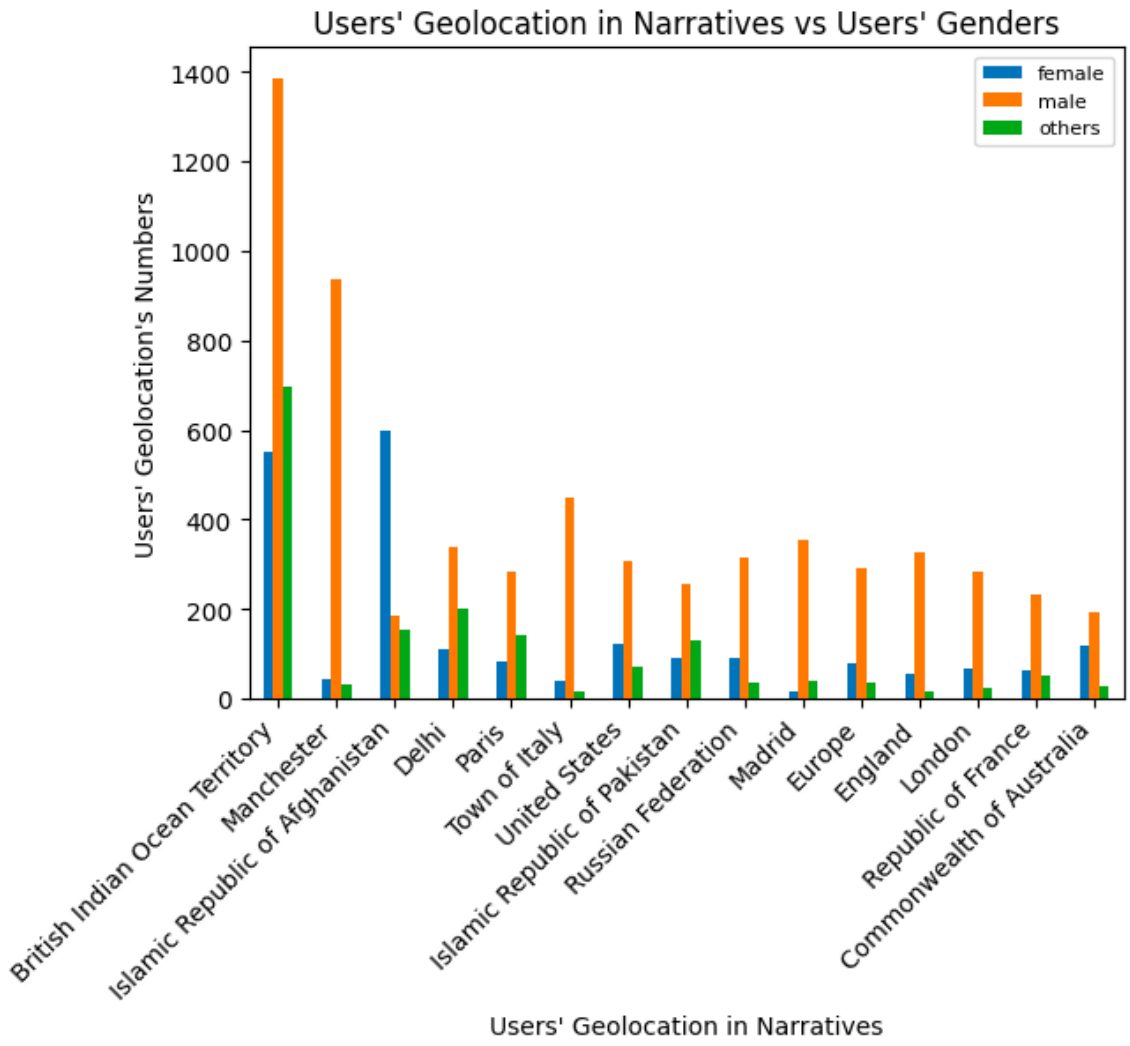
# Report – Assignment 2 Team 3

Jimin Ding  1556886350
Xiaoyu Dong 2468117466
Hui Qi 3206742781
Mingyu Zong 3484496941

**1. Are there any age or gender or topic based correlations by location in the posts?**

We selected the posts containing location entities and dropped those whose users hide their genders to analyze the relationships between them. From the figure below, for each geolocation except the Islamic Republic of Afghanistan, the users' gender distribution has more males than females. It is reasonable considering that the users of this PixStory application are more males than females. However, many more female users mention the "Islamic Republic of Afghanistan" in their posts. We checked several posts having this geolocation, and many of them are related to "horrific attack", "policy", "economy", and "death". This finding might lead to an essential and pondering research question of whether females are more concerned, care about, and express their thoughts about human rights,  freedom, and national peace in their private posts on social media.

