Raspagem e Visualização do preço de tickers

Jefferson Silvério

2022-04-21

Relatório

Esse relatório foi feito em uma manhã e tarde, com a duração aproximada de 4-5 horas totais. Apesar de ter familiaridade com R ainda não havia feito web scrappig e nem construido um app Shiny. Então, utilizei maior parte do tempo do relatório na busca de materiais e pacotes para realizar o scrapping, bem como das páginas que continham as informações desejadas. Não tive muita dificuldade no scrapping em si, depois de saber quais ferramentas usar e como usá-las.

Encontrei parte das informações das empresas em uma página da Wikipédia e o restante encontrei no site do Yahoo Finances. O site do yahoo finances também foi simples de raspar já que ele possui uma estrutura da URL bem definida, o que torna a busca dos dados de cada empresa bem simples.

Um problema que encontrei ao raspar os dados da página do Yahoo, porém, foi que fazendo a raspagem da tabela de fechamento eram devolvidos apenas 100 registros, mesmo aparecendo mais dados no navegador. Assim, optei por usar a URL de download que existe na própria página e que retorna um *csv* do intervalo de tempo escolhido. Além disso, como a bolsa opera só em dias úteis tive que usar o pacote *bizdays* para fazer o cálculo da data dos 200 dias úteis anteriores ao dia atual. Com esses dois problemas resolvidos foi possível terminar o relatório em PDF sem mais problemas.

Infelizmente não consegui terminar o app Shiny a tempo. Assim como o scrapping ainda não tinha experiência com essa ferramenta. Comecei a ler alguns materias da documentação e tutoriais mas não tive tempo para implementar uma solução. Espero que apesar disso o relatório em PDF esteja satisfatório.

Além disso ficou faltando uma limpeza mais minuciosa nos dados da empresa como: separar o nome dos executivos por vígula ou dividir em diferentes colunas de acordo com o cargo e retirar o ano que está entre parêntesis na coluna do número de empregados. Essas tarefas poderiam ser feitas usando R ou aplicando um regex nessas colunas, porém a solução a ser feita (dividir em colunas ou mantér apenas uma) dependeria do uso que se quer dar a esses dados.

Além desse relório também criei um script na pasta *script* que baixa e salva todos os dados das empresas em um novo csv.

Tarefas

1. Informações da empresa:
    - ☒ Nome completo da empresa
    - ☒ Endereço completo da empresa
    - ☐ Telefone
    - ☒ Setor
    - ☒ Indústria
    - ☒ Número de funcionários
    - ☒ Nome dos principais executivos

☒ Obter dados sobre o valor ajustado de fechamento das ações dos últimos 200 dias de cada ticker

- ☒ Obter dados sobre o volume de ações negociados nos últimos 200 dias de cada ticker
- ☒ Processar e manipular os dados obtidos para um formato de dado fácil de ser processado.
- ☒ Apresentar em um mesmo gráfico, as informações de preço de cada ticker ao longo do tempo, mas somente o último dia de cada mês.
- ☐ Implementar uma interface Shiny para apresentar os gráficos criados nas tarefas 4 e 5.
- ☐ O usuário deve ter a opção de escolher se ele deseja visualizar ou o preço das ações ou o volume negociado

Informações das empresas

Listar página da wikipedia dos tickers selecionados

O melhor caminho para buscar as informações das empresas foi por meio da Wikipedia. Assim, iniciando por essa lista das 500 maiores empresas na Wikipedia podemos buscar a página da Wikipedia de cada uma das empresas por meio do simbolo

symbol	security	gics_sector	gics_sub_industry	headquarters_location	link
GOOG	Alphabet (Class C)	Communication Services	Interactive Media & Services	Mountain View, California	http://en.wikipedia.org/wiki/Alphabet_Inc.
AMZN	Amazon	Consumer Discretionary	Internet & Direct Marketing Retail	Seattle, Washington	http://en.wikipedia.org/wiki/Amazon_(company)
AAPL	Apple	Information Technology	Technology Hardware, Storage & Peripherals	Cupertino, California	http://en.wikipedia.org/wiki/Apple_Inc.
CSCO	Cisco	Information Technology	Communications Equipment	San Jose, California	http://en.wikipedia.org/wiki/Cisco
INTC	Intel	Information Technology	Semiconductors	Santa Clara, California	http://en.wikipedia.org/wiki/Intel

Buscar informações das empresas na Wikipedia

A partir das informações acima e do link das páginas da wikipedia de cada empresa buscamos o restante das informações. Nas páginas da Wikipedia encontramos o número de empregados e os nomes dos principais executivos de cada empresa. Ao final adicionamos essas novas informações à tabela apresentada anteriormente.

symbol	name	gics_sector	gics_sub_industry	headquarters_location	headquarters
GOOG	Alphabet (Class C)	Communication Services	Interactive Media & Services	Mountain View, California	Googleplex, Mountain View, California, U.S.
AMZN	Amazon	Consumer Discretionary	Internet & Direct Marketing Retail	Seattle, Washington	Seattle, Washington, U.S.
AAPL	Apple	Information Technology	Technology Hardware, Storage & Peripherals	Cupertino, California	1 Apple Park Way, Cupertino, California, U.S.
CSCO	Cisco	Information Technology	Communications Equipment	San Jose, California	San Jose, California, United States[1]
INTC	Intel	Information Technology	Semiconductors	Santa Clara, California	Santa Clara, California, U.S.

