Abstractive Summarization

Mandate-2

1. Problem Statement :

Given a set of tweets pertaining to a trending topic, create an abstractive prose summary of the tweets. Do not just string the tweets together to form the summary. The summary will need to paraphrase and/or say more than what is directly said in the tweets. Propose a rubric to evaluate the accuracy of your summarization.

2.Generation of Dataset :

- For corpus collection it's important to identify the relevant source and the relevance of collected content to the summary. As we have defined our problem statement over twitter tweets so we are collecting data from twitter.
- We scrape tweets of a particular hashtag .
- For that we have used an available automated platform like Apify (https://console.apify.com/) to extract the data as a CSV file of a particular hashtag or urls etc . We can also use available libraries like Python's Snscrape , BeautifulSoup ,Tweepy, Scrapy etc using twitter API.
- Corpus generated -
  1. #rammandir - 629 X 181
  2. #elonmusk -1000 X 251
- Some other already available datasets -

1. https://github.com/kavgan/opinosis-summarization (Graph algorithm based summarization framework
2. https://github.com/guyfe/Tweetsumm (A dataset focusing on summarization of dialogs, which represents the rich domain of Twitter customer care conversations and many more.)

3.Preprocessing of Dataset :

Since we have created the corpus so we need lots of data cleaning and preprocessing before applying models.

- Data Cleaning : Since our dataset contains multiple columns of fields like date, user profile , number of likes etc which are useless for our problem so we have dropped those columns and kept only one column consisting of tweets.
- Exploratory Data Analysis : To analyze the corpus.
- Preprocessing of Text : We have done following preprocessing steps
  1- Lower case conversion of all tweets
  2- Removal of all urls from tweets
  3- Removal of hashtags and mentions
  4- Removal of emojis
  5- Removal of tweets belonging to other language except english (unnecessary rows removal)
  6- Removal of punctuations

Following above steps decreases our dataset from 1000 X 251 to 686 X 1 . Only 1 column consisting of tweets and all tweets are of english language only.
Python libraries that we have used -

- Pandas
- Numpy
- Nltk
- Re
- Cleantext
- Langdetect

```
[17] data=data.drop(['conversation_id','created_at','favorite_count'],axis=1)

data.head()

[19] data=data.drop(data.iloc[:, 1:], axis=1)

[20] data.head()
```

	full_text
0	@DrEliDavid @elonmusk Can one of you clowns as...
1	@elonmusk Lie
2	@wmanastasio @saylor @elonmusk When we are FOR...
3	@CandeOchoaA @dariocelise @omarcervantesr9 @lo...
4	@elonmusk is again the richest man on earth. T...

```
[21] data.shape

(1000, 1)
```

```
[ ] def processing(text):
        text=lower_case(text)
        text=remove_emojis(text)
        text=remove_links(text)
        text=remove_hashtags(text)
        #text=remove_stopword(text)
        return text

#sample example
processing("How tommorrow  https://www.freecodecamp.org/news/python-web-scraping-tutorial/ #rammm @rammm 😂

'how tommorrow    '

[ ] processed_data=[]
    for i in range(len(data)):
        processed_data.append(processing(data['full_text'][i]))

[ ] !pip install langdetect
```

Lexical Processing - In NLP lexical processing refers to the process of analyzing words in a text. It is done to transform the raw , unstructured

text data into structured data which we can further analyze . For abstractive text summarization lexical preprocessing is one of the important steps .
It includes - Tokenization , Lemmatization , Stemming , Part of Speech Tagging (POS) ,Word sense disambiguation, Word Embeddings etc.

4. Tokenization-

Using Python's NLTK library we have tokenized the final cleaned and preprocessed corpus .Basically tokenization means to break sentences (tweets) into further smaller units called tokens . Here we are doing word tokenization.
Apart from NLTK we can also use Keras, Gensim to accomplish the task.
For example, the text "He is crying" can be tokenized into 'He', 'is', 'crying'.

