INE5454 - Web crawler

Gustavo Fukushima de Salles

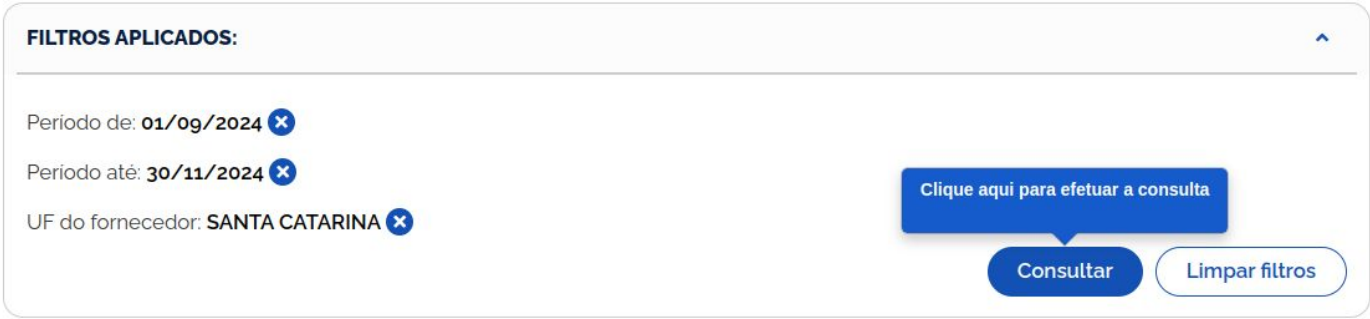
Introdução

- Desenvolvimento de crawler e scripts consumidores de API
- Crawler extrai dados de notas fiscais eletrônicas
- Fonte: Portal da Transparência da CGU
- Scripts fazem requisições da API de contratos e licitações
- Fonte: Portal da Transparência da Prefeitura de Florianópolis

Objetivo do crawler

- Receber datas inicial e final
- Receber informações de UF (opcional)
- Fazer requisição da API da página com as chaves de acesso
- Acessar cada página individual da NF-e a partir da chave de acesso
- Fazer o scraping de dados relevantes da página
- Fazer requisição da API da próxima página
- Repetir requisição e scraping enquanto houver mais registros
- Gerar arquivo JSON com dados de NF-e

