UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMATICA
Programa de Pós Graduação em Ciência da Computação
CMP 259 – Descoberta de Conhecimento em Banco de Dados
Prof[a]. Karin Becker
Alunos: André Vieira Pigatto – 186041
Matheus Marrone Castanho - 207269

Trabalho Prático 2017 – Etapa 1

1. Compreensão do Negócio e Objetivos de Negócio

A bolsa de valores é o mercado organizado onde se negociam ações de sociedades de capital aberto e outros valores mobiliários, atuando, normalmente, como uma sociedade anônima, visando lucro através de seus serviços. Uma de suas principais funções é a divulgação imediata e detalhada das operações realizadas; essas, por sua vez, são, principalmente, operações de compra e venda de ações.

O valor de uma ação varia de forma contínua e estocástica, dependendo, de uma forma geral, da lei da oferta e procura: quanto maior o número de operações de compra de uma determinada ação, maior será seu valor de mercado; o contrário, também é verdadeiro. Sendo assim, simplificadamente, sabe-se que o valor de uma ação é diretamente dependente do interesse de investidores por aquela ação, ou seja, depende de um julgamento humano que, em geral, se baseia em uma série de fatores que levam o investidor a acreditar que o valor de uma determinada ação irá subir, e, por consequência, resultar em um lucro. Dessa forma, diversos fatores podem influenciar na decisão de compra ou venda de uma ação, dentre eles estão: os resultados da empresa já obtidos em um determinado período, a tendência de valor da ação, determinada através do comportamento gráfico que representa o valor da ação nas ultimas transações, o panorama geral político e econômico, dentre outros.

Uma das formas de determinar o panorama geral, é através das notícias apresentadas nos veículos de comunicação, como, por exemplo, jornais. Sendo assim, através das notícias de um jornal, o investidor pode fazer uma predição de tendências de mercado baseado no teor das últimas notícias e no seu conhecimento a priori. Dessa forma, como o valor das ações dependem justamente das operações de compra e venda, se um investidor fizer uma operação de um grande volume de ações ou se vários investidores fizerem várias operações, o valor daquela ação sofrerá variações; tendo em vista que a justificativa para essas operações pode ter sido baseada em uma notícia veiculada nos meios de comunicação, acredita-se que as notícias podem exercer uma influência sobre o valor da bolsa de valores.

Sendo assim, o objetivo de negócio primário desse trabalho é analisar se as notícias exercem, de fato, influência sobre valor da bolsa de valores, assim como encontrar quais as classes de notícias (Economia, Esportes, Política) que apresentam maior influência sobre as variações.

2. Objetivos de Mineração

O principal objetivo de mineração, é tentar encontrar uma relação entre o assunto das principais notícias de jornal publicadas ao longo de um dia, através da análise textual de suas manchetes, e a variação do valor médio das ações de empresas através de um indicador da bolsa de valores, para aquele mesmo dia. Ou seja, o objetivo é descobrir se as notícias tem alguma influência sobre o indicador da bolsa de valores, e, em caso positivo, qual o gênero da notícia que gera essa influência.

3. Escolha dos dados

Para possibilitar a análise do problema, é necessário utilizar alguma base de dados que contenha todas as variáveis de interesse organizadas e relacionadas entre si, dessa forma,

combinou-se 2 *datasets* principais para construir a base de dados que, posteriormente, será minerada; são eles:

a. Fonte das manchetes de notícias:

As manchetes das notícias foram retiradas de um *dataset* feito a partir da base de dados *Reddit WorldNews*, e contém as manchetes das 25 notícias mundiais mais quentes do dia, ranqueadas pelos próprios usuários do *Reddit*, para cada dia, no intervalo de 08/08/2008 a 07/01/2016. O *dataset* original pode ser encontrado em: https://www.kaggle.com/aaron7sun/stocknews/downloads/RedditNews.csv

b. Fonte dos dados da bolsa de Valores

O indicador da bolsa de valores escolhido para a análise é o indicador *Dow Jones Industrial Average*, que é baseado na cotação média das ações das 30 maiores e mais importantes empresas dos Estados Unidos. Dessa forma, obteve-se os dados diários de abertura, fechamento, alta e baixa do indicador DJI para o mesmo período das notícias, 08/08/2008 a 07/01/2016, através do *Yahoo Finance*, disponível no link abaixo:

https://finance.yahoo.com/quote/%5EDJI/history?period1=1218164400&period2=1467342000&interval=1d&filter=history&frequency=1d

