Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

DCC057 Mineração de Dados

Trabalho Prático 4 — Relatório de Carregamento e Descrição dos Dados

Alexander Thomas Mol Holmquist

30 de março de 2021

1 Objetivo

Este relatório trará informações descobertas durante a primeira fase do projeto, com relação ao carregamento e descrição dos dados. Serão expostos aspectos muito importantes para todo o projeto, como onde encontrar os dados, como carregá-los nas ferramentas, e visualizações.

2 Localização e Codificação dos Dados

Os dados devem ser dispostos no diretório "data". Não é viável carregá-los todos no repositório remoto do projeto na plataforma GitHub, devido ao tamanho. Por isso, foi incluído na mesma tão somente uma referência de onde encontrá-los online.

É muito provável que este projeto se restrinja ao arquivo "prices.csv", que contém 7 colunas, e 851264 registros. Cada coluna foi descrita na proposta do projeto, mas aqui é repetido o seu significado.

- **date** data do registro
- **symbol** símbolo da empresa
- **open** preço inicial — momento de abertura do mercado
- **close** preço final — momento de fechamento do mercado
- **low** preço mínimo
- **high** preço máximo
- **volume** quantidade de ações negociadas

É importante notar que o arquivo "prices.csv" dá uma tabela indexada por tempo. A frequência dos registros é diária. O primeiro dia é 04 de janeiro de 2010, e o último dia é 30 de dezembro de 2016. Como explicado na Seção 3, foi necessário transformar a coluna com os dias para conter valores inteiros, a partir de 0.

Nem todos os dias tem um registro correspondente. Somente são registrados os dias úteis, e a quantidade de registros não é uniforme para cada dia da semana. Isto é, alguns dias da semana tem mais registros que outros. Apesar disso, a disparidade provavelmente não é causada pela diferença entre os dias da semana.

Todos os arquivos do banco de dados estão no formato CSV. Os campos estão separados por vírgula, e registros distintos são separados pelo caractere de nova linha. A primeira linha dos arquivos define os nomes das colunas.

3 Manipulações

A única manipulação que se mostrou necessária para esta fase, para que se pudesse visualizar os dados, foi a translação da coluna "date" para o intervalo [0, 2552]. Obviamente, a frequência da coluna transladada continua diária (com exceção dos finais de semana).

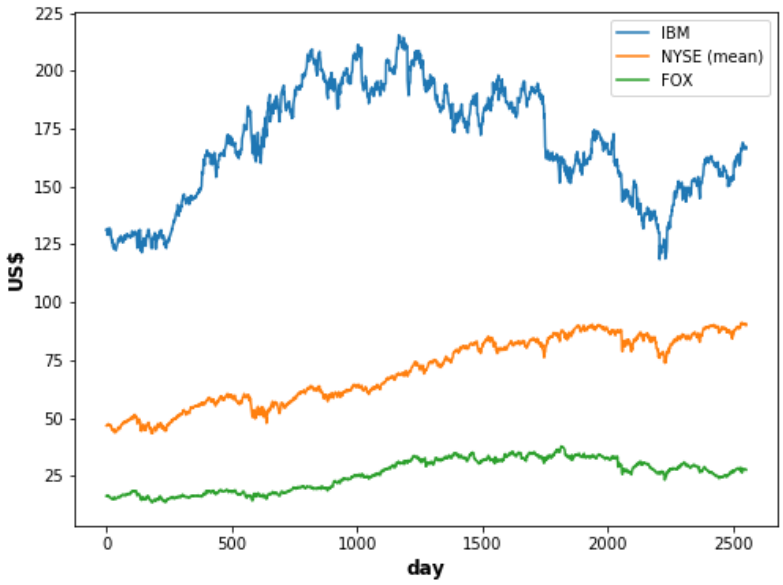

Figura 1: Preço das ações no momento de abertura de mercado, para as empresas conhecidas IBM e Fox. Para a bolsa de Nova York como um todo, a média entre todas as empresas é considerada.

4 Gráficos

Algumas visualizações foram produzidas pelo passo 4 desta fase. Decidiu-se utilizar as empresas IBM e Fox como exemplo, para ilustrar as diferenças entre empresas. A Figura 1 mostra que as ações da IBM, desde o primeiro dia, são muito mais caras que as da Fox. Além disso, é observado que as ações da Fox parecem altamente correlacionadas com a média geral (em laranja) da bolsa de Nova York. O comportamento para a IBM é, por outro lado, quase independente de tal preço médio; talvez a razão principal é que ela é uma empresa do ramo tecnológico, e portanto os investimentos tendem a ser mais flutuantes.

A Figura 2 mostra, lado a lado, o volume de transações diárias para as duas empresas sob consideração. Note que não há muita diferença. Entretanto, ao contrário dos preços, parece haver uma alta correlação entre o volume de transações da IBM e da Fox.

5 Considerações

Uma observação muito importante que surgiu dessa análise primária dos dados foi que os pares de colunas “open” e “close”, “low” e “high”, são altamente correlacionados. Devem ser consideradas as ideias:

1. Criação de duas colunas artificiais com os valores (open + close) / 2 e (open - close), para substituir as atuais “open” e “close”. Assim, serão mantidos os graus de liberdade, mas com informações potencialmente melhor filtradas para o modelo.

2. Algo similar com as colunas “low” e “high”.

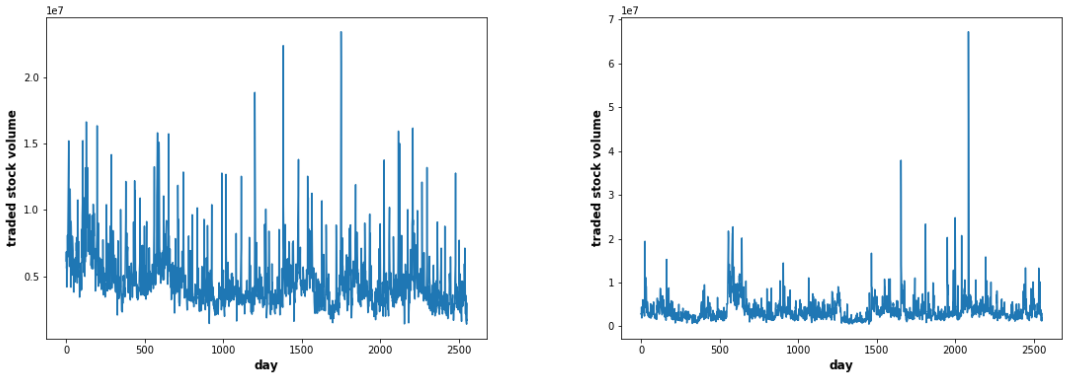

Figura 2: À esquerda e direita, o volume de transações diárias para a empresa IBM e Fox, respectivamente.

6 Código fonte produzido

Todo o código fonte produzido nesta fase, e que pode ser reutilizado em fases posteriores, pode ser encontrado no diretório “src/loading”.