Uso de web crawling e BI para listagem de nomeados em cargos públicos

Natália Quirino de Oliveira

nataliaquirino@gmail.com

Abstract. *This paper describes an approach for crawling government websites for finding open and public information about public agent nominations and displaying relevant info on Power BI dashboards.*

Resumo. *Este paper descreve uma abordagem para realizar web crawling sites governamentais para se encontrar dados abertos e públicos sobre nomeações para cargos públicos e exibir informações relevantes em dashboards do Power BI.*

1. Informações gerais

Informações governamentais abertas nem sempre estão disponíveis de forma estruturada. Neste caso, pode surgir a necessidade de se obter conteúdos de maneira não estruturada e manipular e consolidar os dados em novos formatos. O escopo que escolhemos é nomeação para cargos em algumas cidades capixabas. Um ponto forte da abordagem é que pode ser ampliado gradativamente, com a incorporação de novas origens de dados.

2. Extração de dados

Primeiramente, os diários de cada fonte de dados são obtidos através do comando wget. Cada cidade é uma fonte de dados e terá seu diretório.

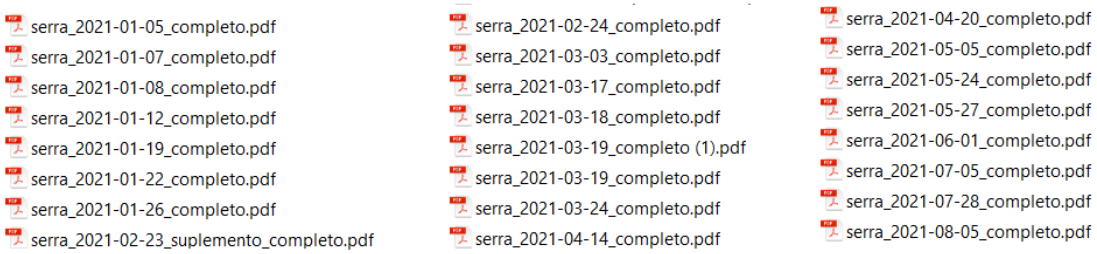
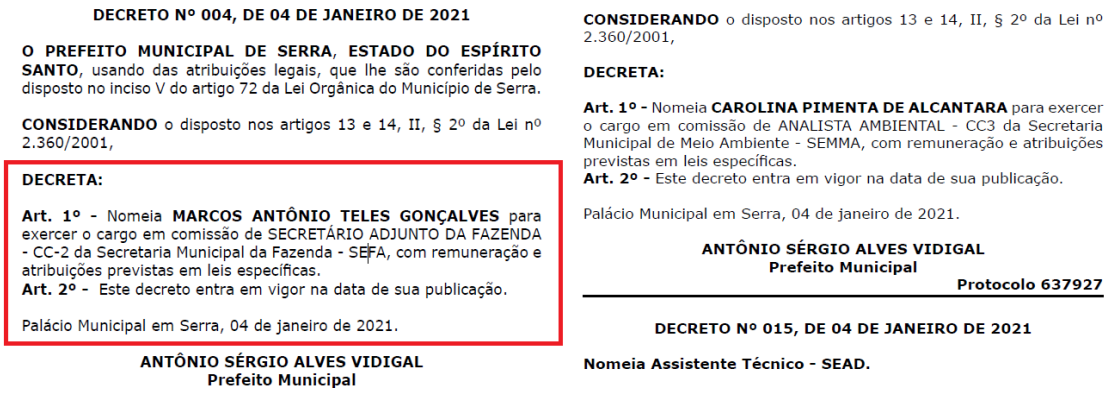


Figura 1. Diretório da cidade Serra - ES após o download dos diários oficiais

Utilizamos apenas o diário oficial do executivo ou legislativo de cada cidade para fim de protótipo, visto ser uma prova de conceito. Segue um dos arquivos pdf com exemplo:



DECRETO Nº 004, DE 04 DE JANEIRO DE 2021

O PREFEITO MUNICIPAL DE SERRA, ESTADO DO ESPÍRITO SANTO, usando das atribuições legais, que lhe são conferidas pelo disposto no inciso V do artigo 72 da Lei Orgânica do Município de Serra.

CONSIDERANDO o disposto nos artigos 13 e 14, II, § 2º da Lei nº 2.360/2001,

DECRETA:

Art. 1º - Nomeia **MARCOS ANTÔNIO TELES GONÇALVES** para exercer o cargo em comissão de SECRETÁRIO ADJUNTO DA FAZENDA - CC-2 da Secretaria Municipal da Fazenda - SEFA, com remuneração e atribuições previstas em leis específicas.
Art. 2º - Este decreto entra em vigor na data de sua publicação.

Palácio Municipal em Serra, 04 de janeiro de 2021.