Como um dos objetivos da proposta é analisar qual é o gênero das notícias que influencia na bolsa de valores, foi necessário classificar cada uma das notícias do *dataset* em relação ao gênero (Economia geral, esportes, dentre outros). Para isso, será utilizado um classificador, que, por sua vez, será treinado a partir de um conjunto de dados já rotulados que apresenta a notícia e o gênero ao qual ela pertence. Os dados de treinamento foram feitos a partir de *headers* de notícias publicadas no jornal The New York Times (no período de 1996 a 2006) e já foram rotulados com um código que denota o gênero e subgênero de cada uma de acordo com uma tabela fornecida pelo autor. O *dataset* e a tabela de códigos são fornecidos pela universidade Davis da Califórnia (UCDavis) e podem ser acessados através dos links abaixo:

http://psfaculty.ucdavis.edu/boydstun/Supplementary_Information_for_Making_the_News_files/nyt_ftpg_1996_2006.csv

http://psfaculty.ucdavis.edu/boydstun/Supplementary_Information_for_Making_the_News_files/NYT%20Front%20Page%20Policy%20Agendas%20Codebook.pdf

4. Formatação do *dataset* de trabalho

Para possibilitar a exploração dos dados, compilou-se os dados de cada um dos *datasets* em uma tabela única, onde a primeira coluna apresenta a data, a segunda, o valor de abertura da bolsa, da terceira a sexta, o valor de alta, baixa, fechamento e variação percentual do índice e da sétima em diante as manchetes das notícias em ordem decrescente de importância. A variação percentual diária do índice não estava presenta no *dataset* original, e, portanto, foi calculada a partir dos valores de abertura e fechamento diários. Sendo assim a o *dataset* gerado tem o formato apresentado na Tabela 1.

Tabela 1 – Formato do *dataset* de trabalho.

Date	Open	High	Low	Close	Var	Healine 25	...	Headline 1
01/07/2016	17924,24	18002,38	17916,91	17949,37	0,001402	A 117-year	...	Ozone Layer
...	...	...	...	...	...	...	...	...
08/08/2008	11432,09	11759,96	11388,04	11734,32	0,26437	Nim Chimp	...	Marriage

5. Análise Preliminar dos dados

Tanto a preparação dos dados quanto a análise preliminar foi realizada utilizando-se duas ferramentas: Python e o software R. A primeira análise foi feita com base na estrutura dos dados, onde observou-se que a tabela final tem 1989 linhas de dados, onde, cada linha é composta por 31 elementos, e, portanto, sabe-se que o banco de dados é constituído por 61659 elementos, onde 49725 elementos são as manchetes das notícias, 9945 elementos são os dados do índice DJI e 1989 são as datas. Nota-se que, pegando-se o período real, de 08/08/2008 a 07/01/2016, calcula-se um número de 2884 dias, no entanto, após examinar os dados, notou-se que os dados da bolsa e algumas das notícias não estavam disponíveis para algumas datas desse intervalo, portanto, na etapa de pré-processamento, analisou-se os dados brutos, deixando-se apenas os dados em que havia correspondência entre as datas de todos os bancos de dados originais.

Posteriormente à etapa inicial de pré-processamento, realizou-se uma análise estatística dos dados referentes aos valores dos índices DJI, onde encontrou-se, para o período de interesse, o valor médio 13459 pontos para abertura, 13463 para o fechamento, 13541 para o pico diário e 13373 para o vale diário; além de valores máximos e mínimos de 6547 e 18315 pontos para abertura, 6547 e 18312 para o fechamento, 6710 e 18351 para o pico diário e 6470 e 18273 para o vale diário. Além disso, analisou-se os dados respectivos às variações diárias, onde constatou-se uma variação diária média de 0,0366%, máxima de 10,9% e máxima negativa de 7,783%. O histograma da variação diária é apresentado na Figura 1.