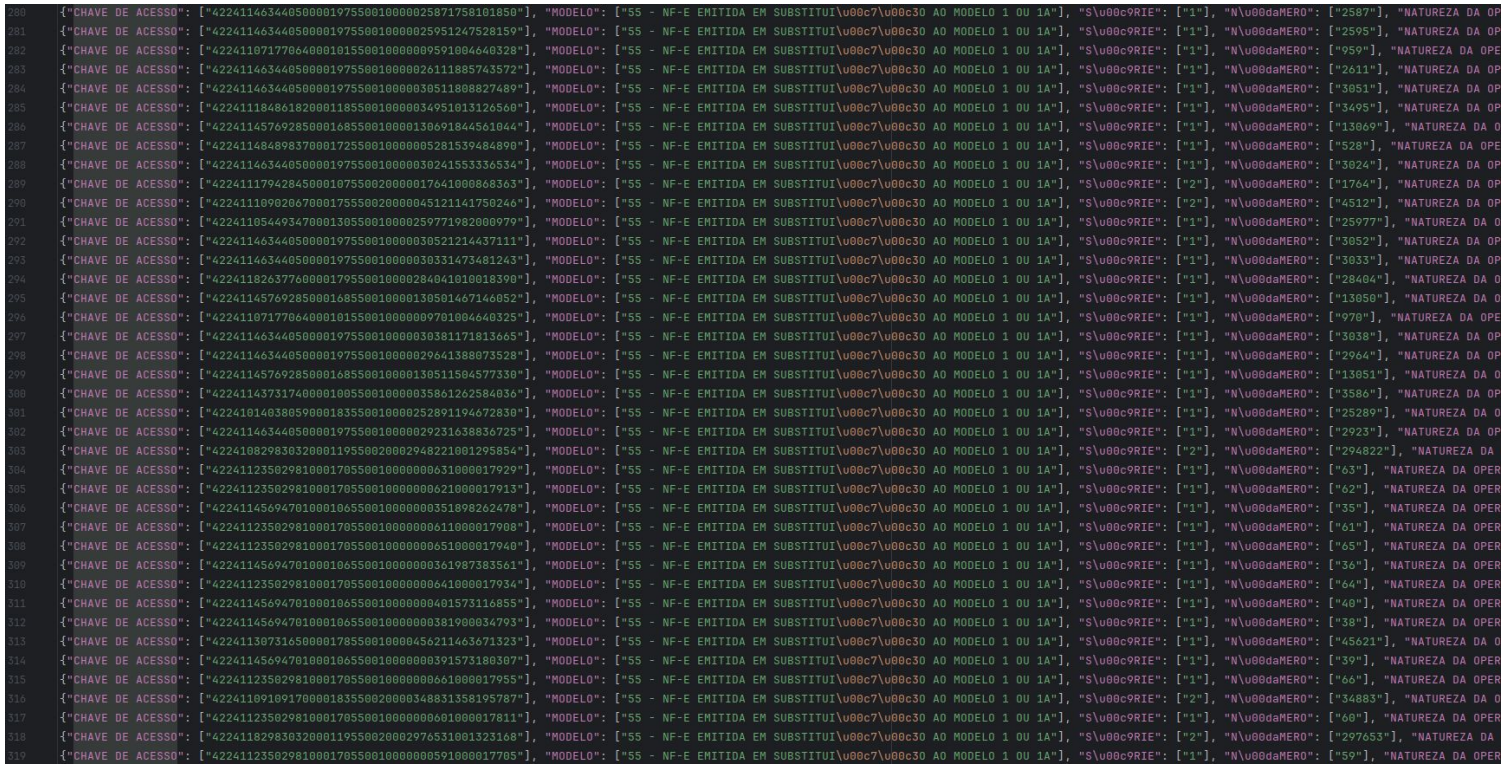
Exemplo dos dados de entrada



▼ General

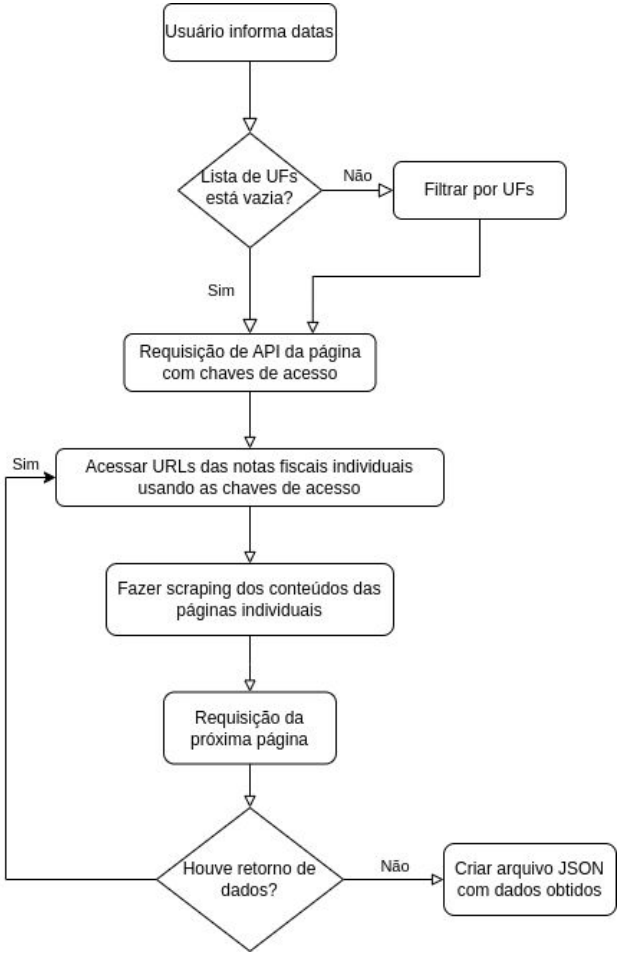
Request URL:	https://portaldatransparencia.gov.br/notas-fiscais/consulta/resultado?paginacaoSimples=true&tamanhoPagina=10&offset=0&direcaoOrdenacao=asc&colunaOrdenacao=municipioFornecedor&de=01%2F09%2F2024&ate=30%2F11%2F2024&ufFornecedor=SC&colunasSelecionadas=linkDetalhamento%2CorgaoSuperiorDestinatario%2CorgaoDestinatario%2CnomeFornecedor%2CcnpjFornecedor%2CmunicipioFornecedor%2CufFornecedor%2CchaveNotaFiscal%2CvalorNotaFiscal%2CdataEmissao%2CtipoEventoMaisRecente%2Cnumero%2Cserie&_=1733179020594
Request Method:	GET
Status Code:	200 OK
Remote Address:	[2600:9000:28b5:2e00:12:199c:72c0:93a1]:443
Referrer Policy:	strict-origin-when-cross-origin

