based in same city.
This project can also be used by universities to improve and regulate their twitter posts depending upon the response statistics.

**Selecting Universities to Analyze**

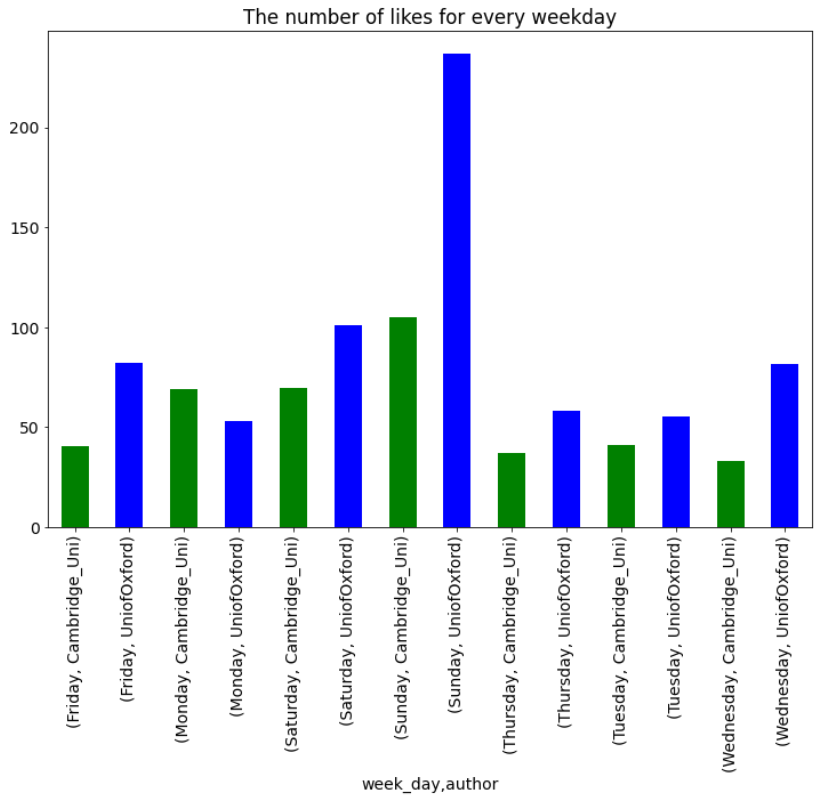
Being a student in Turku (Finland), the initial goal was to analyze two important universities based in Turku which are Åbo Akademi University and University of Turku. But since these universities are based in Finland, some of the posts were in Finnish or Swedish and analyzing hashtags and words was difficult due to my lack of knowledge in these languages.

Therefore, the universities chosen for analyses are two renowned universities based in United Kingdom namely, University of Oxford (@UniofOxford) and Cambridge University (@Cambridge_Uni). The following are primary reasons for choosing these universities:

1. They belong to English speaking regions; hence the tweets are guaranteed to be in English language.
2. They are highly popular among the youth, with thousands of followers, so analyzing statistics would be interesting.

## Data Analysis – Beneficial to University

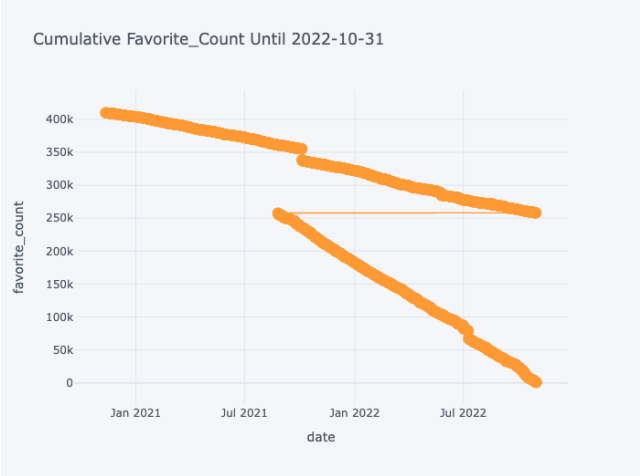
1. ***What day of the week do shared tweets attract more attention for each university?***




***Observation 1:*** *Both universities attract most attention on their shared tweets on* ***Sunday****. This is expected since people are usually free on Sundays.*
*However,* ***University of Oxford*** *receives more than double the number of likes on Sunday as compared to* ***Cambridge University****.*

```
********************************************************************************
Please, choose date interval to see some statistical information about tweets.
    start_date  01/10/2022
    end_date    31/10/2022
********************************************************************************
UniofOxford published 269 tweets between 2022-10-01 and 2022-10-31.
These tweets have got 23664 likes and retweeted by 4635 totally.
--------------------------------------------------------------------------------
On the other hand, Cambridge_Uni published 116 tweets during this time.
These tweets have got 4710 likes and retweeted by 1130.
********************************************************************************
```