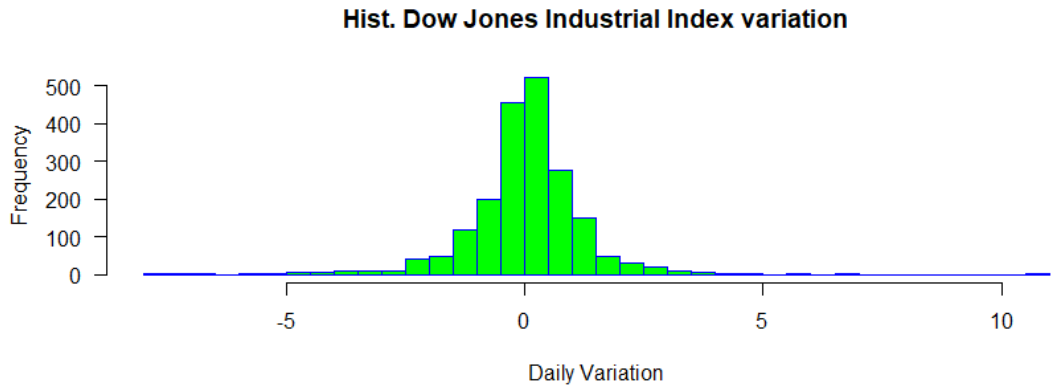

Figura 1 - Histograma da variação diária do índice DJI.

Através da análise do histograma, nota-se que a maior frequência de acontecimentos fica em torno de zero, o que indica que variações maiores, tanto para valorização do índice quanto para desvalorização são menos comuns. Para 75% dos casos, a variação diária ficou dentro dos intervalos de -0,439% e 0,569%. A Figura 2 apresenta o comportamento do variação ao longo do tempo e a Figura 3 apresenta o valor de abertura diário do índice DJI ao longo do tempo.

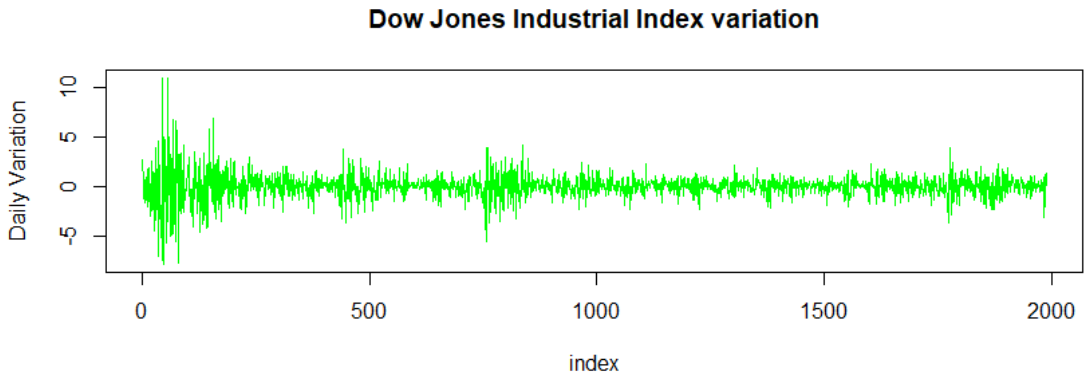

Figura 2 – Comportamento da variação diária ao longo do tempo.

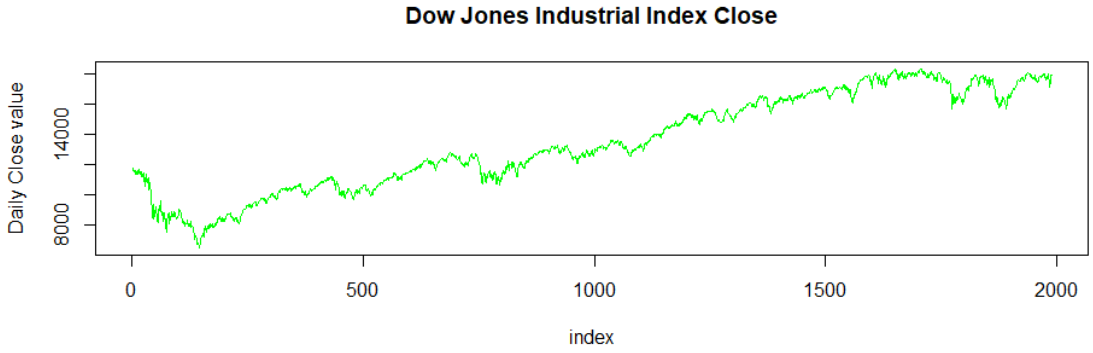

Figura 3 – Valor de fechamento diário do índice DJI no período de 2008 a 2016.

