Cryptocurrencies scraper

May 2022

1 Participants

- Zofia Cieślińska, 395117
- Adrian Hirt, K-14561
- Kai Xing, 389862

2 Topic and webpage

We decided to scrape for 100 most popular cryptocurrencies from the webpage **coinmarketcap**. It contains over 10000 various cryptocurrencies and their market statistics. We scraped basic statistics for each of the 100 most popular currencies:

- name of the currency,
- total market value of circulating supply (aka market cap),
- price in dollars (value),
- total circulating volume,
- change in the price at the time of scraping.

3 Scraper mechanics

The scrapers first gather one hundred links to cryptocurrency pages from the main page (with the possibility to scrape for more - but the default setting is to scrape only first page with 100 currencies). Then from each page they scraped values listed above. Due to anti-scraping measures taken by the website, we needed to slow down our scrapers to avoid being blocked. Before slow-down Scrapy was running in a few seconds (until it got blocked after a few too close runs). The waiting time is set to 3 seconds both is Scrapy and Selenium. Because of it, the fastest scraper is Beautiful Soup, with 17 seconds runtime, both Scrapy and Selenium running in minutes.

3.1 Beautiful Soup

For each link found using findall method and regex a soup is created. Listed above values are found and save to a CVS file.

3.2 Selenium

The links and values were found using XPaths, and an additional waiting time was added after accessing new URL.

3.3 Scrapy

Two spiders were created, one for getting links and one for retrieving the values. All was found using XPaths. Delay was added in the settings, as well as a custom user agent.

4 Output

The final output is a CSV file with columns corresponding to the values listed in the introductory part of the report.

5 Data analysis

The data was cleared and converted to numeric format where possible (in "data analysis" folder there are two files - one with raw scraped data and one with data cleared using a python script). On the cleared data a few simple analyses were performed - means and standard deviations of the values as well as simple visualisations comparing all of the currencies.

5.1 Some statistics

Mean values of the scraped data, listed for the data scraped by scrapy spider:

- change: 1.715206e+01
- market_cap: 1.277146e+10
- value: 8.142975e+02
- volume: 1.498343e+13

Standard deviations of the above means:

- change: 1.335060e+01
- market_cap: 6.341691e+10
- value: 4.466345e+03
- volume: 1.074468e+14

5.2 Some plots

Scatterplots comparing different values across currencies can be found **on the github page**. They present:

- relation of change to the total volume divided by market cap times 100, which point sizes varying according to the value of the currency,
- relation of market cap with respect to the value,
- comparison of the total volume of all the currencies, with point sizes representing total value and colors - the change in value.

6 Division of the work

After spending most of the given time on unsuccessful attempts to make Scrapy Splash work on airbnb site (visible in the history of the github repository), we decided in the last week to scrape a different site - but due to an accident which happend this week, Kai was not able to participate fully in the final project.

- Zofia Cieślińska:
  - Scrapy part of the project, data analysis, final report
- Adrian Hirt:
  - Beautiful Soup and Selenium parts of the project, management of github repository
- Kai Xing:
  - Selenium part of the first version of the project