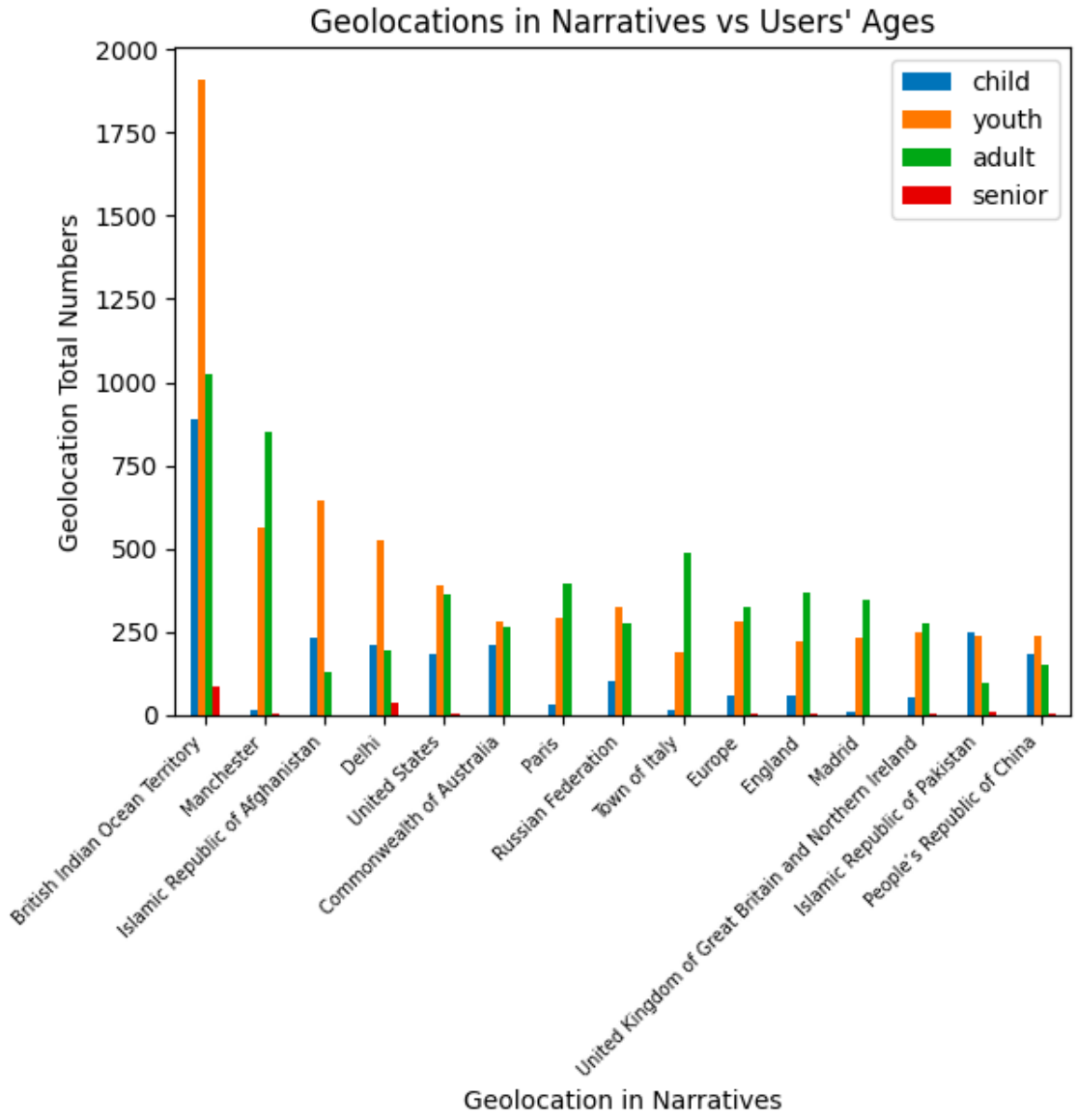
Moreover, posts including Manchester mostly come from males, while only a few posts about it from females and the other gender. We checked several users' narratives, and most are related to "Manchester United", which is a football club. Considering that users whose interest topic is football are mostly males from assignment one, we are likely to understand this phenomenon. This phenomenon might inspire the operation administration at PixStory to further analyze and create attractive marketing methods by building space for football lovers and constructing specific events on social media to celebrate the games.



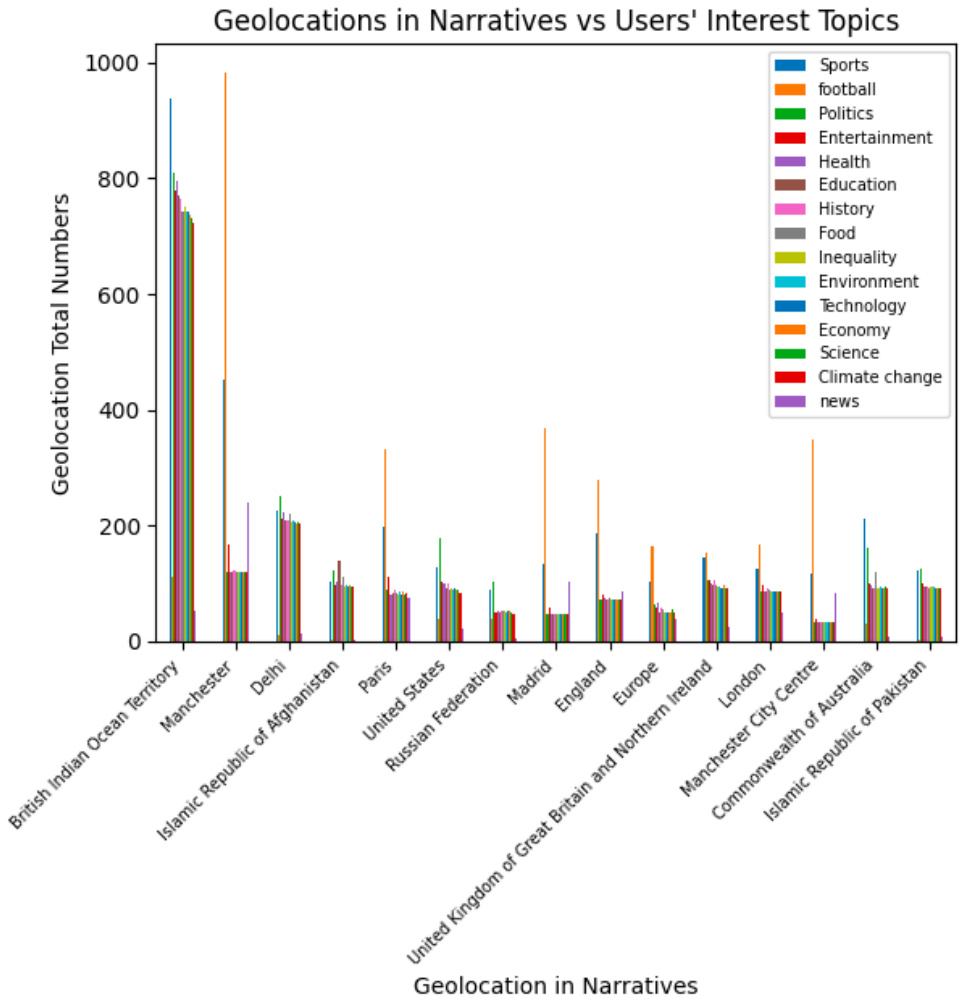

We are interested in the relationship between users' geolocation mentioned and their ages. We classify users into four age groups: ages between 5 and 14 are children; ages between 15 and 24 are youths; ages between 25 and 64 are adults; ages between 65 and 100 are seniors. Some users with ages over 100 or below 5 are removed from this analysis since this data is spurious, as some users might set false and exaggerated information. We pick the 15 location entities most users include among all the geolocations from texts.

