Projet NLP

Table of content

- Introduction of our project
- Twitter API
- Text classification model
- Improvements
- Conclusion

Introduction

Twitter API model

This function allow us to get tweets using the twitter API.

The parameters are the language, search word, date and the number of tweets you want.

Unfortunately, the twitter API permission has been udptated.

```
{"consumer_key":"fibpny904RMcf2bYRUwM3tInF",
"consumer_secret":"yCGZQKs91AtfAqPq9pLFTui7MRkkDRaUgVVwtxfUHJxsLMssnK",
"bearer_token":"AAAAAAAAAAAAAAAAAAAAA9OtgEAAAAALkbClLDb1H%2BVKQHsAQoA2s
"access_token":"1785316205508575233-yuZd8tmGgFHuGxhE8rmW48LsbyRkw0",
"access_token_secret":"4mrPBprOBgKtzL72qv7qNWWqCzh7jGHltfnk6Q8ajcs3G"}
```

```python
def get_API(credentials):
    auth = tw.OAuthHandler(credentials['consumer_key'], credentials['consumer_secret'])
    auth.set_access_token(credentials['access_token'], credentials['access_token_secret'])

    return tw.API(auth, wait_on_rate_limit=True)

credentials = load_credentials("/content/twiter_API.json")
api = get_API(credentials)

import tweepy as tw
def search_tweets(search_word, date_since, limit=20):
    """
        search_word: the topic we want to search about (e.g. covid19)
        data_since: the date from when we want the information related to "search_word"
        limit: used to restrict the number of tweets to be returned. If not specified, it will retrive
    @return:
        tweets_cursor: a cursor used to paginate through the large retrieved data
    """
    tweets_cursor = tw.Cursor(api.search_tweets,
                q=search_words,
                lang="fr",
                since=date_since).items(limit)

    return tweets_cursor

search_words = "#onepiece"
date_since = "2024-04-29"

ONE_cursor = search_tweets(search_words, date_since)
```

Twitter Web scrapping

Twitter web scrapping is a possible free solution.

But it is’nt a stable solution and Elon Musk said he’s reducing the power of twitter scrapping tools.

```
!pip install git+https://github.com/JustAnotherArchivist/snscrape.git

import snscrape.modules.twitter as sns
import itertools

search = "one piece"

scraped_tweets = sns.TwitterSearchScraper(search).get_items()

sliced_scraped_tweets=itertools.islice(scraped_tweets, 10)

df= pd.DataFrame(sliced_scraped_tweets)[['date','content']]
```

3.bis

Creation of a Text classification model

=>Let’s take a look at our model in the notebook

Improvements

- Fine-tune a pre-trained model instead of creating a model from scratch
- Use of stemming/ Lemmatization on the data
- Using parallel computing to reduce the training time => Increase the epoch and batch size

Conclusion

Thank you for your **Attention**