Exemplo dos dados de saída



```
{"CHAVE DE ACESSO": ["42241146344050000197550010000025871758101850"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["2587"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241146344050000197550010000025951247528159"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["2595"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241107177064000101550010000009591004640328"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["959"], "NATUREZA DA OPE
{"CHAVE DE ACESSO": ["42241146344050000197550010000026111885743572"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["2611"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241146344050000197550010000030511808827489"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3051"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241118486182000118550010000034951013126560"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3495"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241145769285000168550010000130691844561044"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["13069"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241148489837000172550010000005281539484890"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["528"], "NATUREZA DA OPE
{"CHAVE DE ACESSO": ["42241146344050000197550010000030241553336534"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3024"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241117942845000107550020000017641000868363"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["2"], "N\u00daMERO": ["1764"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241110902067000175550020000045121141750246"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["2"], "N\u00daMERO": ["4512"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241105449347000130550010000025977198200097 9"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["25977"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241146344050000197550010000030521214437111"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3052"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241146344050000197550010000030331473481243"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3033"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241182637760000179550010000284041010018390"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["28404"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241145769285000168550010000130501467146052"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["13050"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241107177064000101550010000009701004640325"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["970"], "NATUREZA DA OPE
{"CHAVE DE ACESSO": ["42241146344050000197550010000030381171813665"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3038"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241146344050000197550010000029641388073528"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["2964"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241145769285000168550010000130511504577330"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["13051"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241143731740000100550010000035861262584036"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["3586"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241014038059000183550010000252891194672830"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["25289"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241146344050000197550010000029231638836725"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["2923"], "NATUREZA DA OP
{"CHAVE DE ACESSO": ["42241082983032000119550020002948221001295854"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["2"], "N\u00daMERO": ["294822"], "NATUREZA DA
{"CHAVE DE ACESSO": ["42241123502981000170550010000000631000017929"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["63"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241123502981000170550010000000621000017913"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["62"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241145694701000106550010000000351898262478"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["35"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241123502981000170550010000000611000017908"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["61"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241123502981000170550010000000651000017940"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["65"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241145694701000106550010000000361987383561"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["36"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241123502981000170550010000000641000017934"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["64"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241145694701000106550010000000401573116855"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["40"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241145694701000106550010000000381900034793"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["38"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241130731650000178550010000456211463671323"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["45621"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241145694701000106550010000000391573180307"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["39"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241123502981000170550010000000661000017955"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["66"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241109109170000183550020000348831358195787"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["2"], "N\u00daMERO": ["34883"], "NATUREZA DA O
{"CHAVE DE ACESSO": ["42241123502981000170550010000000601000017811"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["60"], "NATUREZA DA OPER
{"CHAVE DE ACESSO": ["42241182983032000119550020002976531001323168"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["2"], "N\u00daMERO": ["297653"], "NATUREZA DA
{"CHAVE DE ACESSO": ["42241123502981000170550010000000591000017705"], "MODELO": ["55 - NF-E EMITIDA EM SUBSTITUI\u00c7\u00c3O AO MODELO 1 OU 1A"], "S\u00c9RIE": ["1"], "N\u00daMERO": ["59"], "NATUREZA DA OPER
```

Descrição do crawler



Obrigado!

