**Reddit and stock prices database**

**Stocks**

In the database you can find the prices of the following stocks:

* + AAPL (Apple Inc., benchmark)
  + AMC (AMC Entertainment Holdings Inc.)
  + BBBY (Bed Bath & Beyond Inc.)
  + FIZZ (National Beverage Corp.)
  + GME (GameStop Corp.)
  + LGND (Ligand Pharmaceuticals Inc.)
  + NOKIA (Nokia Corp.)
  + VIAC (Viacom CBS Inc.)

All these stocks, except AAPL, have been affected by some Reddit conversations in the first months of 2021. All these stocks are primarily traded in New York (New York Stock Exchange or Nasdaq), except Nokia whose main market is the Helsinki Stock Exchange (Nordic Nasdaq).

**Stock prices**

Stocks are traded continuously between Opening and Closing. In New York, trading starts at 09:30 and ends at 16:00 (American Eastern Time), from Monday to Friday. The dataset contains prices between September 2020 and July 2021. In the table, you can find the exceptions to this rule, represented by national holidays, for the considered period:

|  |  |  |
| --- | --- | --- |
| **NYSE Holiday** | **Date** | **Trading hours (NY time)** |
| Labor Day | Monday, September 7 | Closed |
| Thanksgiving Day | Thursday, November 26 | Closed |
| Thanksgiving Day | Friday, November 27 | Half-day (09:30 – 13:00) |
| Christmas Day | Thursday, December 24 | Half-day (09:30 – 13:00) |
| Christmas Day | Friday, December 25 | Closed |
| New Year’s Day | Friday, January 1 | Closed |
| Martin Luther King, Jr. Day | Monday, January 18 | Closed |
| Washington's Birthday | Monday, February 15 | Closed |
| Good Friday | Friday, April 2 | Closed |
| Memorial Day | Monday, May 31 | Closed |
| Independence Day | Monday, July 5 | Closed |

Similarly, for the Nordic Nasdaq (Helsinki) the trading hours are between 10:00 and 18:30 Finnish time, from Monday to Friday except the following holidays:

|  |  |  |
| --- | --- | --- |
| **Nordic Nasdaq Holiday** | **Date** | **Trading hours (NY time)** |
| Christmas Day | Thursday, December 24 | Closed |
| Christmas Day | Friday, December 25 | Closed |
| New Year’s Eve | Thursday, December 31 | Closed |
| New Year’s Day | Friday, January 1 | Closed |
| Epiphany | Wednesday, January 6 | Closed |
| Good Friday | Friday, April 2 | Closed |
| Easter Monday | Monday, April 5 | Closed |
| Ascension Day | Thursday, May 13 | Closed |
| Midsummer | Friday, June 25 | Closed |

The Opening and the Closing prices are determined through auctions, the intermediate prices (between 09:30 and 16:00 in New York) are determined through continuous trading.

In detail, the datasets are:

* **daily\_close\_prices.csv**: price at the end of the trading hours for the considered stocks, as defined by the Closing auction
* **daily\_open\_prices.csv**: price at the beginning of the trading hours for the considered stocks, as determined by the Opening auction
* **hourly\_prices.csv**: traded price at each hour for the considered stocks; all the times are in Eastern American times (New York). Only prices included in the aforementioned trading hours should be considered

To link stock prices to Reddit posts, one must be aware of the time zone differences and perform a transformation if needed.

**Reddit posts and comments**

The Reddit database contains all the posts published on the *wallstreetbets* subreddit between 1st September 2020 and 6th June 2021, containing the name (tag) of any of the aforementioned stocks in their title. For each stock, all the submissions are contained in one single json file; in the table you can find the list of features associated to any post. Among them, you find the time and day on which the post was published. All those times are expressed in GMT; hence you must be careful when comparing them with the (hourly) stock prices. For example, you may want to transform everything into American time, or CET (Amsterdam) time.

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Type** |
| all\_awardings | List of awards and coins associated to the post | List |
| author | Author name | String |
| author\_fullname | Author id | String |
| created\_utc | Submission timestamp (UTC time) | Integer |
| created | Submission timestamp (UTC time) | Integer |
| full\_link | Submission webpage | String |
| gildings | Gilding (special awards) | Dictionary |
| id | Submission id | String |
| link\_flair\_css\_class | Submission class (e.g. question, news, meme, etc) | String |
| link\_flair\_text | Submission type (e.g. Discussion, Meme, etc) | String |
| num\_comments | Number of comments until the day of download | Integer |
| num\_crossposts | Is the submission a crosspost from another subreddit? | Integer |
| over\_18 | Age restriction tag | Boolean |
| permalink | Permanent submission link (part of url) | String |
| pinned | Is the submission pinned in the subreddit? | Boolean |
| removed\_by\_category | Subject who removed the submission, if any (e.g. moderator) | String |
| retrieved\_on | Submission download timestamp (UTC time) | Integer |
| score | Submission score (difference between upvotes and downvotes) | Integer |
| selftext | Submission text | String |
| subreddit | Subreddit id (always “wallstreetbets”) | String |
| subreddit\_id | Subreddit id (always “t5\_2th52”) | String |
| subreddit\_subscribers | Number of subreddit users at the moment of submission creation | Integer |
| title | Submission title | String |
| total\_awards\_received | Total awards received by the submission | Integer |
| upvote\_ratio | Ratio between upvotes and downvotes | Float |
| url | Submission webpage | String |

For each stock, you find a zipped folder containing several json files, each showing the comments associated to a certain submission. You can check the identifier of the corresponding submission by looking at the name of the json file. In the table, you find the list of features contained in any json file.

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Type** |
| all\_awardings | List of awards and coins associated to the comment | List |
| author | Author name | String |
| author\_fullname | Author id | String |
| created\_utc\_time | Submission timestamp (UTC time) | Integer |
| created | Submission timestamp (UTC time) | Time |
| associated\_award | Associated award | String |
| gildings | Gilding (special awards) | Dict |
| id | Comment id | String |
| link\_id | Associated submission id | String |
| awarders | List of awarders | List |
| top\_awarded\_type | Type of top awarded tags | String |
| total\_awards\_received | Total awards received by the comment | Integer |
| treatment\_tags | List of tags associated to the comment | List |
| permalink | Permanent submission link (part of url) | String |
| comment\_type | Type of comment, if defined | String |
| retrieved\_on | Comment download timestamp (UTC time) | Integer |
| score | Comment score (difference between upvotes and downvotes) | Integer |
| body | Comment text | String |
| subreddit | Subreddit id (always “wallstreetbets”) | String |
| subreddit\_id | Subreddit id (always “t5\_2th52”) | String |
| locked | Can the comment receive any further replies? | Boolean |
| no\_follow | Has the comment the ‘no follow’ flag? | Boolean |
| parent\_id | Previous level comment/submission id | String |

In detail, the datasets are:

* **submissions\_wallstreetbets\_XXXX\_start20200901\_end20210706.json**: reddit submissions in the subreddit wallstreetbets between 1st September 2020 and 6th June 2021 containing the stock XXXX in the title
* **comments\_XXXX.tar.gz**: zipped folder containing the list of comments associated to each of the submission containing the stock XXXX in the title