***Observation 2:*** *Above are the statistics for the month of* ***October****, which clearly shows that University of Oxford has substantially a greater number of likes, and retweets than Cambridge University.*




***Observation 3:*** *Cumulative of Favorites of both universities date wise till 31st October.*
*Older tweets have more likes than the recent ones. One possible explanation for this trend is that more people read and like the tweets with time.*

2. ***What is the relationship between the number of likes of tweets and other variables?***

The following fields are used to define part of the day according to time:

- Early morning 5 ~ 8
- Morning 8 ~ 11
- Late morning 11 ~ 12
- Early afternoon 13 ~ 15
- Late afternoon 15 ~ 17
- Early evening 17 ~ 19
- Late evening 19 ~ 21
- Night 21 ~ 4

University Name: Cambridge_Uni

| | day_part | favorite_count |
|---|---|---|
| 1 | Early Evening | 101.555172 |
| 3 | Late Afternoon | 47.260647 |
| 6 | Morning | 44.554527 |
| 0 | Early Afternoon | 40.288158 |
| 7 | Night | 38.810811 |
| 5 | Late Morning | 36.493075 |
| 4 | Late Evening | 34.209677 |
| 2 | Early Morning | 27.683333 |

University Name: UniofOxford

| | day_part | favorite_count |
|---|---|---|
| 4 | Late Evening | 310.544444 |
| 1 | Early Evening | 166.153584 |
| 2 | Early Morning | 115.869565 |
| 0 | Early Afternoon | 66.618406 |
| 5 | Late Morning | 62.836538 |
| 6 | Morning | 61.788587 |
| 3 | Late Afternoon | 61.739048 |
| 7 | Night | 47.250000 |

***Observation 1:*** *More likes are received in the **evening** for both the universities. This is again an expected behavior since people usually have spare time to check their social sites after work in the evening.*

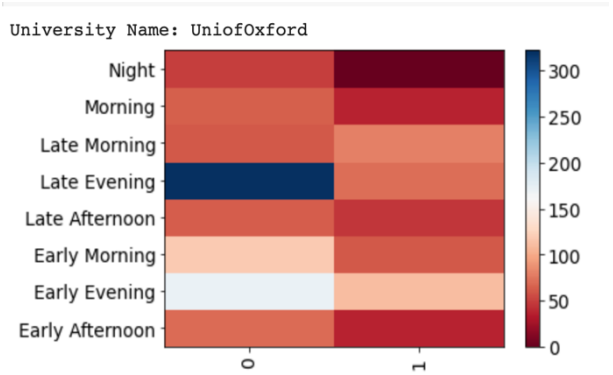
*However, **University of Oxford** received more liked during each part of the day as compared to **Cambridge University**.*

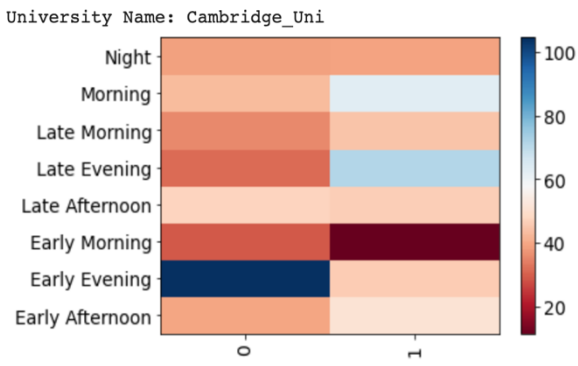
University Name: UniofOxford

| | favorite_count | |
|---|---|---|
| Negative | 0 | 1 |
| day_part | | |
| Early Afternoon | 70.027534 | 37.010870 |
| Early Evening | 171.033457 | 111.458333 |
| Early Morning | 118.340909 | 61.500000 |
| Late Afternoon | 62.965164 | 45.567568 |
| Late Evening | 321.686047 | 71.000000 |
| Late Morning | 61.238596 | 79.703704 |
| Morning | 64.350515 | 36.686869 |
| Night | 49.304348 | 0.000000 |

University Name: Cambridge_Uni

| | favorite_count | |
|---|---|---|
| Negative | 0 | 1 |
| day_part | | |
| Early Afternoon | 39.491525 | 51.134615 |
| Early Evening | 104.578182 | 46.133333 |
| Early Morning | 28.857143 | 11.250000 |
| Late Afternoon | 47.346296 | 46.276596 |
| Late Evening | 31.620690 | 71.750000 |
| Late Morning | 35.631902 | 44.514286 |
| Morning | 43.074444 | 63.055556 |
| Night | 38.805556 | 39.000000 |