*Part of Speech Tagging (POS) - We used the NLTK library to tag each token with the part of speech it belongs to. It is used to assign grammatical information to each word.
Ex- [('car', ' NN')] - Here NN represents a noun.

```
[63] POS_processed_data=[]
     for x in f_processed_data :
       tokens = word_tokenize(x)
       pos_tags = nltk.pos_tag(tokens)
       POS_processed_data.append(pos_tags)
```

```
POS_processed_data[0]

[('can', 'MD'),
 ('one', 'CD'),
 ('of', 'IN'),
 ('you', 'PRP'),
 ('clowns', 'VBP'),
 ('ask', 'VB'),
 ('why', 'WRB'),
 ('he', 'PRP'),
 ('went', 'VBD'),
 ('to', 'TO'),
 ('epstein', 'VB'),
 ("'s", 'POS'),
 ('island', 'NN'),
 ('so', 'RB'),
 ('many', 'JJ'),
 ('times', 'NNS'),
```

5.Embeddings -

We are transforming words into a numerical representation as float vectors.
Word Embedding Techniques -Word2Vec, Glove
Vectorization- TF-IDF , Bag of words (useful for text classification, clustering etc)
Both Word Embedding and Vectorization transforms human understandable english words into machine readable vectors but Word Embedding captures semantic relationship between the words and hence it is preferred for summarization tasks.

Python libraries that we have used -

- Spacy

```
[ ] import spacy
    # Load the pre-trained spaCy model
    nlp = spacy.load("en_core_web_sm")

    /usr/local/lib/python3.8/dist-packages/torch/cuda/__init__.py:497: UserWarning: Can't initialize NVML
      warnings.warn("Can't initialize NVML")

[ ] def get_word_embeddings(text):
        doc = nlp(text)
        embeddings = []
        for token in doc:
            # if not token.is_stop:
            embeddings.append(token.vector)
        return embeddings

[ ] get_word_embeddings(t_processed_data[0][2])

    [array([-0.5494554 , -0.56055266,  1.8407754 , -1.0994548 , -0.52720046,
            -0.84043884, -0.26371843, -0.12708335,  0.40108663,  1.6293848 ,
            -0.28367448, -0.11369698, -1.2266849 , -0.29445678,  0.20818062,
```

6. PLM -

Some famous pre trained models used for abstractive summarization are-

1. Transformer based models -BERT ,GPT2 , GPT3, PEGASUS, T5.
2. LSTM Based Models - Seq2Seq
3. Attention Based Models - TextRank, BERTSUM

As of now we are preferring GPT2 (Generative Pre Trained Transformer) as it has been trained on very large corpus of data and can generate natural language text. Also GPT2 is not task specific like the BERT model which is task specific.
As tweets are often written in more human friendly language instead of following proper grammatical rules hence GPT2 being capable enough to generate more human-like and coherent text.
We can also try the PEGASUS model for summarization.

7.Fine Tuning :

Fine Tuning is required to modify the weights in pre trained language to suit our context . We will tune the hyper parameters of an already trained models like GPT2 trained on large corpus etc to fit our corpus give better results.

8.Challenges faced :

- Gathering data through scraping results in lots of unnecessary data , tweets of other languages and some spam tweets too.
- For converting tweets written into languages like hindi , spanish , chinese etc in english, we need to understand the basic context of those languages and use of translation libraries is required.
- Whether we need to remove stop words or not is still in doubt as for generating a summary we need stop words too. So we can't completely remove them.
- There might be few tweets with some slang words (words more common and popular in normal speaking and has no such historical English background) . We need some tools to consider them also.

9. Conclusion :

- Understood the process of creating a dataset of tweets.
- Figured out the basic preprocessing steps.
- Used Lexical Processing as taught in class during Mandate 2.
- Explored different python libraries to solve the problem.
- Explored different pre-trained models for training our dataset.

10. References :

- https://aclanthology.org/2020.coling-main.504/
- https://pypi.org/project/twitterscraper/0.2.7/
- https://imerit.net/
- https://github.com/agarwaltanmay/text-summarizer/blob/master/Code/newqry.py
- https://www.turing.com/kb/5-powerful-text-summarization-techniques-in-python