In narratives containing some European countries or locations, like Manchester, Paris, Europe, England, Madrid, and the United Kingdom of Great Britain and Northern Ireland, whose authors are more likely to be adults than youths. What's more, the authors of posts mentioning location entities of British Indian Ocean Territory, the Islamic Republic of Afghanistan, and Delhi are more youths than adults. These three locations are in Midwest and southern Asia. It could lead to

further research to determine how this correlation could help us identify users' behaviors and identities while building a more peaceful and positive social media environment.

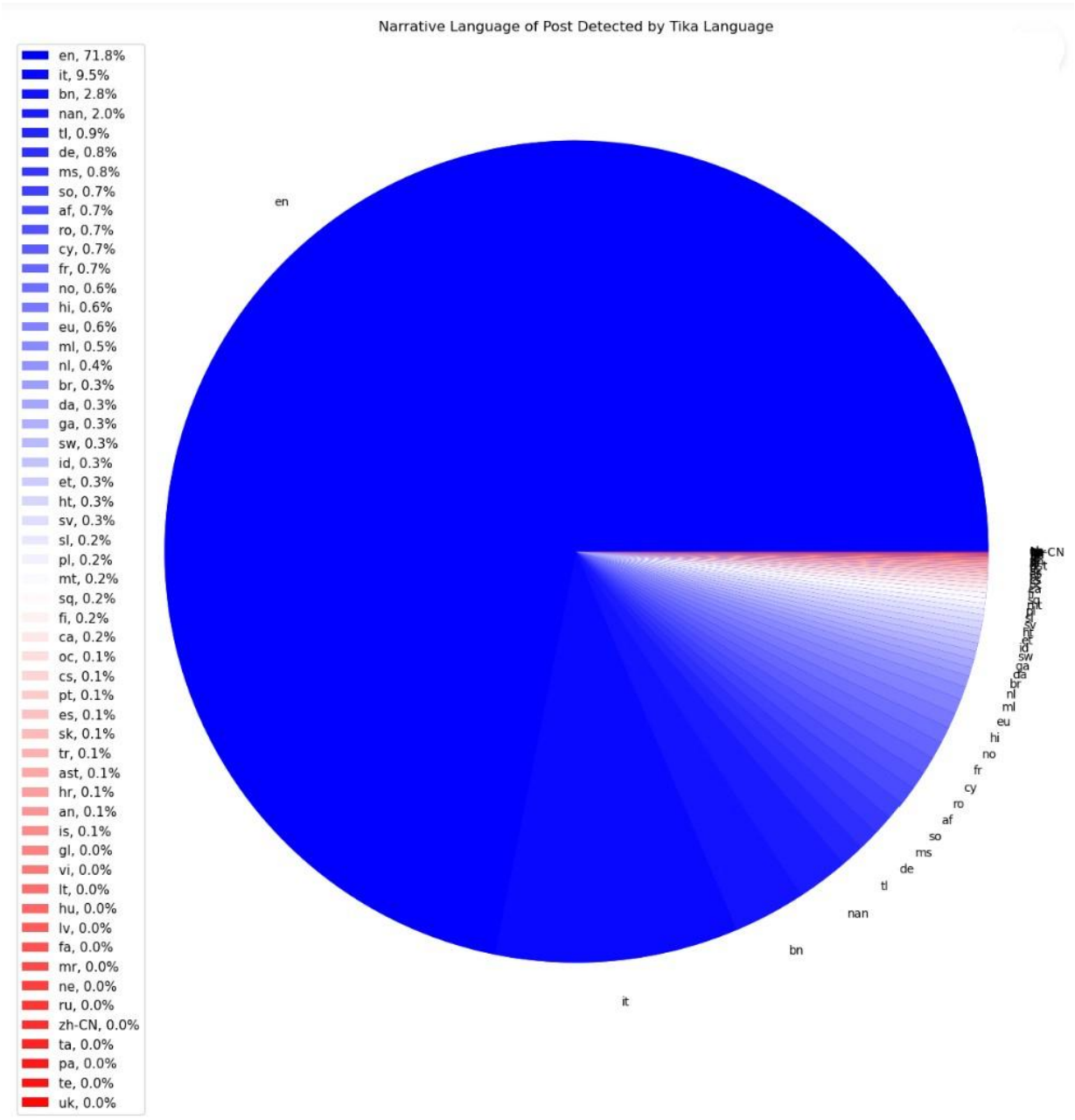



From the figure about users' interests and geolocation mentions, we first selected the top 15 locations mentioned in narratives. Then, under this condition, we picked the 15 most popular users' interests and plotted the graph below. Users who wrote Manchester in their posts have a favor in football, which Manchester here likely refers to Manchester United Football Club. Moreover, users are more interested in football than other topics for those who mention Paris, Madrid, England, Europe, and Manchester City Centre in the application. Football events likely happened in some of those areas, like Six Nations. However, British Indian Ocean Territory seems to have little relationship with sports. Therefore, we need more information to gain a better conclusion about correlations between users' interests and the locations mentioned.




