CS410 Project Progress Report, CollegeEvents

NetID: Chiragr2 Name: Chirag Rastogi

1) Which tasks have been completed?

Data collection from multiple uiuc discord servers and from multiple instagram accounts. Example for instagram accounts:

acm.uiuc_insta_post	ACM@UIUC
inlivingcoloruiuc_insta_post	In Living Color
alpfa_uiuc_insta_post	ALPFA UIUC
aaja_uiuc_insta_post	AAJA UIUC
uiucakdphi_insta_post	UIUC
imagination_uiuc_insta_post	ImagiNation UIUC
unicef_uiuc_insta_post	UNICEF UIUC
uiuc_geoclub_insta_post	UIUC GeoClub
nsbe_uiuc_insta_post	NSBE UIUC
tsa_uiuc_insta_post	TSA UIUC
illinoistrialteam_insta_post	Illinois Trial Team
amnestyuiuc_insta_post	Amnesty International at UIUC
illinoisplhs_insta_post	UIUC Pre-Law Honors Society
ashauiuc_insta_post	Asha for Education UIUC
ieee.uiuc_insta_post	IEEE@UIUC
illinoisfsa_insta_post	UIUC Fraternities & Sororities
hack4impactuiuc_insta_post	Hack4Impact UIUC
kprojectuiuc_insta_post	K-Project UIUC

Preliminary time, location extraction: I will go into details while referring to this example:


```
{'ID': '2909171396040038898_207675938', 'startingTime': '2022-08-20 13:31:34',
'UserName': 'canopyclub', 'Name': 'The Canopy Club', 'media_type': 2,
'Location': '', 'Description': "ð\x9f§¡ð\x9f\x92\x99ð\x9f§¡ð\x9f\x92\x99 The
new semester officially kicks off tomorrow with our Unofficial Quad Day After
Party. Find us out on the quad and then hit the club right after. It's free
before 8pm so be there early. Three stages, all genres, food trucks and drink
specials. Let's do this UIUC!
\n\nâ\x97\x8f\nâ\x97\x8f\nâ\x97\x8f\n\n#CanopyClub #UIUC #Illini #Chambana
#Campus #CampusLife #StudentLife #FallSemester", 'Image_text': ['']}
```


ID: 2909171396040038898_207675938_insta_post
Name: The Canopy Club
startingTime: 2022/08/21 8:00 PM
Location: UIUC
Description: The new semester officially kicks off tomorrow with our Unofficial Quad Day After Party. Find us out on the quad and then hit the club right after. It's free before 8pm so be there early. Three stages, all genres, food trucks and drink specials. Let's do this UIUC! Life
Website/Link: []

Time Detection:
For time detection, I simply use a bert token classifier for temporal tagging. Given below are the list of tags I obtain. As seen in the example above, the classifier tags [8,pm] as time and [tomorrow] as date. I wrote some code to take the post time into consideration and find the date for words like today,tomorrow, monday,tuesday, etc.

```
O -- outside of a tag
I-TIME -- inside tag of time
B-TIME -- beginning tag of time
I-DATE -- inside tag of date
B-DATE -- beginning tag of date
I-DURATION -- inside tag of duration
B-DURATION -- beginning tag of duration
I-SET -- inside tag of the set
B-SET -- beginning tag of the set
```

Pending task: I chose the first time and date for a given message however I need to work on a ranking function that can choose the most appropriate time and date in the case of multiple values. This is probably not possible with simple statistical models and I may have to finetune the BERT model after creating a training dataset.

Location Detection:

For Location detection, I simply use Named entity recognition. Given below are the list of tags I obtain. As seen in the example above, the classifier tags UIUC as LOC or ORG.

tag	meaning
PER	person name
LOC	location name
ORG	organization name
MISC	other name

Pending task: I chose the first LOC or ORG for a given message however I need to work on a ranking function that can choose the most appropriate time and date in the case of multiple values. This is probably not possible with simple statistical models and I may have to finetune the NER model after creating a training dataset.