Crawler

- Desenvolvido em Python (scrapy, json)
- Fonte para o projeto CÉOS
- É possível coletar dados de todas as UFs (somente SC no arquivo)
- Dados disponibilizados em JSON

Dados qualitativos

- Nomes, CPF/CNPJ, município e UF de emitente e destinatário
- Órgão superior do destinatário e órgão da entidade vinculada
- Evento mais recente e data respectiva
- Natureza e destino da operação

Dados quantitativos

- Valor monetário da nota fiscal

Consumidor de APIs

- Desenvolvido em Python (scrapy, json)
- Fonte para o projeto CÉOS
- Dados de Florianópolis
- Dados disponibilizados em JSON

Dados qualitativos

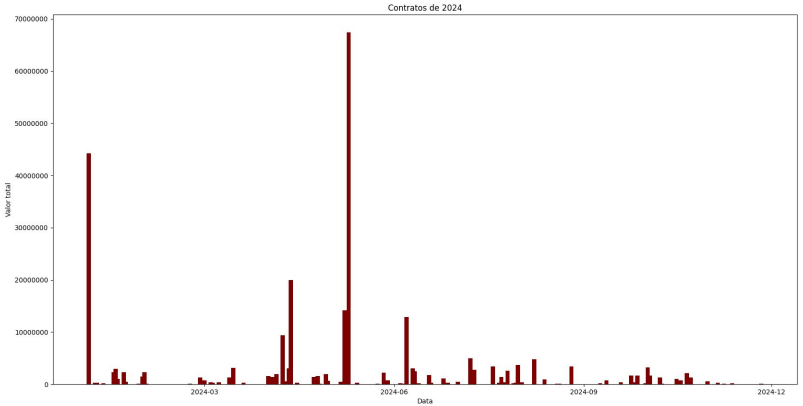
- Número, resumo do objeto, unidade gestora, lista de itens, lista de empenhos
- Data de assinatura, início da vigência, vencimento, fornecedor (contrato)
- Modalidade, finalidade, forma de julgamento, advogado, datas de emissão e abertura, lista de vencedores, lista de textos, lista de contratos (licitação)

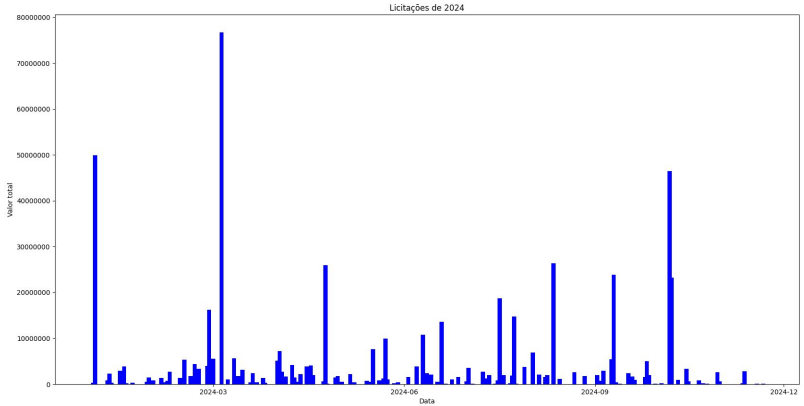
Dados quantitativos

- Valor total, valor unitário de um item, quantidade de itens
- Valor total dos itens (contrato)
- Valor unitário e quantidade oferecidos pelos vencedores (licitação)