**2. What is the most prevalent language in the posts, and least prevalent?**

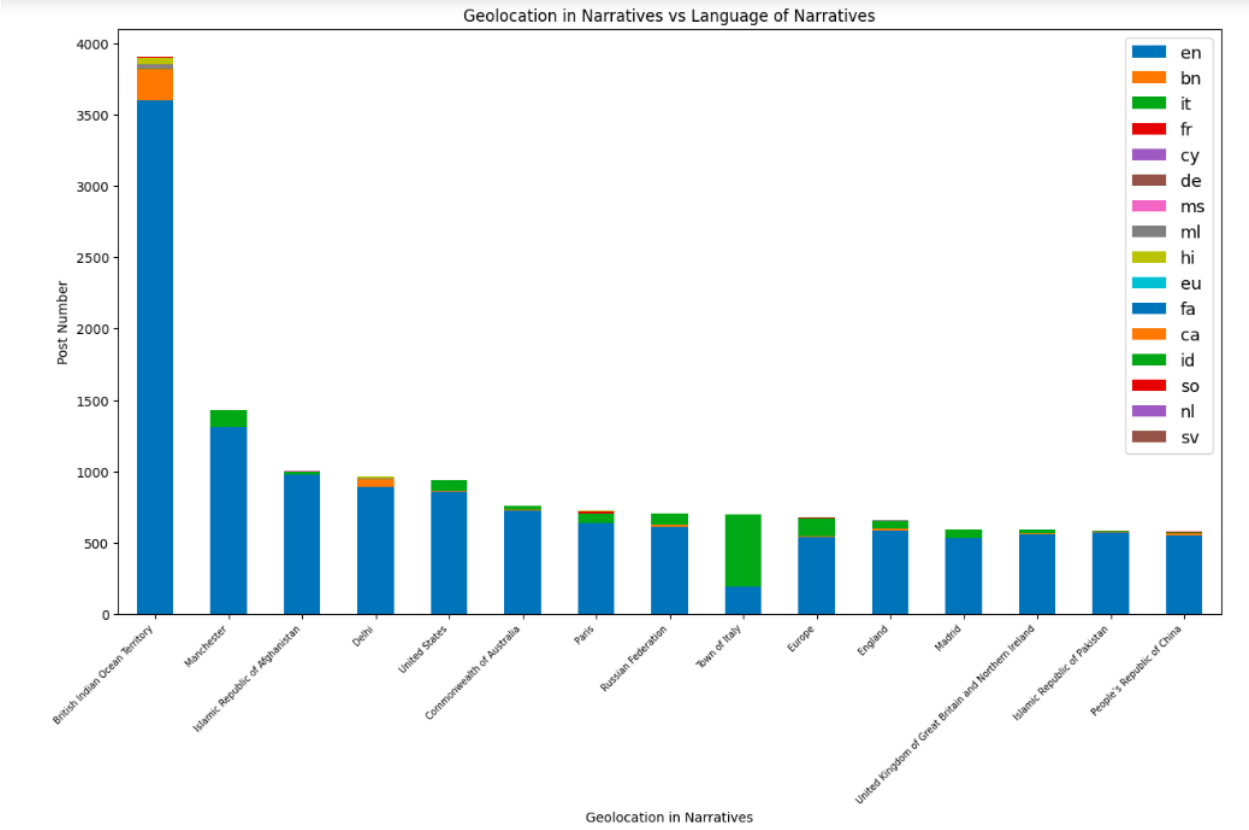
According to Tika Language, the most prevalent languages are English (71.8%), Italian (9.5%) and Bengali (2.8%). Most of the remaining languages are between 0.1% and 1.0%. The least prevalent languages are Galician (gl), Vietnamese (vi), Lithuanian (lt), Hungarian (hu), Latvian (lv), Persian (fa), Marathi (mr), Nepali (ne), Russian (ru), Chinese (zh-CN), Tamil (ta), Punjabi (pa), Telugu (te), Ukrainian (uk), they are all around  0.0%.




**3. Is there a correlation between post language and identified mentioned locations?**

We selected the top 15 mentioned geolocation, and the top 15 primarily used language in the figure below to investigate their relationships. English is the most broadly used language in the PixStory application. Remarkably, British Indian Ocean Territory has a considerable amount of English posts compared to other geolocation mentioned. In addition, users wrote in Bengali, containing more narratives about British Indian Ocean Territory and Delhi. Delhi is the national capital of India. Moreover, according to Wikipedia, in India, Bengali is the official language of West Bengal, Tripura, and the Barak Valley region of Assam. Therefore, the usage of the Bengali

language is closely related to the geolocation mentioned in narratives because most posts written in Bengali are associated with India, like British Indian Ocean Territory and Delhi. Furthermore, narratives containing towns of Italy have more users written in Italian than English. It might imply that there is a considerably large number of users of this application from Italy, where they have a social hub more restricted to people from Italy. Also, geolocations in Europe, like Paris, Europe, and England, have many posts written in Italian. In short, there is a correlation between post languages and identified locations.




**4. Are there correlations between the sporting events, or the entertainment events with locations?**

We picked the top 15 entertainment events and want to determine which geolocation is mentioned most in the narrative under each film event. Our entertainment event includes "No events during the post's day", represented by "NA" in our data and below. The first entity in each row or line in the figure below is the entertainment event, and the second entity is the most geolocation mentioned during the event period. At the same time, the third entity is the number of posts that satisfy the entertainment and geolocation in front of it.

The Pan African film & Arts Festival 2022 happened in Los Angeles, which was not in Australia while we checked in the dataset that only 2 posts mentioned "Los Angeles" during this film

event. What's more, there are many "British Indian Ocean Territory" locations in the figure which are not related to the event.

Among all the film events below, only SXSW 2022 successfully matched the geolocation results from the posts. SXSW 2022 happened in Austin, TX, United States, which reached the most geolocation entities "United States" mentioned during those periods.

Since only one event out of 15 events achieved a successful match, there is not enough evidence from this dataset and our analysis strategy to find the correlations between entertainment events and locations in the narratives.

```
NA  :  British Indian Ocean Territory 1581.0

Pan African Film & Arts Festival 2022  :  Commonwealth of Australia 97.0

Santa Barbara International Film Festival 2022, Glasgow International Film Festival 2022  :
British Indian Ocean Territory 175.0

Pan African Film & Arts Festival 2022, Atlanta Film Festival 2022, Milwaukee Film Festival 20
22  :  Commonwealth of Australia 165.0

Sundance Film Festival 2022  :  Manchester 133.0

SXSW 2022  :  United States 80.0

Melbourne Documentary Film Festival 2022  :  British Indian Ocean Territory 210.0

52nd International Film Festival of India  :  Los Angeles 92.0

2021 Sundance Film Festival  :  Russian Federation 106.0

2021 Metro Manila Film Festival (MMFF)  :  Manchester 97.0

Venice International Film Festival 2022, Telluride Film Festival 2022  :  Manchester 67.0

59th New York Film Festival, 2021 Vancouver International Film Festival, 26th Busan Internati
onal Film Festival  :  Islamic Republic of Afghanistan 344.0

Venice International Film Festival 2022  :  Manchester 65.0

Docaviv 2022, International Film Festival Cologne 2022, Doc Edge 2022  :  British Indian Ocea
n Territory 31.0

Berlin International Film Festival 2022  :  Commonwealth of Australia 72.0
```

Similarly, We picked the top 15 sports events, including "No sport events during the post's day" represented by "NA", and want to determine which geolocation is mentioned most in the narrative under each sport event. In the figure below, the first entity in each row or line in the figure below is the sporting event, and the second entity is the most geolocation mentioned

during the event period. At the same time, the third entity is the number of posts that satisfy the sports event and the geolocation in front of it.