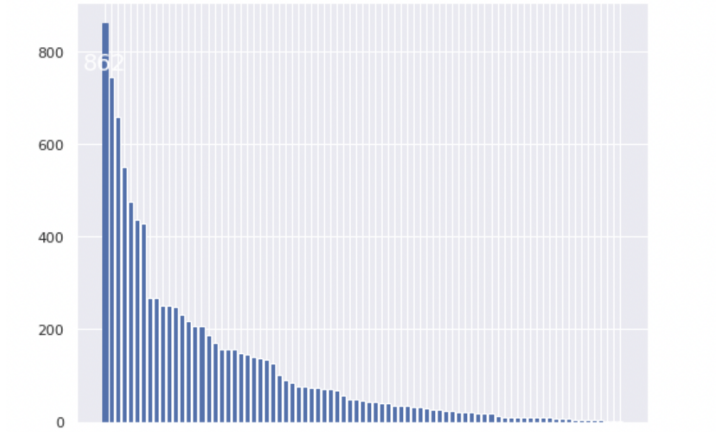
Challenges:

Please let me know if there is a better way to perform location extraction as location is very dependent on the context of the account. The only approach I see is finetuning the model and for that I would like to ask, would 500 labeled documents be enough?

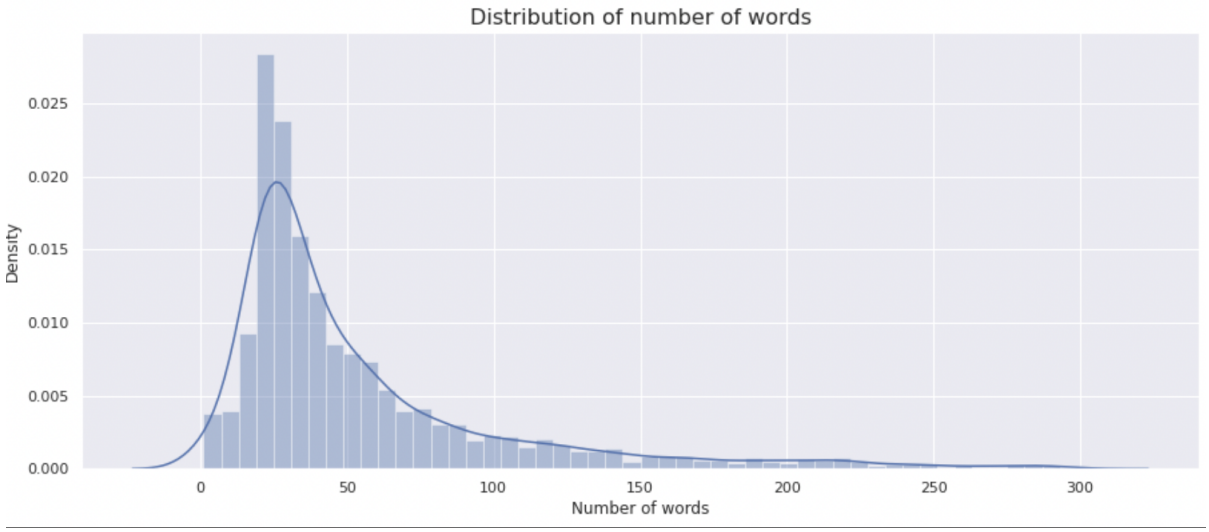
References:
https://huggingface.co/satyaalmasian/temporal_tagger_BERT_tokenclassifier
https://huggingface.co/flair/ner-english-large

Classification combining LDA and Word2Vec (Or Bert)

