EINDHOVEN UNIVERSITY OF TECHNOLOGY

Correlations between stocks using Spark

2IMD15 Data Engineering - Milestone 1

Group 1: Daniel Teixeira Militao (1486314), Manon Wientjes (1398903), Pawel Budzynski (1511734), Tong Zhao (1416790), Ugne Laima Ciziute (1495186)

Dataset and correlation functions

The dataset we used is stock data from January until mid April 2020[1]. We want to find interesting correlations between different stocks. The variables used to calculate correlations are: opening, highest, and lowest prices of each stock on an hourly basis for working hours and days. The correlation functions used were Pearson correlation and Mutual information.

Pre-processing

In order to be able to compute correlations, the vectors need to be of the same length. Hence, we needed to filter out some stocks and handle missing values by interpolating and extrapolating. Pre-calculations are done to speed up computation of correlations. The pipeline in Figure 1 depicts the steps taken during pre-processing.

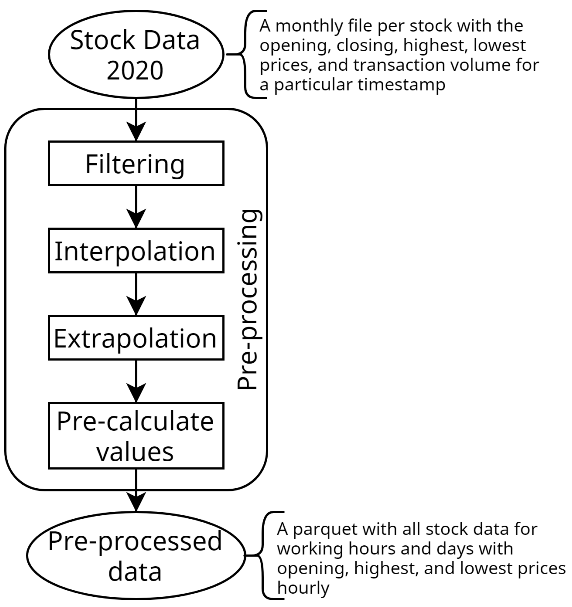

Figure 1: Pre-processing pipeline.

Correlations calculation

Correlations are calculated according to Figure 2. First groups of stocks are created to ensure that each worker gets a similar workload. To prevent redundant comparisons between group combinations, duplicate combinations are filtered out. In the end, $flatMap$ is used to calculate the correlations between every combination. Since there are multiple stocks in every group, we need to make sure no redundant computations are done. We do this using a nested for loop and testing some logical predicates.

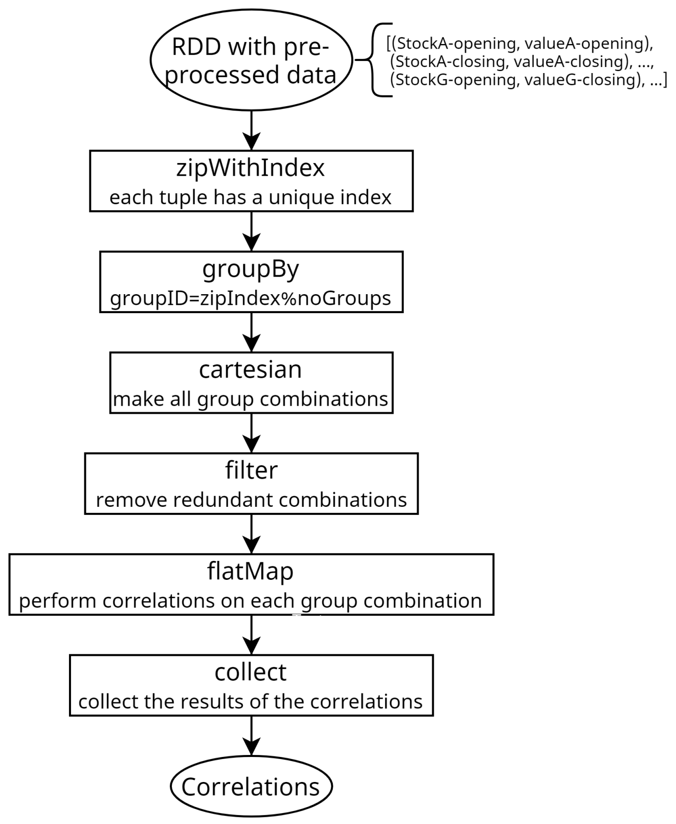

Figure 2: Correlations calculation pipeline.