Six Nations is the Rugby union competition between England, France, Ireland, Italy, Scotland, and Wales. Although this sport is not directly related to Manchester, Manchester might indirectly relate to this competition since it is in northwestern England. We checked other sports events and geolocation in the results below, but all events did not happen in the geolocation mentioned in most posts within the same periods. Therefore, based on this method, there is little relationship between sports events and geolocations from the narratives. But still, it is possible that we could get more evidence if we find other analysis strategies and improve the data collection.

```
NA  :  British Indian Ocean Territory 1581.0

Six Nations  :  Manchester 363.0

Australian Open, Africa Cup of Nations  :  Manchester 272.0

World Snooker Championship, Invictus Games  :  Commonwealth of Australia 224.0

US Open, World Volleyball Championships (men)  :  Manchester 137.0

Six Nations, Winter Olympics  :  Hollywood 103.0

Six Nations, Winter Paralympics, ODI World Cup for women  :  British Indian Ocean Territory 1
90.0

Six Nations, ODI World Cup for women  :  United States 80.0

World Championships (women)  :  British Indian Ocean Territory 177.0

Africa Cup of Nations  :  British Indian Ocean Territory 55.0

Paralympic Games  :  Islamic Republic of Afghanistan 190.0

World Snooker Championship  :  Commonwealth of Australia 38.0

IIHF World Championship, South-East Asian Games, World Masters Games (Summer)  :  British Ind
ian Ocean Territory 96.0

French Open  :  United States 23.0

Nordic World Ski Championships  :  Russian Federation 51.0
```

**5. Do the Detoxify scores and associated GLAAD and ADL or sarcasm flags line up? Is there any relationship between the flags and the identified Detoxify scores?**

To explore whether or not Detoxify scores are associated with GLAAD and ADL or sarcasm flags, we extracted 4 sub-dataframe from the complete tsv: ‘hate’ contains all rows being labeled as hate speech, ‘nothate’ contains all rows that are not labeled as hate speech, ‘sarcasm’ has all instances in which we detected sarcasm using keywords, and ‘notsarcasm’ is the complementary dataset of ‘sarcasm’. As a result, ‘hate’ has 1246 rows, ‘nothate’ has 93754, ‘sarcasm’ has 3498, and ‘notsarcasm’ has 91502 rows.

We decided to set the cutoff to be 0.5 for all 7 different scores (‘Toxicity’, ‘Severe_Toxicity’, ‘Obscenity’, 'Identity_Attack', 'Insult', 'Threat', 'Sexual_Explicit'). By doing so, we noticed that there is only one problematic post in the ‘hate’ dataframe; while in the ‘nothate’ dataframe, 517 out of 93754 are considered problematic. More specifically, the only problematic post recognized by Detoxify in ‘hate’ is toxic, obscene, and insulting. Generally hate speech can also belong to any of the 7 categories from Detoxify, but those with the highest scores from each category are not labeled as hate speech by our previous keywords matching approach. Similar things happened to the ‘sarcasm’ and ‘notsarcasm’ datasets. Detoxify scores indicate only 46 posts in the ‘sarcasm’ dataset are problematic. On the contrary, 472 posts in the ‘notsarcasm’ are recognized by it, which is roughly ten times the number we have for the ‘sarcasm’ dataset. Other statistics show that posts in ‘notsarcasm’ have higher average scores for ‘Toxicity’, ‘Obscenity’, 'Insult', and 'Threat'. Although ‘sarcasm’ has higher average scores for the other few columns, the only one resulting in a significant difference is ‘Sexual_Explicit’. However, given that the ‘Sexual_Explicit’ highest score is derived from a post in ‘notsarcasm’, we still can’t say with confidence that the pattern matching method agrees with the Detoxify approach on identifying sexually explicit contents.

| | Toxicity | Severe_Toxicity | Obscenity | Identity_Attack | Insult | Threat | Sexual_Explicit |
|---|---|---|---|---|---|---|---|
| count | 1.000000 | 1.00000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| mean | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |
| std | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| min | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |
| 25% | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |
| 50% | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |
| 75% | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |
| max | 0.959638 | 0.00933 | 0.923579 | 0.016089 | 0.648131 | 0.000523 | 0.003306 |

```
Number of toxic post: 1
Number of severely toxic post: 0
Number of obscene post: 1
Number of indentity attacking post: 0
Number of insulting post: 1
Number of threatening post: 0
Number of sexual explicit post: 0
```

Statistics of ‘hate’ dataframe

| | Toxicity | Severe_Toxicity | Obscenity | Identity_Attack | Insult | Threat | Sexual_Explicit |
|---|---|---|---|---|---|---|---|
| count | 517.000000 | 517.000000 | 517.000000 | 517.000000 | 517.000000 | 517.000000 | 517.000000 |
| mean | 0.674935 | 0.004720 | 0.264285 | 0.031538 | 0.285213 | 0.028175 | 0.193153 |
| std | 0.227223 | 0.019292 | 0.319771 | 0.104491 | 0.320616 | 0.113786 | 0.297567 |
| min | 0.104771 | 0.000006 | 0.000198 | 0.000203 | 0.002077 | 0.000027 | 0.000043 |
| 25% | 0.526376 | 0.000230 | 0.007663 | 0.001944 | 0.023750 | 0.000247 | 0.001104 |
| 50% | 0.689446 | 0.000778 | 0.075555 | 0.005056 | 0.123025 | 0.000618 | 0.006875 |
| 75% | 0.876301 | 0.002164 | 0.536766 | 0.012491 | 0.517514 | 0.001326 | 0.414062 |
| max | 0.996834 | 0.203792 | 0.991341 | 0.818293 | 0.989978 | 0.893525 | 0.979708 |

```
Number of toxic post: 427
Number of severely toxic post: 0
Number of obscene post: 143
Number of indentity attacking post: 11
Number of insulting post: 137
Number of threatening post: 8
Number of sexual explicit post: 127
```