key_people	number_of_employees
John L. Hennessy (Chairman)Sundar Pichai (CEO)Ruth Porat (CFO)	156,500 (Dec 2021)
Jeff Bezos(executive chairman)Andy Jassy(president and CEO)	1,608,000 (Dec. 2021)[1]U.S.: 950,000 (June 2021)[4]
Tim Cook (CEO)Arthur D. Levinson (chairman)Jeff Williams (COO)Luca Maestri (CFO)	154,000 (2021)
Chuck Robbins(CEO & Chairman)	79,500 (2021)[2]
Omar Ishrak (Chairman)Pat Gelsinger (CEO)David Zisner (CFO)[1]	121,100 (2021)[2]

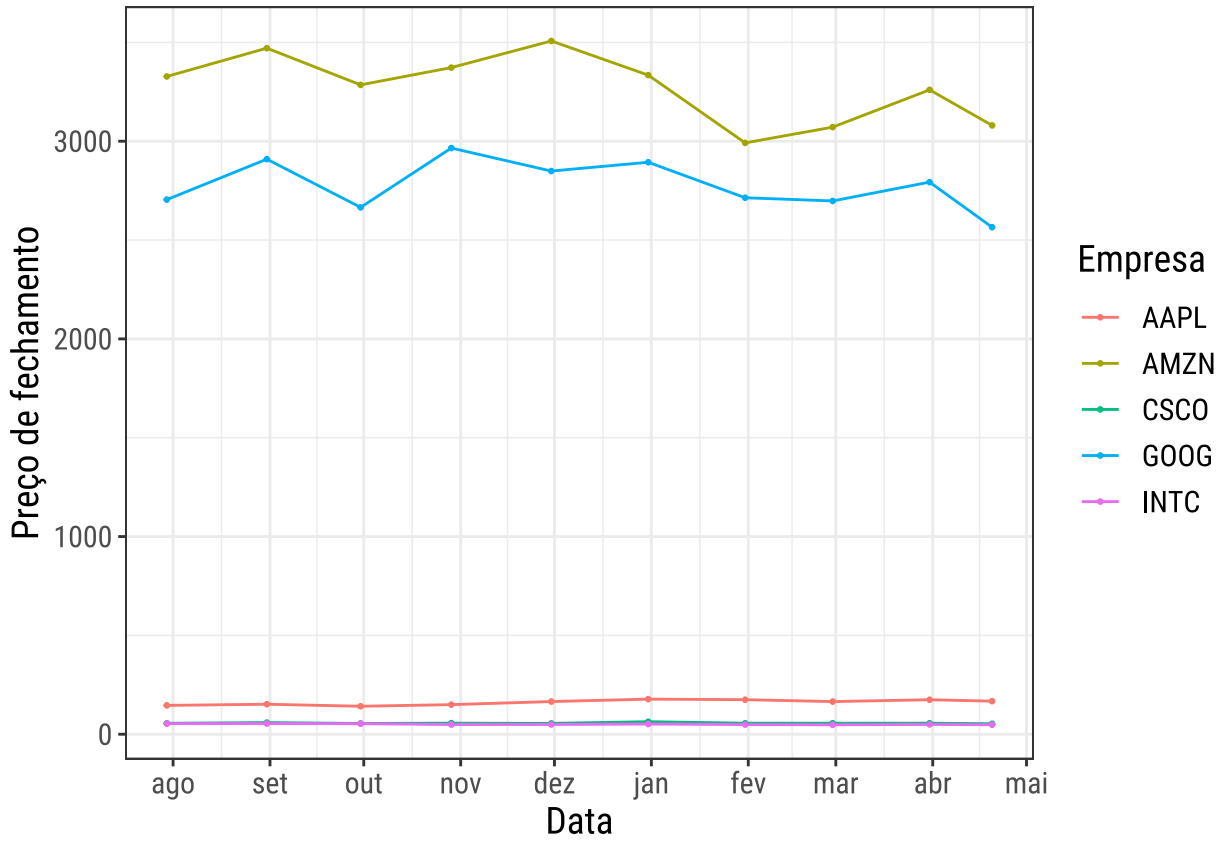
Busca valores de fechamento

Os dados do fechamento dos últimos 200 dias dos tickers é calculado com base no calendário de dias úteis do Brasil. Esses dados são baixados diretamente do site do Yahoo Finances. Um exemplo do formato de dados obtidos é apresentado abaixo. São quatro colunas: symbol (sigla de empresa), data, valor de fechamento e volume de venda no dia.

Gráfico de variação do preço de cada ticker

O gráfico abaixo mostra o valor de fechamento dos tickers de cada empresa no último dia de cada mês.

symbol	date	close	volume
AAPL	2021-07-06	142.02	108181800
AAPL	2021-07-07	144.57	104911600
AAPL	2021-07-08	143.24	105575500
AAPL	2021-07-09	145.11	99890800
AAPL	2021-07-12	144.50	76299700
AAPL	2021-07-13	145.64	100827100



Referências

- https://blog.curso-r.com/posts/2018-02-18-fluxo-scraping/
- https://blog.curso-r.com/posts/2018-03-19-scraper-cetesb/
- https://www.dataquest.io/blog/web-scraping-in-r-rvest/
- http://wilsonfreitas.github.io/posts/bizdays-dias-uteis-no-r.html

Páginas para raspagem

- https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
- https://finance.yahoo.com/quote/