ANTÔNIO SÉRGIO ALVES VIDIGAL
Prefeito Municipal

CONSIDERANDO o disposto nos artigos 13 e 14, II, § 2º da Lei nº 2.360/2001,

DECRETA:

Art. 1º - Nomeia **CAROLINA PIMENTA DE ALCANTARA** para exercer o cargo em comissão de ANALISTA AMBIENTAL - CC3 da Secretaria Municipal de Meio Ambiente - SEMMA, com remuneração e atribuições previstas em leis específicas.
Art. 2º - Este decreto entra em vigor na data de sua publicação.

Palácio Municipal em Serra, 04 de janeiro de 2021.

ANTÔNIO SÉRGIO ALVES VIDIGAL
Prefeito Municipal

Protocolo 637927

DECRETO Nº 015, DE 04 DE JANEIRO DE 2021

Nomeia Assistente Técnico - SEAD.

Figura 2. Exemplo de trecho de diário oficial (parte de interesse em vermelho)

Cada diário possui sua própria formatação, por isso, cada fonte deve ter sua expressão regular. Por vezes, podem ser necessárias duas, a fim de obter o nome e cargo separadamente. Nosso objetivo é um bloco de texto que engloba uma ou várias nomeações, possuem um final delimitado (por exemplo, a string "REF.")

```
expRegular = "NOMEAR(.*?)CARGO(.*?)REF."

expRegular2 = "Nomeia(.*?)Art."

x1 = re.search(expRegular, Text, flags=(re.DOTALL | re.IGNORECASE ))
x2 = re.search(expRegular2, Text, flags=(re.DOTALL | re.IGNORECASE ))
```

Figura 3. Exemplo de regex da cidade Serra - ES

```
expRegular = "NOMEIA(.*?)\n(.*?)\n(.*?)\n"

x1 = re.search(expRegular, aux, re.DOTALL)

# Obtem o cargo
expRegularCargo1 = x1.group()+"\s{0,}para\s{0,}(.*)\s{0,}cargo\s{0,}(.*)\s{0,}(.*)\s{0,}(.*
```

Figura 4. Exemplo de regex para obter o nome e cargo separadamente - cidade Vila Velha - ES

```
expRegular = "\,\s{0,}[A-Z]{3,}\s[A-Z]{2,}\s[A-Z]{2,}(.*?)\,"

x1 = re.search(expRegular, aux, re.DOTALL)
```

Figura 5. Exemplo de regex de diários estaduais do ES

```
        expRegular = "(.*?)\s-\s[0-9]+\sHORAS\n"
        cargos = re.findall(expRegular, Text,re.MULTILINE)
```

Figura 6. Exemplo de regex de diários Vitória - ES

Após a execução das buscas pelas expressões regulares em cada pdf, um arquivo txt é criado com cada ocorrência (podem haver várias em um mesmo pdf), apenas com a parte relevante em texto livre.

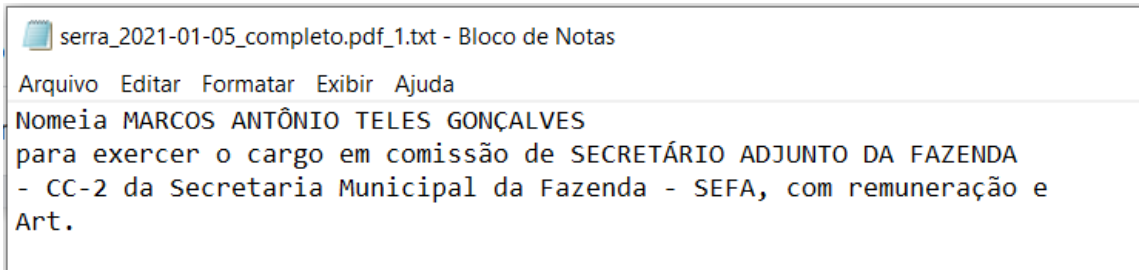


Figura 7. Exemplo de txt com o trecho de interesse da figura X extraído

O próximo passo é, do parágrafo obtido, separar apenas o nome e cargo do nomeado em duas strings, para que seja possível finalmente gerar o csv estruturado.

```
expRegular = r",\n{0,}\s{0,}[A-Z\u00C0-\u00FF\s.]+\n{0,}\s{0,},"
l1 = re.findall(expRegular, Text, flags=(re.DOTALL ))
print(l1)

# Obtem o cargo, parando em um ponto
expRegular = "(cargo)(.*?)\."
l2 = re.findall(expRegular, Text, flags=(re.DOTALL ))
print(l2)
```

Figura 8. Exemplo de expressão regular para geração dos dados estruturados

Cada arquivo de texto gera um csv de apenas uma linha com aquela nomeação estruturada.