Pela análise das Figuras 2, nota-se uma mudança no comportamento do valor do índice DJI; pode-se perceber que, no início do gráfico, período relativo ao ano de 2008, a variação diária apresentava picos bem maiores do que no final do período de análise, referente ao ano de 2016, o que representa uma estabilidade maior do valor do índice. Nota-se também, que o valor da bolsa subiu substancialmente no período analisado. Esse comportamento, muito provavelmente, tem uma correlação com a crise que afetou o país em 2007-2008, e que pode ter gerado essas grandes oscilações, devido à incerteza dos investidores em relação à economia do pais; além disso, nota-se um vale, abaixo de 8000 pontos, na Figura 3, nesse período inicial.

6. Definição do Plano Preliminar

Até o presente momento tem-se os dados importados e organizados em R e Python, no entanto, ainda não é possível fazer uma análise da correlação entre a classe das notícias e o efeito no índice DJI. Para possibilitar a análise, tem-se como plano realizar as seguintes tarefas:

a. Criação de um Classificador e rotulação das manchetes

O processo de rotulação manual de 61659 manchetes de notícias é inviável e não faz o menor sentido, por esse motivo, a melhor opção para rotular as manchetes com os gêneros aos quais cada uma pertence é através de um classificador textual. Esse por sua vez, classifica as manchetes com base em um conhecimento obtido a priori, através do treinamento. Dessa forma, criar-se-á um classificador e treinar-se-á o mesmo utilizando o *dataset* do New York Times, conforme apresentado anteriormente, para que o mesmo identifique o tipo da notícia de acordo com o texto de sua manchete. Depois de testar o classificador, utilizar-se-á o mesmo para rotular as manchetes das notícias da base de dados de criada, gerando uma nova matriz, onde, no lugar das colunas das manchetes, serão colocados os códigos que indicam o assunto da notícia.

b. Análise dos dados rotulados

A segunda fase será a análise estatística dos dados rotulados, para que sejam definidas as características da base de dados, como, os percentuais de cada gênero de notícias que compõe a base, a distribuição temporal desses gêneros, dentre outros.

c. Análise de correlação das notícias com o índice DJI

A terceira fase será a análise estatística e de correlação entre os dados dos gêneros de notícias com as variações e tipos de variação (positiva e negativa) da bolsa.

d. Mineração dos dados

A quarta fase será a de geração de regras de associação entre as variações do índice e o gênero das notícias, a definição do suporte e confiança mínimos para a validade da regra e a filtragem de regras que não demonstram padrão significativo. Além disso, será realizada a análise das regras que podem demonstrar algum padrão significativo entre as manchetes e a os valores da variação da bolsa.

e. Resposta da Pergunta proposta

Por fim, pretende-se responder à pergunta se as notícias tem, ou não, uma relação com as variações no índice DJI da bolsa de valores americana e quais os gêneros de notícia que impactam mais nos valores.

7. Conclusões

A partir da análise do negócio, formulou-se os objetivos de negócio, ou seja, as perguntas que se deseja responder através da técnica de mineração de dados. A partir dos objetivos, analisou-se quais seriam os dados necessários para possibilitar a resposta à pergunta. Definidos os dados necessários para possível resolução dos problemas, buscou-se *datasets* que contivessem as informações. A partir dos *datasets* encontrados, analisou-se a estrutura de dados de cada um e realizou-se as transformações necessárias, de forma a possibilitar a união dos mesmos para formar a base de dados que será explorada. Após a estruturação da base de dados, analisou-se estatisticamente as variáveis de interesse e explorou-se algumas características interessantes sobre o comportamento do índice DJI da bolsa de valores americana.

A partir dos dados analisados, elaborou-se um plano de passos que serão utilizados ao longo do processo de descoberta de conhecimento, sendo o primeiro deles, a elaboração de um classificador, para rotular as manchetes das notícias. Por fim, após a classificação, será realizada uma nova análise estatística e de correlação, para posterirormente partir para a mineração propriamente dita e a descoberta do conhecimento. Sendo assim, nota-se que o processo de descoberta de conhecimento é um processo que depende de várias etapas, que, muitas vezes, devem ser repetidas para que se possa, efetivamente, alcançar os objetivos de negócio.