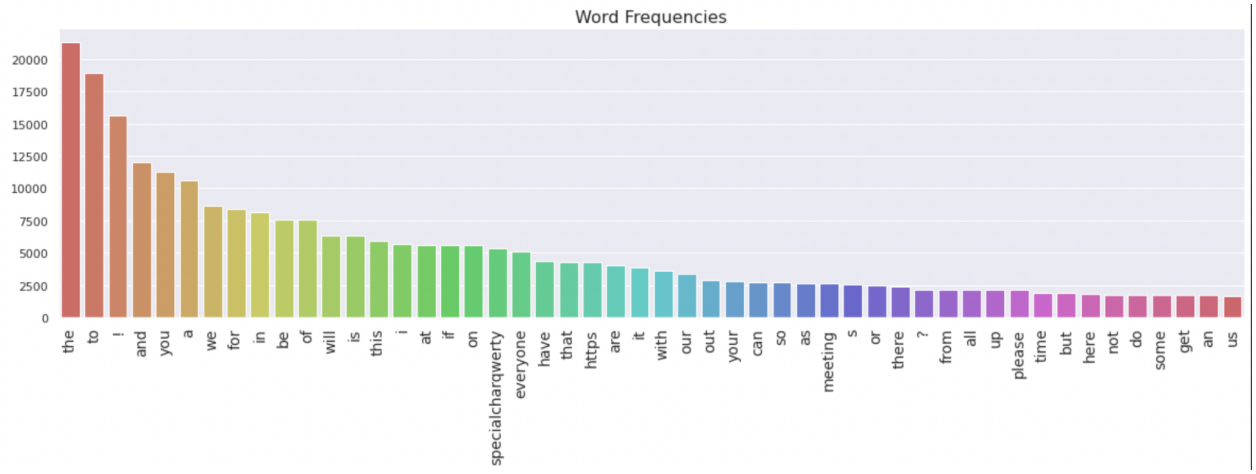
In the table below, we can see the number of documents (messages) per server. As we can see, this number is not uniform, and I need to work on creating a more uniform distribution. However, this may not be a problem as I will be using unsupervised clustering for different categories.

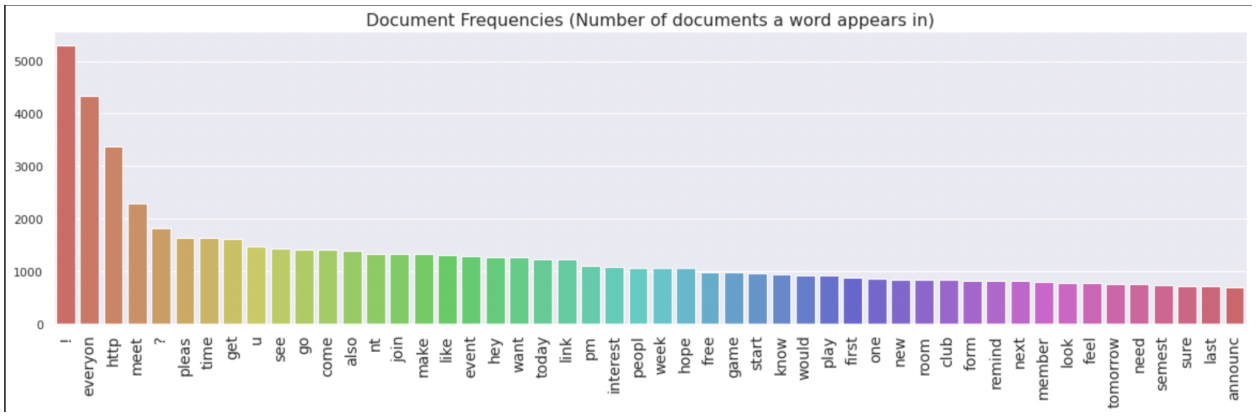


I limited documents to less than 300 words per document to remove outliers.



The next step was feature creation. Here I have tokenized the text, removed stop words, used stemming and finally vectorized the words. We can see a before and after view of Word frequencies. As we can see, the, to, etc has been removed and the words have been stemmed.





Pending:
The technique I am referring to trains the model using labels, however, due to the absence of labels I am trying to simply use LDA with word2vec to create clusters.
I still have to create the clusters and label a few documents to get categories. If needed, I will create labeled data.

Reference: https://www.kaggle.com/code/vukglisovic/classification-combining-lda-and-word2vec