

Descripción de la solución

Con el objetivo de brindar una base sólida de persistencia y de utilizar un framework compatible POO, se decidió realizar una API REST con Django, utilizando una base de datos MySQL.

Con el objetivo de poder obtener noticias relevantes de distintas fuentes de información se utilizó NewsAPI (NewsAPI Inc, 2021) para obtener los enlaces a las noticias y la biblioteca NewsPaper (CodeLucas, 2014) para realizar el web scrapping de forma sencilla. Para evitar tener inconvenientes con la protección Cloudflare de los distintos sitios de noticias, se realizaron varios cambios sobre la biblioteca previamente mencionada.

Una vez descargado el contenido de los distintos artículos web relativos a la criptomoneda a analizar, se realiza un pre-procesamiento con expresiones regulares. A continuación, se procede a Tokenizar cada artículo, identificar las palabras más frecuentes y encontrar las NERs (Name Entity Recognition) utilizando la biblioteca NLTK (NLTK, 2022). Por último, cada artículo es analizado con el modelo vaderSentiment (CJHUTTO, 2014), en donde se clasifica el mismo como: POSITIVO, NEUTRAL o NEGATIVO. Toda esta información se persiste.

Con las distintas “opiniones” otorgadas por los artículos se computa un *promedio* resultante para la criptomoneda en ese momento específico. El valor de dicho promedio se encuentra entre -1 (muy negativo) y 1 (muy positivo). Este resultado se persiste en conjunto con el precio actual de la criptomoneda que se extrae de la API de Coingecko (MAN-C, 2021).

Una vez finalizados los análisis de sentimientos, la información almacenada permite computar los siguientes datos:

- Polaridad de las distintas fuentes de información: En base a las emociones detectadas en los distintos artículos de una misma fuente, se determina de forma tentativa si la misma es: CRYPTO FRIENDLY, NEUTRAL, CRYPTO ENEMY.
- Posible dirección del precio de una criptomoneda: En base al análisis actual se puede aproximar la dirección del precio de la moneda.

Justificaciones

Cantidad de criptomonedas: Se decidió soportar una muy amplia variedad de monedas y tokens. Utilizando en concreto la totalidad de monedas disponibles en Coingecko.

Web Scraping: Debido a que se decidió diseñar un sistema compatible con muchas monedas, era importante utilizar una herramienta universal para poder realizar el scraping, ya que era imperativo poder utilizar una cantidad variable de fuentes. Para esto se decidió utilizar un motor de búsqueda como NewsAPI y la biblioteca NewsPaper que realiza el scraping propiamente dicho.

Análisis de sentimientos: La decisión de que modelo de PNL utilizar fue realmente compleja. Como punto de partida tome la decisión de no ligar la solución al uso de APIs externas con sus conocidos limitantes, por lo que me incline por el uso de modelos pre-trained. El artículo “Prediction of stock values changes using sentiment analysis of stock news headlines” [1] fue el que me ayudo a decidirme por “Vader Sentiment”.

Referencias

[1] Nemes, L., & Kiss, A. (2021). Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3), 375–394.
doi:10.1080/24751839.2021.187425.