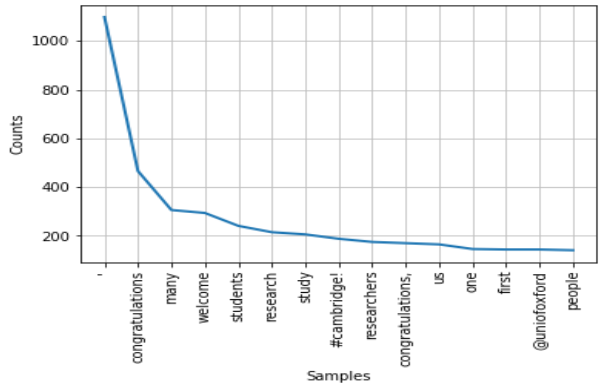

***Observation 2:*** *University of Oxford receives maximum likes on Negative comments in Early Evening (shown by skin color).*

*Cambridge University receives maximum likes on Negative comments in Late Evening (shown by light blue color).*

*However, the majority of like counts for both universities are on non-Negative comments and in the evening (shown by dark blue).*

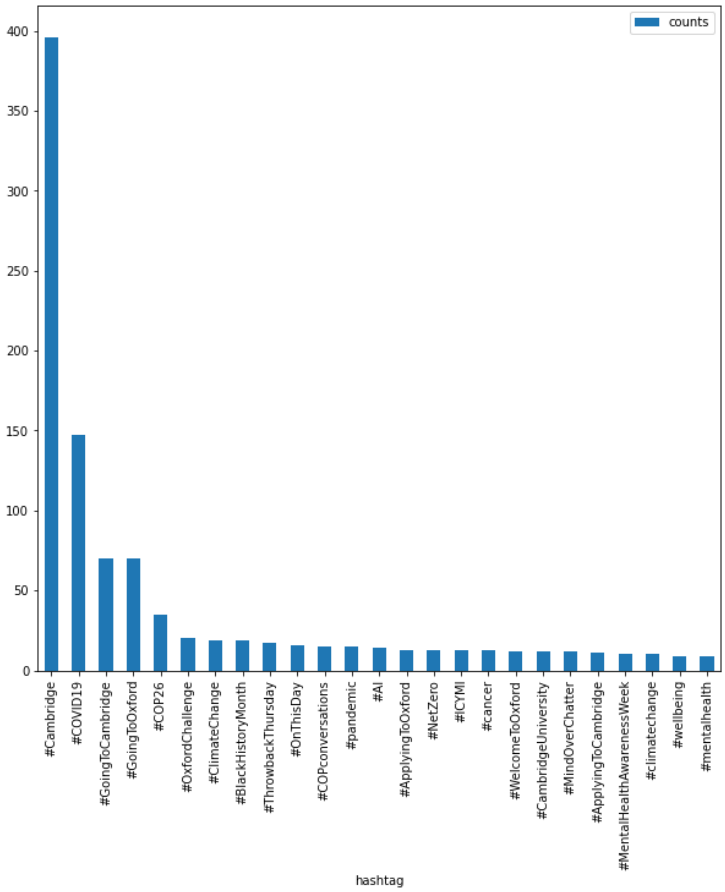
**Data Analysis – Beneficial to Student**

1. ***Which word and hashtag were used most in shared tweets?***




***Observation:*** *The analysis clearly shows that the universities mostly tweet when they are congratulating. This shows that the university is highly appreciative of the achievements of its staff and students. The universities also give special attention to proper welcoming. It can be of new students, professors etc.*
*Apart from these, like all universities, these 2 universities also tweet about students, studies and research.*

2. ***Which Hashtags do universities use mostly?***

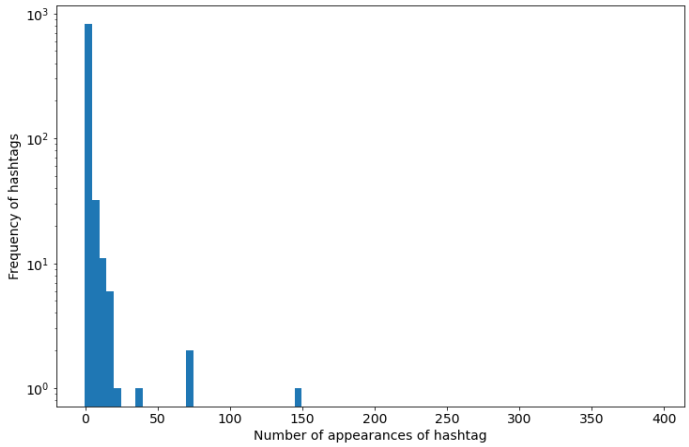



***Observation 1:*** *Out of the two universities, Cambridge university hashtag is more used than Oxford university. The top 5 hashtags are:*