Statistics of ‘nothate’ dataframe

| | Toxicity | Severe_Toxicity | Obscenity | Identity_Attack | Insult | Threat | Sexual_Explicit |
|---|---|---|---|---|---|---|---|
| count | 46.000000 | 46.000000 | 46.000000 | 46.000000 | 46.000000 | 46.000000 | 46.000000 |
| mean | 0.501504 | 0.006913 | 0.163199 | 0.045363 | 0.093611 | 0.001086 | 0.448252 |
| std | 0.274283 | 0.030057 | 0.303007 | 0.113547 | 0.207113 | 0.001025 | 0.283464 |
| min | 0.141950 | 0.000110 | 0.001393 | 0.001734 | 0.002077 | 0.000142 | 0.000588 |
| 25% | 0.294260 | 0.000458 | 0.010941 | 0.006080 | 0.004761 | 0.000613 | 0.137056 |
| 50% | 0.406188 | 0.000725 | 0.026720 | 0.010102 | 0.008144 | 0.000792 | 0.540824 |
| 75% | 0.703541 | 0.001749 | 0.119006 | 0.017642 | 0.077195 | 0.001187 | 0.622375 |
| max | 0.990867 | 0.203792 | 0.991341 | 0.524751 | 0.924350 | 0.006177 | 0.959312 |

```
Number of toxic post: 18
Number of severely toxic post: 0
Number of obscene post: 6
Number of indentity attacking post: 2
Number of insulting post: 3
Number of threatening post: 0
Number of sexual explicit post: 32
```

Statistics of ‘sarcasm’ dataframe

| | Toxicity | Severe_Toxicity | Obscenity | Identity_Attack | Insult | Threat | Sexual_Explicit |
|---|---|---|---|---|---|---|---|
| count | 472.000000 | 472.000000 | 472.000000 | 472.000000 | 472.000000 | 472.000000 | 472.000000 |
| mean | 0.692441 | 0.004516 | 0.275533 | 0.030158 | 0.304654 | 0.030757 | 0.167890 |
| std | 0.215180 | 0.017915 | 0.321013 | 0.103487 | 0.323779 | 0.118775 | 0.287072 |
| min | 0.104771 | 0.000006 | 0.000198 | 0.000203 | 0.002294 | 0.000027 | 0.000043 |
| 25% | 0.537275 | 0.000183 | 0.006686 | 0.001863 | 0.036282 | 0.000228 | 0.000857 |
| 50% | 0.703929 | 0.000792 | 0.089180 | 0.004455 | 0.141490 | 0.000547 | 0.005486 |
| 75% | 0.883370 | 0.002221 | 0.544501 | 0.012002 | 0.526686 | 0.001353 | 0.188866 |
| max | 0.996834 | 0.203616 | 0.989618 | 0.818293 | 0.989978 | 0.893525 | 0.979708 |

```
Number of toxic post: 410
Number of severely toxic post: 0
Number of obscene post: 138
Number of indentity attacking post: 9
Number of insulting post: 135
Number of threatening post: 8
Number of sexual explicit post: 95
```

Statistics of ‘notsarcasm’ dataframe

Overall, out of 95000 instances, hate speech detection by pattern matching and Detoxify agreed on 1 positive case and 93237 negative cases, sarcasm detection by pattern matching and Detoxify agreed on 46 positive cases and 91030 negative cases. According to these numbers, we claim that there is no strong alignment or association between either hate speech flags or sarcasm flags and the Detoxify scores across 7 categories.

Codes for analysis are stored in Detoxify_Analysis.ipynb under the folder ‘Step11 - Detoxify’.

**6. Do the image captions accurately represent the image?**

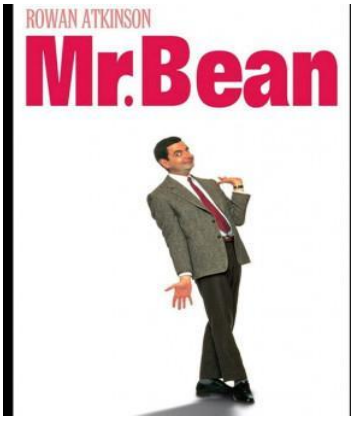




Image captions can provide valuable information on the broader patterns of images, such as background, color, and shape. However, they may not be accurate when it comes to identifying specific details like smaller objects, text, and abstract symbols. For example, in the first image, the caption correctly identifies the beach as the background, but reports the presence of a person that does not exist in the image. In the second image, the caption accurately detects a man in a suit but misses important text information "Mr. Bean" in the image. Similarly, in the third image, the caption mistakenly interprets the daisy as a pair of scissors. For symbols in image four, caption detection does not perform properly. It describes a cartoon panda as “a man holding a white frisbee in his right hand” instead. The color is correct, but the content is different.

