**Big Data Programming Project Report**

Ahmed Elhadidy   
CU202101774

The Knowledge Hub Universities

CS L5-S02 2021/2022

Dr. Batoul Haidar

14/05/2022

# Introduction

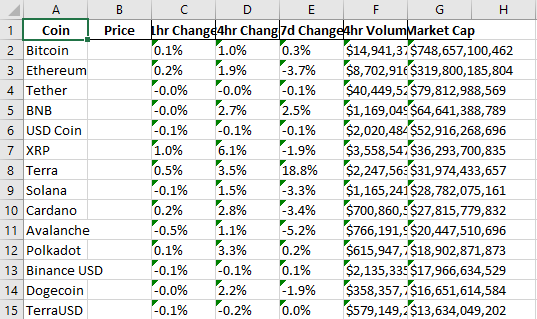
Machine learning is the backbone of data science. Data science is a study of massive amounts of data, the art in this subject is that the data is extracted from raw, structured, and unstructured data that is processed using complex algorithms, different scientific methods, and technologies. After filtering and studying this type of data, it ends up with delivering a data output that tells a story or can help uncover insights, patterns and discover new trends. Data science is fulfilled through a cycle of steps, first is data acquisition; extracting data can be done from databases, online repositories, logs, and even websites that are visible to the public. As a matter of fact, everything around us is made up of data but it needs to be explored thoroughly in order to reach a meaning. Moving next to data preparing which involved gathering the data and cleaning it to transform it into meaningful data, then exploratory data analysis is performed to study the data thoroughly and provide models and visualizations based on the studies. The purpose of this paper is to demonstrate what was done throughout this module’s project. The module’s project involved Big Data, where multiple datasets were extracted to perform studies over specific cryptocurrency coins. Throughout this project, one will apply the different theories and steps of data science and perform a complete data extraction whilst performing analysis and extracting useful information. This report discusses what happened throughout this project within detail.

# Phase 1

During the first phase, it was required to select a topic that the project would revolve around. The topic to be used is relevant to the modern events and would have a value hence why cryptocurrency was chosen as the topic for this project. Since the topic is mainstream, it would be easy to gather data for this particular topic. The intended outcome was to understand how fluctuating the crypto market is and provide drawings and statistics based on a specific course of time.  
The first approach was searching Google for data related to cryptocurrency and cryptocurrency coins; the most suitable result was that provided by a website called CoinGecko. The website had an HTML table that was sorted in a descending order by the market cap. For most of the current cryptocurrency coins, the table had details that included the price, the price change over the course of 1 hour, 24 hours, and 7 days. Not only did this website include crucial data relevant to the goal of the project, but it was easy to scrap data from. Therefore, the first scrapping was approached using BeautifulSoup, which is a Python library used for web scraping thereby pulling data from HTML and XML files and outputting the data in a readable manner.

  
*Figure 1.1 – Sample Code*

In this code used for the first scrapped dataset, a for loop was used to loop over the pages of the website; then the HTML of a page was parsed, and the table was mainly selected since it contained the relevant data. The data of the table was then selected by identifying the class of each row and appending the data in the data frame. Each class was assigned its name in the dataframe and the result was outputted in an excel file.

  
*Figure 1.2 – Sample Data*

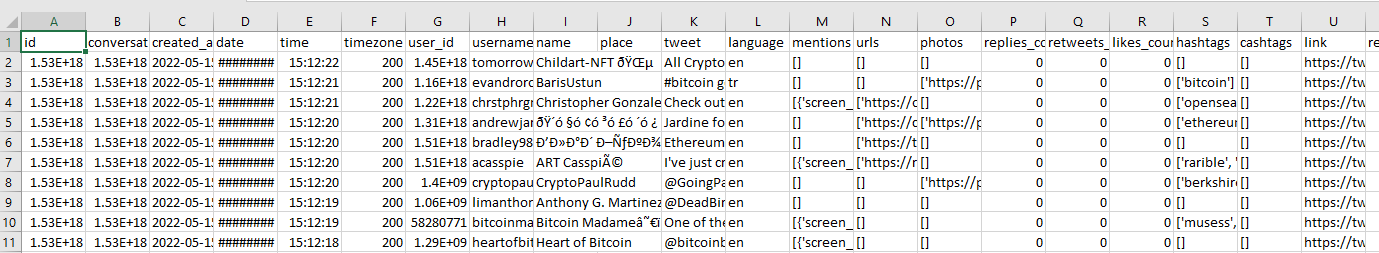
The output of this phase provided an outline the project and it provided data relevant to the thesis. By the end of the phase, data gathering for this topic helped gain additional information and ideas that will be beneficial in the rest of the phases.

