cryptosentiment

kacper klensporf

Założenie

Scrapowanie danych w celu ewaluacji ogólnego sentymentu odnośnie kryptowalut.

Tweet scraping

Podstawowym źródłem miałyby być tweety. Jako, że spora część całego kontentu w tej dziedzinie pochodzi z bot-kont, bądź jest po prostu spamem, użyłem listę użytkowników, których tweety chciałbym scrapować.

Tweet scraping v2

Do samego użyłem opensource’owego projektu dostępnego na githubie. Problemem jednak były jego limitacje. Na każde konto przypadały jednak pewne ograniczenia:

- the request limit is updated every 15 minutes for each endpoint individually
- each account have 50 search requests / 15 min

Z tego względu niezbędne było utworzenie kilku kont do scrapowania.

https://github.com/vladkens/twscrape

Przetwarzanie tweetów

Celem było doprowadzenie tweetów do takiego stanu, w jakim były użyte do fine-tuningu BERT'a:

- usunięcie URL'i, hashtagów, cashtagów
- przekonwertowanie do małych liter
- weryfikacja długości
- sprawdzenie czy tweet nie jest retweetem

Dodatkowo, zaimplementowałem rozwiązanie osobnej tokenizacji tweetów, a następnie wgrywanie do modelu już gotowych inputów. Metoda ta wbrew oczekiwaniom okazała się wolniejsza niż użycie prostego pipeline.

```
n – liczba postów
p – czas wykonania w pipe [s]
k – czas wykonania samych klasyfikacji [s]

 n      p       k
225 | 57.68 | 52.30
450 | 100.81| 105.55
750 | 164.04| ***
```

Analiza sentymentu tweetów

W celu analizy sentymentu poszczególnych tweetów, wykorzystałem jeden z modeli dostępny na huggingface.co - CryptoBERT. Jest to model językowy oparty na BERT (Bidirectional Encoder Representations from Transformers.

Jak pisze autor:

> The model has been tested, out-of-sample, on a set of around 200K stocktwits posts, and delivered superior classification performance when compared to Vader or other BERT-based classifiers. The accuracy and F1-score for 3-class problem (bearish, neutral and bullish) was around 70% if I remember correctly. This greatly outperformed VADER's predictions, which delivered accuracy closer to 50%.

Analiza sentymentu tweetów

Za pomocą modelu oceniamy, oraz zapisujemy do bazy danych sentyment wszystkich zescrapowanych tweetów, które zostały utworzone danego dnia.

Kolejnym krokiem jest wyliczenie ogólnego dziennego sentymentu. W tym przypadku założyłem, że będzie on reprezentowany w skali 0/1 gdzie:

$$s = p / (p + n)$$

s - score, p - tweety oznaczone jako pozytywne, n - tweety oznaczone jako negatywne

Prezentacja wyników

Do prezentacji wyników, napisałem prosty front z użyciem vue.js