A snippet of our generated corpus :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	conversati	created_a	favorite_c	full_text	hashtags/(	hashtags/1	hashtags/2	hashtags/3	hashtags/4	hashtags/5	hashtags/6	hashtags/7	hashtags/8	hashtags/9	hashtags/1	hashtags/1	hashtags/1	hashtags/1	hashtags/1	hashtags/1	hashtags/1	id	media/0/n	me
2	1.63E+18	2023-02-2	0	@DrEliDavid @elonmusk Can one of you clowns ask why he went to Epstein's Island so many times?																		1.63E+18		
3	1.63E+18	2023-02-2	0	@elonmusk Lie																		1.63E+18		
4	1.63E+18	2023-02-2	0	@wmanastasio @saylor @elonmusk When we are FORCED to walk around with "smart" phones & AI is unleashed on the www. Then we are lost before 2026-2029. The phones & AI must l																		1.63E+18		
5	1.63E+18	2023-02-2	0	@CandeOchoaA @dariocelise @omarcervantesr9 @lopezobrador_ @elonmusk Como les gusta repetir pendejadas																		1.63E+18		
6	1.63E+18	2023-02-2	0	@elonmusk is again the richest man on earth. The media acts like he should be ashamed of his accomplishments. The media sux's dog chit.																		1.63E+18		
7	1.63E+18	2023-02-2	0	@elonmusk @mtracey Maybe you can call in the @avengers to finish off the job. It reminds me of when captain america has to take on the government in winter soldier.																		1.63E+18		
8	1.63E+18	2023-02-2	0	@WR4NYGov @elonmusk I love how elons old version tech is still world class and he just stops using it because he's got better stuff now 😂😂😂 countries would kill to get his leftover tech																		1.63E+18		
9	1.63E+18	2023-02-2	0	@qtf @elonmusk It's a danger to our National Security and that was these guys job and now they will try anything to evade responsibility. It's been almost 8 years of an attempt to disenfranchise the																		1.63E+18		
10	1.63E+18	2023-02-2	0	eu perdoaria todos os crimes do elon musk se ele proibisse as palavras monogamia e não-monogamia nessa rede social, não aguento mais ler baboseira																		1.63E+18		
11	1.63E+18	2023-02-2	0	@TerryJulianShow @billlawr @ProdigalSonIV @ProgrammerDude @elonmusk @McFaul It also owes a lot more to the natives of USA still live on reservations, and who have been forgotten																		1.63E+18		
12	1.63E+18	2023-02-2	0	@elonmusk @laragracenewton																		1.63E+18		
13	1.63E+18	2023-02-2	0	@elonmusk @mtracey Why would they? The likes of Raytheon, Northrop, Honeywell, etc are making money hand over fist. Anyone who has an aerospace engineering degree who doesn't work for Y																		1.63E+18		
14	1.63E+18	2023-02-2	0	@theprism89 @elonmusk I enjoy Twitter much more today than when it was a liberal propaganda foghorn.																		1.63E+18		
15	1.63E+18	2023-02-2	0	Elon																		1.63E+18	https://pb	ph
16	1.63E+18	2023-02-2	0	@goddeketal @CommunityNotes @elonmusk Stepping on Lego is also a. Symptom of Covid																		1.63E+18		

Before Preprocessing 1000 rows 251 columns.

A	B
	0
0	can one of you clowns ask why he went to epstein's island so many times?
1	when we are forced to walk around with "smart" phones & ai is unleashed on the www. then we are lost before 2026-2029. the phones & ai must be given limited capabilities and stripped of power over their own being. simply, we must always ha
2	is again the richest man on earth. the media acts like he should be ashamed of his accomplishments. the media sux's dog chit.
3	maybe you can call in the to finish off the job. it reminds me of when captain america has to take on the government in winter soldier.
4	i love how elons old version tech is still world class and he just stops using it because he's got better stuff now countries would kill to get his leftover tech
5	it's a danger to our national security and that was these guys job and now they will try anything to evade responsibility. it's been almost 8 years of an attempt to disenfranchise the american people. it's time the facts can't stay their responsibility.
6	it also owes a lot more to the natives of usa still live on reservations, and who have been forgotten
7	why would they? the likes of raytheon, northrop, honeywell, etc are making money hand over fist. anyone who has an aerospace engineering degree who doesn't work for you is making more money than they ever had during their indoctrination
8	i enjoy twitter much more today than when it was a liberal propaganda foghorn.
9	elon
10	stepping on lego is also a. symptom of covid
11	that's scary news.
12	there is no space stop it elon earth is flat
13	it's a race thing the jews control the media and they want black and white people to not like each other
14	no, your ego made the offer to buy twitter and laws written by and for rich people forced you to actually buy it. nice try though.
15	elon musk says the us media is 'racist against whites and asians' via
16	this feels like the iphone in 2006
17	so the microsoft guy is a climate expert

After Preprocessing 686 rows 1 column.