# Phase 2

The second phase of the project had some crucial changes that would affect, in a good manner, the outcome of the project. Cryptocurrency is trending and there is no place to keep up with trends like twitter. Cryptocurrencies are tradable assets; the supply and demand factors are what makes cryptocurrencies valuable. Generally, social media can cause cryptocurrency price fluctuations this was a motivation to use an alternative data scraping tool that is also based on Python: twint library. Twint, short for twitter intelligence tool, is an advanced twitter scrapping tool that provides the ability to scrap tweets from twitter profiles. Twint will be used to scrap 60,000 tweets regarding five of the most known, and valuable cryptocurrency coins. This library makes the implementation is relatively easy, as seen in figure 2.1, everything is adjustable and direct.

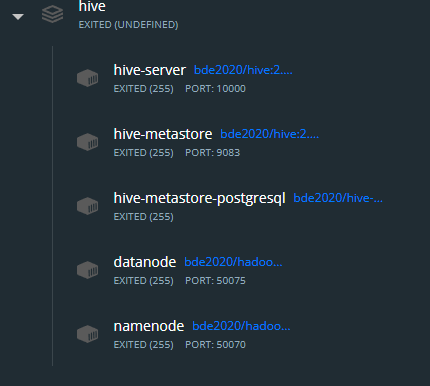
  
*Figure 2.1 – Twint code*

The figure above consists of the configuration of the code that outputs the dataset. As seen in c.Search, the scraping will target five coins: Bitcoin, Ethereum, Cardano, Solana, Ripple since these are some mainstream coins. The output will be structured in a data frame. This is made possible using the pandas library which is a library crucial for data analysis and it contains data structures and operations for data manipulation. The data will then be stored in a CSV file.  
This approach is valuable because Twitter, and social media in general, is a big factor that affects cryptocurrency. After extracting this dataset, the goal was to study and provide the sentiment analysis of the cryptocurrency data over the course of April and May.

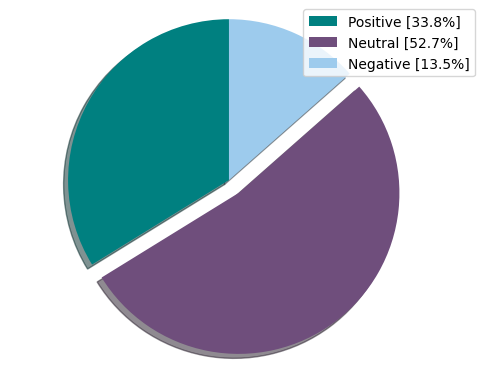
  
*Figure 2.2 – Output sample*

# Phase 3

The third phase was designated to study the data and implement data visualizations that would provide logical representation for the tweets gathered. Another achievement in this phase was use Docker server so that we host a Hive container that can hold the dataset. Docker is an open-source technology to package an application and all its dependencies into a container. Apache Hive, configurated using Hadoop is an open-source framework, fault-tolerant data warehouse system that provides the option to efficiently store large datasets. This step was done in several steps. First was to create a directory structure with that consists of the code and data on the local machine, configured for Hive. Then a docker tool was composed to define and run simultaneous docker applications; meanwhile, the Hadoop configuration file is used to set the variables for a working environment. Finally, in order to translate the dataset into a Hive database, a HQL script was used that is imported and executed by the hive server.

  
Figure 3.1 – Hive Container

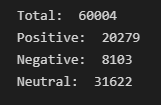
The second step in this phase was to implement exploratory data analysis (EDA) after filtering and deciding which data is relevant after the Hadoop phase. EDA is a method of visualizing data that summarizes the purpose of the main data. The output of this analytical data is resembled using data visualization methods like statistical graphs. The graphs outputted from this dataset explore the data in a systematic way, providing patterns that show how cryptocurrency is actively discussed by Twitter users and how the interaction varies from one day to another. It was revealed that in the recent period, most of the tweets were mainly neutral, unlike a while ago. Tweets were positive; however, since the cryptocurrency market is crashing right now, people are starting to react accordingly and this reflects in their actions on twitter which had affect on the dataset.