1. *#Cambridge*
2. *#COVID19*
3. *#GoingToCambridge*
4. *#Going to Oxford*
5. *#COP26*

*Going to Cambridge hashtag is. More popular than Going to Oxford hashtag.*

*(COP26 is the 2021 United Nations Climate Change Conference)*

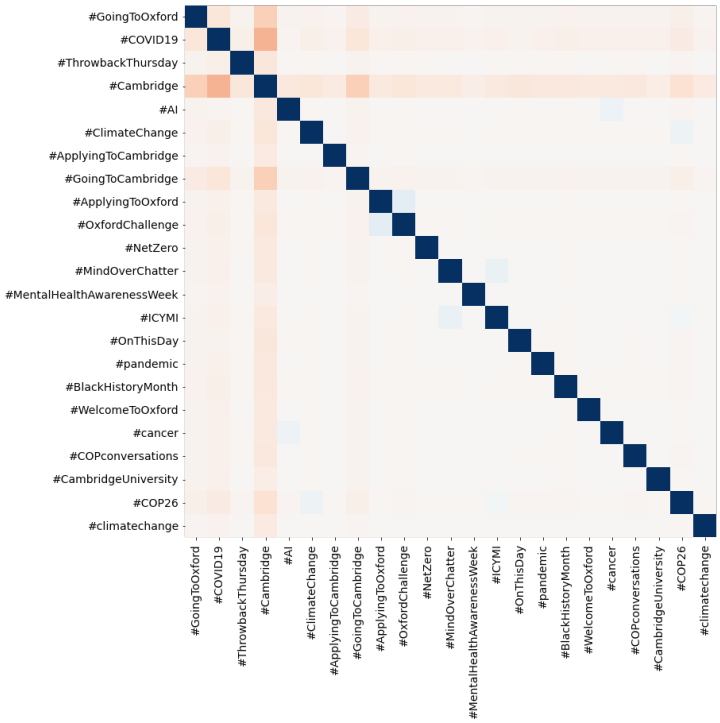



***Observation 2:*** *Hundreds of hashtags appear less than 10 times.*

*Less than 10 Hashtags reappear for around 75 times.*

*Only 1 or 2 Hashtags have a reappear count of 150 times.*

*These 1-2 hashtags with high number of appearances are most probably the university names.*




***Observation 3:*** *The topics with most correlation are (the ones in light blue):*

1. *#AI vs #cancer*
2. *#ClimateChange vs #COP26*
3. *#ApplyingToOxford vs #OxfordChallenge*
4. *#MindOverChatter vs #ICYMI*
5. *#ICYMI vs #COP26*

*(ICYMI – in case you missed it)*

## 3. *Which university is more active during the summer holidays?*

```
**********************************************************************************
Please, choose date interval for summer holidays to see some statistical information about tweets.
```

start_date 22/07/2022

end_date 04/09/2022

```
**********************************************************************************
UniofOxford published 341 tweets between 2022-07-22 and 2022-09-04.
These tweets have got 22310 likes and retweeted by 4871 totally.
----------------------------------------------------------------------------------
On the other hand, Cambridge_Uni published 317 tweets during this time.
These tweets have got 5993 likes and retweeted by 1533.
**********************************************************************************
```

*According to the figures of last year, the summer break in England was from 22 July to 4 September.*

*Both universities almost had the same number of tweets, although University of Oxford had slightly more. However,* ***significant more likes and retweets*** *are seen on tweets by University of Oxford than Cambridge University.*

**Conclusion**

University of Oxford is more popular and active than Cambridge University. One of the reasons can be more population. Another reason might be the difference of courses that both the universities offer.
Since in the recent years, there has been spread of COVID pandemic, it seems natural that #COVID19 being one of the most used hashtags.
It is also observed that special focus has been given to the 2021 Climate Change Conference and it has been actively promoted by both the universities with hashtags like #ClimateChange and #COP26.
The last but not the least, interestingly, work in the field of Artificial Intelligence and cancer have been tweeted about, which shows good progression in the field of technology and healthcare.

**New Learnings**

1. Accessing Twitter API and collecting data.
2. IPython Widgets for interactive data exploration.
3. Sentiment Analysis using TextBlob Library
4. Assigning Dummy values to sentiments
5. Text Pre-processing and using Natural Language Toolkit (nltk)
6. Working with Hashtags
7. Variety of Visualizations (heatmaps, histogram, bar graph, line graph, pivot table)
8. Drawing conclusions from visualizations

**Future Scope**

The project could be further improved to answer more questions related to tweets by universities. Some of them are:

1. Comparison between most used hashtags and words by different universities could be included.
2. Topic Modelling could be done on the text of tweets, to draw more conclusive results.
3. Based on suggestions from university social handle in-charge and the students, more analysis can be added depending upon what questions they would like to be answered.

## Thank You ☺