[1] https://canvas.tue.nl/courses/10287/files/2383551/download?download_frd=1
Video presentation: https://drive.google.com/open?id=1XFU6AW4Rfsfl6JE2Ay6Jf22pqqccpXNy

Experimental performance

The final dataset has 1,998 vectors each having 639 dimensions. Running the program on one machine with 4 cores and 16 GB RAM takes about 2 minutes for both Pearson correlation and Mutual Information.

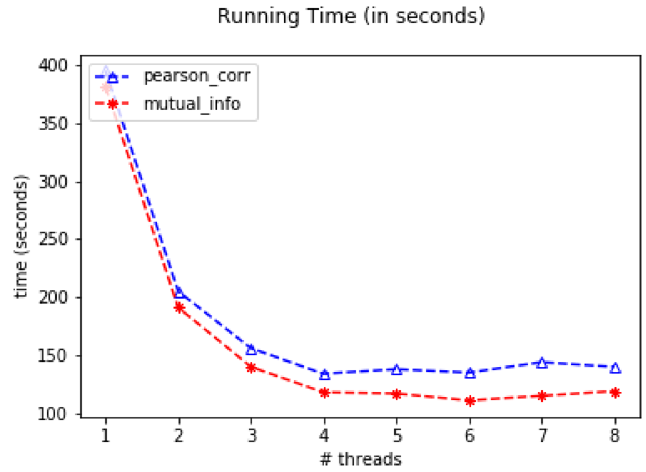

Figure 3: Running time per thread.

Figure 3 shows the relation between the running time and the number of threads. Increasing the number of threads is beneficial up to some point. At this point probably too much RAM is needed and hence the running time will not decrease anymore.

Insights

As we had anticipated, regardless of the correlation function, the top correlations were for the most part, between the same value type, e.g. opening with opening.

The results for Pearson correlation can be seen in Table 1. In the future, we could perhaps filter correlation pairs with values above 0.95 in order to derive more interesting insights.

Table 1: Ten highest Pearson correlations

Pair		Value
Amsterdam_MT-lowest	Madrid_MTS-lowest	0.9999
Amsterdam_MT-highest	Madrid_MTS-highest	0.9999
Amsterdam_MT-opening	Madrid_MTS-opening	0.9999
CME_6A-lowest	Forex_AUD-lowest	0.9999
CME_6A-opening	Forex_AUD-opening	0.9999
CME_6A-highest	Forex_AUD-highest	0.9999
Forex_GBP-lowest	CME_6B-lowest	0.9998
CME_6B-opening	Forex_GBP-opening	0.9998
Forex_GBP-highest	CME_6B-highest	0.9997
Forex_CHFEUR-lowest	Forex_EURCHF-highest	-0.9997

The scores for mutual information (MI) stay relatively low, as seen in Figure 2, raising the question whether the applied mapping was the best choice or if the MI score is meant to be low for all stock pairings. It might be a good choice to further investigate the topic and experiment with different ways of discretization of stock prices.

Table 2: Ten highest Mutual Information correlations

Pair		Value
CME-eMini_NQ-opening	CBOT-mini_YM-opening	0.4579
London_TUI-highest	Xetra_TUAG00-highest	0.4488
London_TUI-opening	Xetra_TUAG00-opening	0.4349
London_TUI-lowest	Xetra_TUAG00-lowest	0.4279
CME-eMini_NQ-lowest	CBOT-mini_YM-lowest	0.4208
CME-eMini_NQ-highest	CBOT-mini_YM-highest	0.3845
vwd-Indications_XPDUSD-lowest	NYMEX_PA-lowest	0.3827
vwd-Indications_XPDUSD-highest	NYMEX_PA-highest	0.3489
US-Indices_DJGT-highest	US-Indices_W1DOW-highest	0.3199
vwd-Indications_XPDUSD-opening	NYMEX_PA-opening	0.3198

Department of Mathematics and Computer Science