Análise dos resultados dos consumidores de API

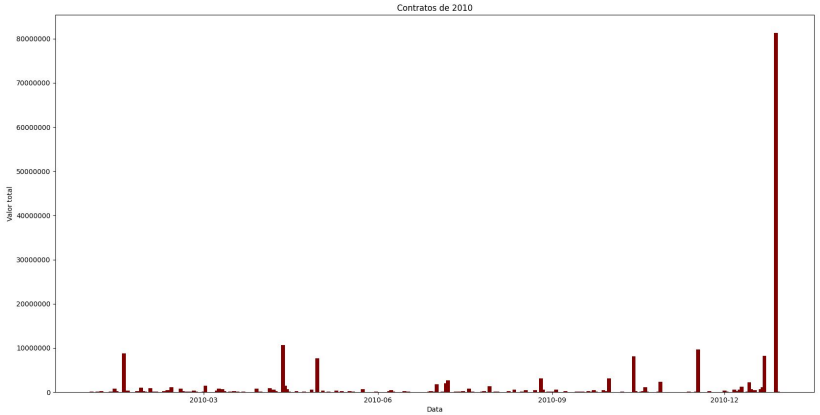
- Geração de gráficos para cada ano (1999 a 2024)
- Valores de contratos/licitações X datas

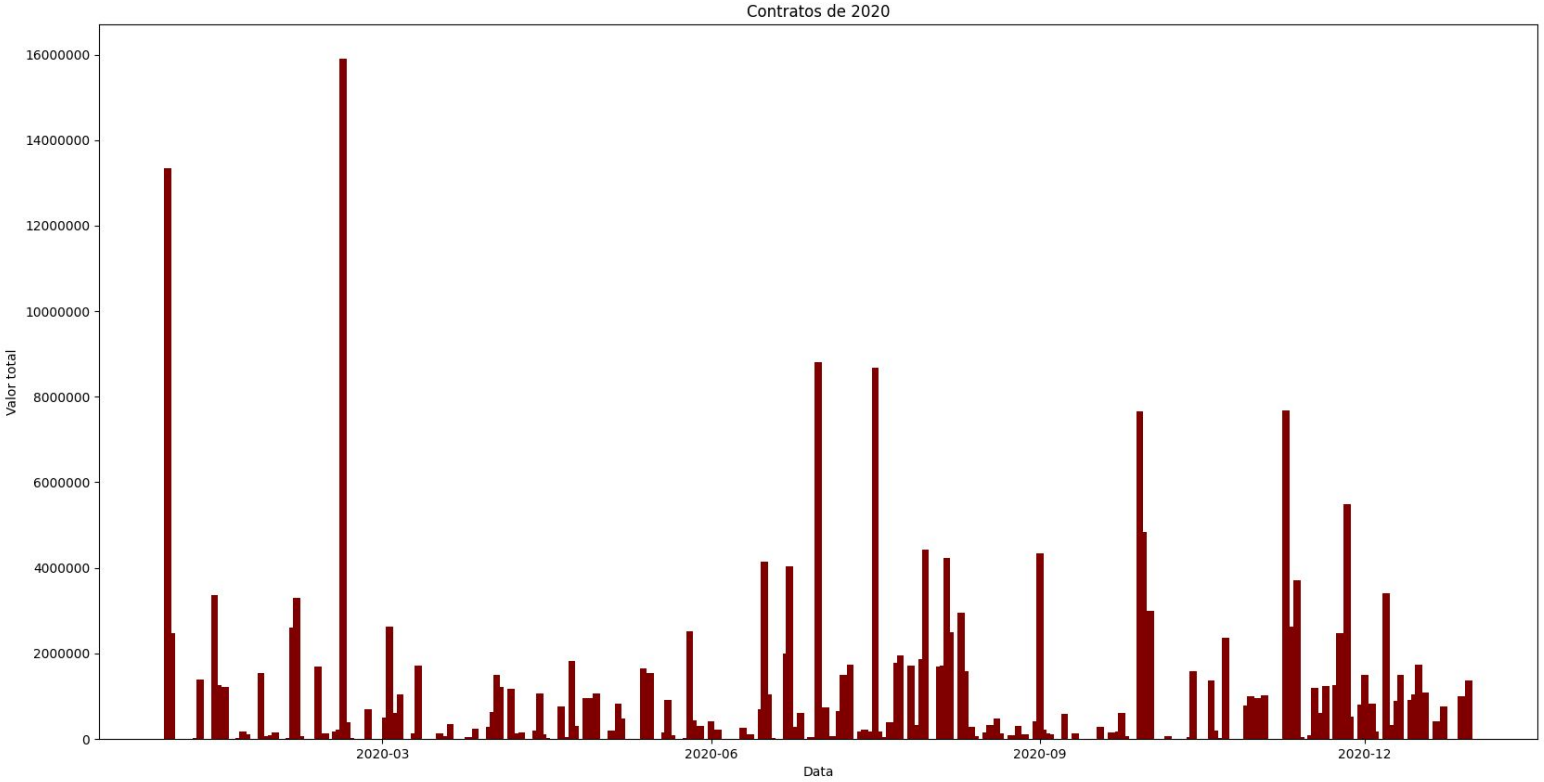




Análise dos contratos

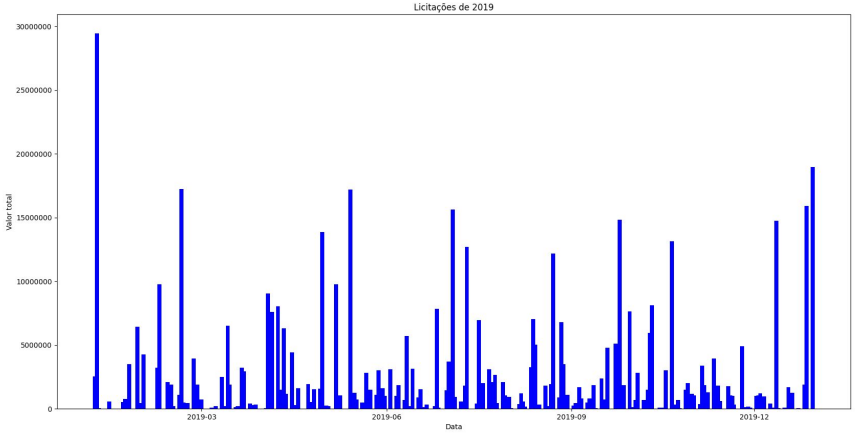
- Mais registros a partir de 2002
- Outliers no final de 2010 e meio de 2017 (valores elevados)
- Em 2020, queda após março até julho

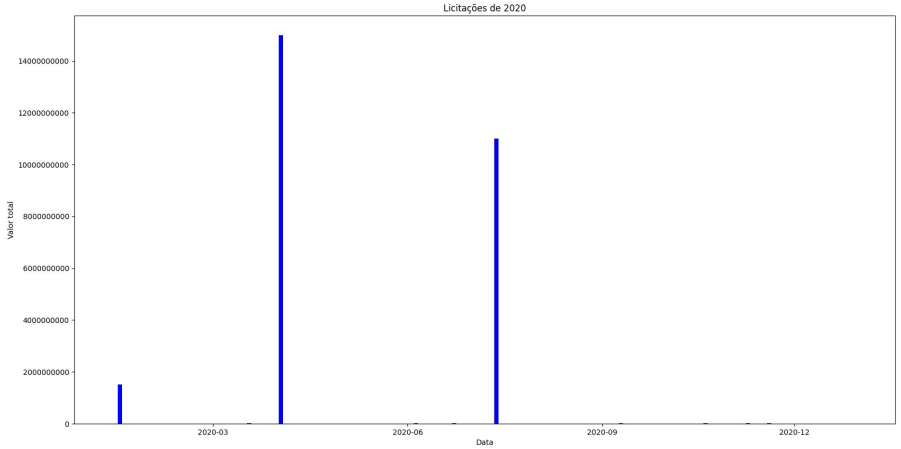



Contratos de 2020
Valor total
16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
2020-03
2020-06
2020-09
2020-12
Data

Análise das licitações

- Poucos registros com valor até 2018
- Gráfico de 2020 menos distribuído em comparação com outros





Casos de falha

- API não permite fazer requisição de mais de um ano (ex. 31/12/2023 até 01/01/2024)
- Contratos do dia 12/08/2024 estão nulos
- Links inválidos no Portal de Transparência da CGU

Conclusão

- Maiores dificuldades foram aprender a utilizar o scrapy e encontrar a fonte dos contratos e licitações
- Gráficos não foram suficientes para identificar tendências
- Espera-se que o trabalho contribua para carga do data warehouse do CÉOS