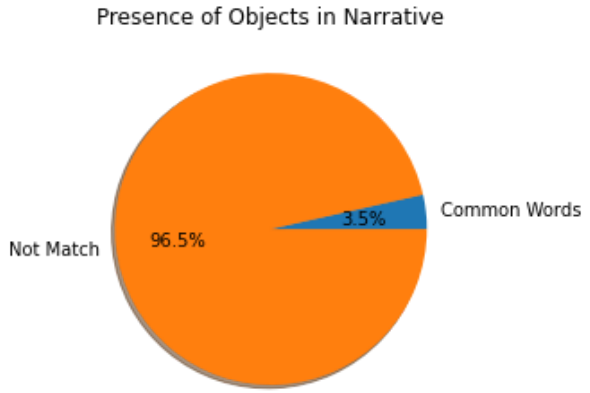
**Thoughts using Tika Image Captioning and Inception Rest service**

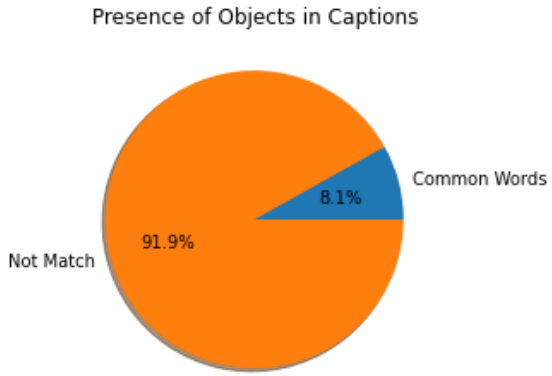
When we use Tika image captioning, it offers a brief overall description of the image. It is significantly faster with multi-threading. However, we can improve its performance by combining it with Tesseract or other parsers to extract text from images and accurately capture more important details. Additionally, by combining caption and object detection, we can get a more precise and accurate description of the image, such as in the third example, where the caption identifies a person holding a pair of scissors, and object detection identifies the presence of a daisy. By combining these results, we can get a more accurate description of the image, such as "a person holding a daisy in their hands."

**7. Are the identified objects present in the image described in the original post and/or the generated caption?**

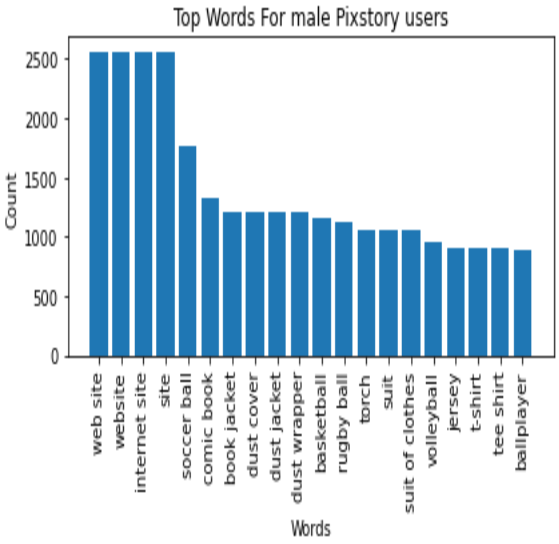
We selected image captions with the highest confidence level. From a dataset of 95,000 posts, we found that 3.5% of the identified objects were in the narrative, while 8.1% of the identified objects were in the generated captions. To identify matches, we converted all words to lowercase and compared them directly with the identified objects, narrative, and captions.

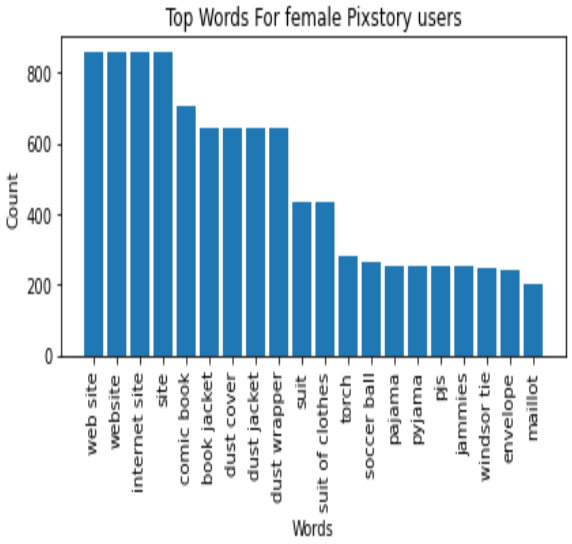
We observed that there were more common words between the identified objects and captions than between the objects and narratives. This could be because the identified objects and captions provide a broad overview of the image, while the narratives contain more detailed information, as discussed in Question 6.

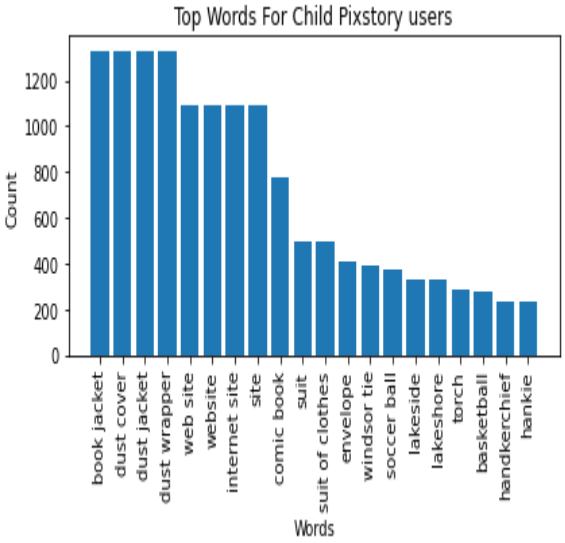


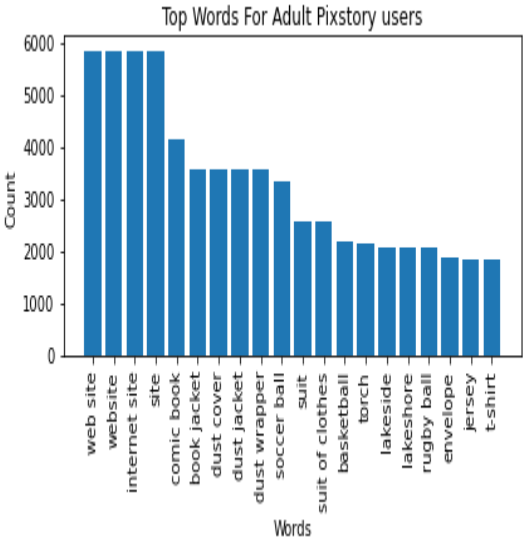



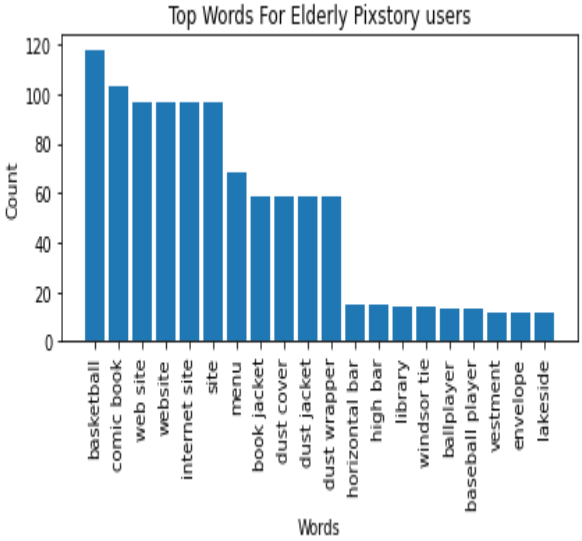

**8. Are there any age, or gender specific trends you see in the text captions or identified objects in the image media?**