  
Figure 3.2 – Pie chart

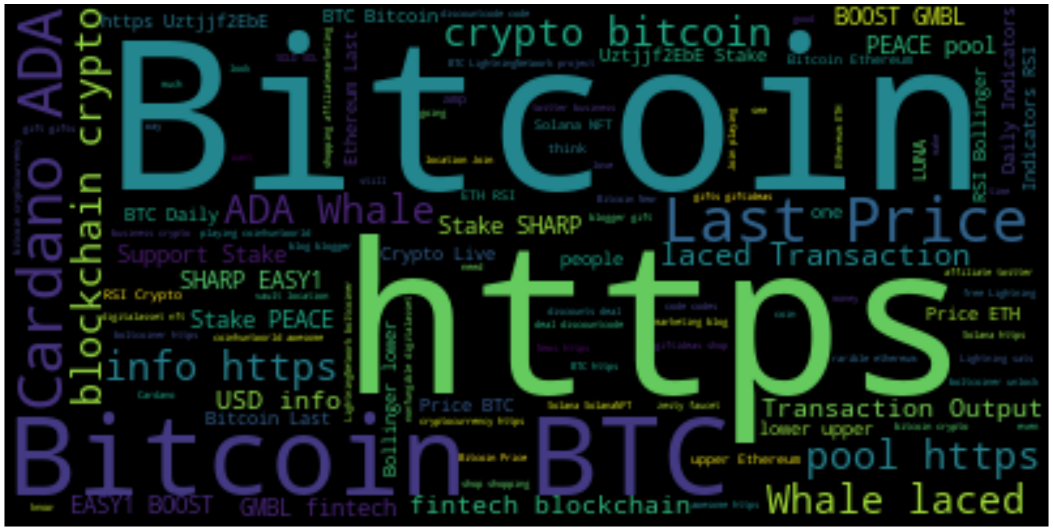
The exploratory data analysis was performed using various libraries like plotly, seaborn, and tooltext. These are visualization libraries and they eased the process of creating visualizations to represent the data included in the datasets.

# Phase 4

In the final phase of the project, the project was finalized, in this phase everything started to come together. In order to understand the output of the project, sentiment analysis was done using Python over the dataset. The result of the sentiment analysis revealed as seen in the figure 4.1, that during the recent period of time the tweets were mostly neutral and the least amount was negative.

  
*Figure 4.1 – Sentiment Analysis*

Adding to that, a word cloud was created that would contain the most repeated words in the tweets available in the dataset.



*Figure 4.2 – Word bank plot*

As a result, a dictionary of word frequencies was initiated, and it sorted the words from starting from the frequently used, the result showed that the most used word in most of the cryptocurrency tweets is Bitcoin. Bitcoin indeed is the most demanded coin, hence why the communication regarding this coin is the most frequent and the most popular. On the other hand, the sorting also outputted the least frequent words but this mainly included words that were barely used and they barely have a logical meaning.



Figure 4.3 – *Word frequency*

# Conclusion

In data science, we use various machine learning algorithms to solve problems. This project helped explore new sectors of machine learning and it provided direct interactions with large, complex data structures and it taught how to extract data that has value. This project was mainly done using Python including libraries like BeautifulSoup, Twint, and plotly. This project was uploaded on [GitHub](https://github.com/Hadidy7/CryptocurrencyDataHadoop) with thorough explanation on how to run the environment.

# ‌References:

Vidya. (2020, July 13). *Tweet Analysis in just < 100 lines of code using Twint, TextHero, SweetViz Python Libraries!* Medium; Analytics Vidhya. <https://medium.com/analytics-vidhya/tweet-analysis-in-just-100-lines-of-code-using-twint-texthero-sweetviz-python-libraries-2c5813cf4d74>

*Cryptocurrency Prices, Charts, and Crypto Market Cap | CoinGecko*. (2022). CoinGecko; CoinGecko. <https://www.coingecko.com/>

‌Data Science | Dr. Batoul Haidar (2022). The Knowledge Hub Universities;

Hrishi Shirodkar. (2021, March 9). *Apache Hive on Docker - Hrishi Shirodkar - Medium*. Medium; Medium. <https://hshirodkar.medium.com/apache-hive-on-docker-4d7280ac6f8e>