Top Words For male Pixstory users
Count
Words
web site
website
internet site
site
soccer ball
comic book
book jacket
dust cover
dust jacket
dust wrapper
basketball
rugby ball
torch
suit
suit of clothes
volleyball
jersey
t-shirt
tee shirt
ballplayer


Top Words For female Pixstory users
Count
Words
web site
website
internet site
site
comic book
book jacket
dust cover
dust jacket
dust wrapper
suit
suit of clothes
torch
soccer ball
pajama
pyjama
pjs
jammies
windsor tie
envelope
maillot


Top Words For Child Pixstory users
Count
Words
book jacket
dust cover
dust jacket
dust wrapper
web site
website
internet site
site
comic book
suit
suit of clothes
envelope
windsor tie
soccer ball
lakeside
lakeshore
torch
basketball
handkerchief
hankie


Top Words For Adult Pixstory users
Count
Words
web site
website
internet site
site
comic book
book jacket
dust cover
dust jacket
dust wrapper
soccer ball
suit
suit of clothes
basketball
torch
lakeside
lakeshore
rugby ball
envelope
jersey
t-shirt


Top Words For Elderly Pixstory users
Count
Words
basketball
comic book
web site
website
internet site
site
menu
book jacket
dust cover
dust jacket
dust wrapper
horizontal bar
high bar
library
windsor tie
ballplayer
baseball player
vestment
envelope
lakeside

We first filtered the data frame by age and gender to obtain the detected objects for each group. We then counted the occurrence of each object and presented the results using a bar chart. We combined words with similar meanings, such as website and internet site; dust cover, dust jacket, and dust wrapper; and handkerchief and hankie.

For male Pixstory users, website, soccer ball, comic book, book jacket, dust cover, basketball, rugby ball, torch, suit, volleyball, jersey, t-shirt, and ballplayer are objects that appear frequently in their images. For female Pixstory users, their top object words are website, comic book, dust cover, suit, torch, soccer ball, pajama,, windsor tie, envelope, and maillot. For children who post images on PixStory, they focus on book jacket, dust cover, website, comic book, suit, envelope, windsor tie, soccer ball, lakeside, torch, basketball, and handkerchief. Adult users' images concentrate on website, comic book, book jacket, soccer ball, suit, basketball, torch, lakeside, rugby ball, envelop, jersey, and t-shirt. Elderly people post images more on basketball, comic book, website, menu, book jacket, horizontal bar, high bar, library, windsor tie, ballplayer, vestment, envelop, and lakeside.

The analysis showed that books and sports equipment are popular among all age groups and genders. Sports equipment, such as soccer balls, basketballs, rugby balls, and volleyballs, are particularly prevalent among male Pixstory users. Clothing is more commonly featured in images posted by females and older age groups. Female clothing tends to be more casual and varied, including items like pajamas and Windsor ties, while older individuals post more images of religious and formal clothing, such as vestments. Outdoor scenes, like lakeshores, are more commonly posted by children and adults, while charts with bars are more prevalent among elderly users. Additionally, fitness equipment is more popular among older individuals who are interested in maintaining their physical health.

**Also include your thoughts about the ML and Deep Learning software like RTG, GeoTopicParser, Detoxify, LangDetect, Tika Image Captioning, etc. – what was easy about using it? What wasn't?**

**About RTG:**
The biggest advantage of using RTG is that there is no translation limitation, which can easily handle batch translation. I have also used Google Translate before, though it is much faster than RTG, Google Translate only allows users to send about 100 requests per hour. However, RTG has significant disadvantages compared with other translation models. First, the translation speed. Compared RTG with Google Translate, which takes 10 seconds to complete a text of about 1,000 characters and RTG takes a minute to run. It is also easy to cause the timeout error when RTG encounters some uncommonly used languages such as Hindi and Bengali. Therefore, users need to manually slice the text and translate it in batches when they encounter the teaching text in a language that is not prevalent.

**About GeoTopicParser:**
Detecting the name entity recognition to find the locations mentioned in narratives is straightforward once the servers are set up. It is also fast running on 95000 instances compared to the run time of other steps in this assignment. It took me about 40 minutes to get the results. However, setting up the servers is complicated and can easily make mistakes. The instructions on the GeoTopicParser are not clear enough and not up to date. Although we do not need to know much about java, the server setting uses java, which caused trouble when I got stuck on the step to run the geotopic-server. I finally solved it by killing all java and restarting the process.

**About Tika Image Captioning and Inception Rest service:**
Mentioned above.

**About Detoxify:**
Developers of Detoxify made the library easy to install and use. On the package's documentation website, they included clear instructions and intuitive code examples which make their product first-time-user friendly. Besides, it works well with inputs in other languages. If a dataset doesn't have contents in uncovered languages, there will be no need to use an extra translator component.

The biggest issue with using the library is its long processing time. On our personal laptop, each post content that's translated into English requires around 3.5 seconds to be processed by the tool. If we perform single-threading processing on the 95000 instances, we should expect the whole procedure to take almost 4 days to complete. So we decided to further divide the dataset into 4 subsets and process them in parallel. In this way we shortened the processing time to ¼ of the original. Considering that datasets in this era can be much larger than the one we have for this task and that multithreading would only be a savior if the computer's CPU has multiple cores, this library is less ideal and less competitive for big data analysis.