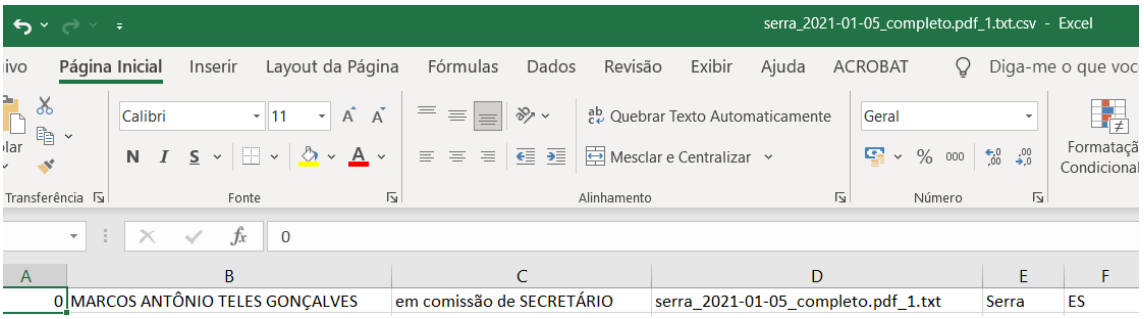


Figura 9. Exemplo de csv gerado a partir de um txt

O consolidado é o merge de todos os csvs, ou seja, de todas nomeações daquela localidade. Reparar que no consolidado gerado, o campo data ainda não está perfeitamente formatado.

A	B	C	D	E	F
posicao	nome	cargo	data	cidade	estado
0	MARCOS ANTÔNIO TELES	em comissão de SECRETÁRIO	serra_2021-01-05_completo.pdf_1.txt	Serra	ES
0	STEVEN FABIANO VIEIRA para	em comissão de ASSESSOR PARA	serra_2021-02-23_suplemento_completo.pd	Serra	ES
0	MAIQUE UELINTON LARES MOTA	em comissão de ASSISTENTE	serra_2021-02-24_completo.pdf_5.txt	Serra	ES
0	ALEX SHANDER SILVA,	em	serra_2021-03-18_completo.pdf_3.txt	Serra	ES
0	JEFERSON BRAVIM DE OLIVEIRA	em comissão de CHEFE DA DIVISÃO	serra_2021-03-24_completo.pdf_4.txt	Serra	ES
0	FABIANY BINDA WRUCK LOUREIRO,	em comissão de SECRETÁRIA	serra_2021-06-01_completo.pdf_3.txt	Serra	ES
0	LINA DE SOUZA CARVALHO, para	de Administração e Recursos Human	serra_2021-07-05_completo.pdf_4.txt	Serra	ES

Figura 10. Exemplo de consolidado (no caso, trecho da cidade de Serra - ES)

Segue o trecho de código que mostra o tratamento de dados após a geração do consolidado. No caso, a coluna data, que nas fontes analisadas, é obtida a partir do nome do arquivo de origem. Novas fontes podem ter outras características.

```
df_serra2 = pd.read_csv('C:\\tcc\\serra2\\CONSOLIDADO_SERRA2_PYPDF2.csv', encoding = 'latin', delimiter = ";")
df_serra2['data'] = df_vix['data'].str.replace("serra_","")
df_serra2['data'].replace(to_replace=[r"_completo(.*?).txt"], value=[""], regex=True, inplace=True)
df_serra2.to_csv('C:\\tcc\\SERRA2_PYPDF2.csv',index=False, sep=";")
```

Figura 11. Exemplo código de de tratamento de dados para Serra - ES

	A	B	C	D	E	F
1	posicao	nome	cargo	data	cidade	estado
2	0	MARCOS ANTÃ”NIO TELES GONÃ‡ALVES	em comissÃ£o de SECRETÃRIO	07/07/2021	Serra	ES
3	0	STEVEN FABIANO VIEIRA para	em comissÃ£o de ASSESSOR PARA	07/07/2021	Serra	ES
4	0	MAIQUE UELINTON LARES MOTA	em comissÃ£o de ASSISTENTE	07/07/2021	Serra	ES
5	0	ALEX SHANDER SILVA,	em	07/07/2021	Serra	ES
6	0		em comissÃ£o de CHEFE DA	07/12/2020	Serra	ES
7	0	FABIANY BINDA WRUCK LOUREIRO,	em comissÃ£o de SECRETÃRIA	07/12/2020	Serra	ES
8	0	LINA DE SOUZA CARVALHO, para	de AdministraÃ§Ã£o e Recursos Huma	07/12/2020	Serra	ES

Figura 12. Exemplo do consolidado após trabalhar dados

No exemplo citado da Serra - ES, houve tratamento da coluna data. Varia de fonte a fonte quais colunas necessitam tratamento.

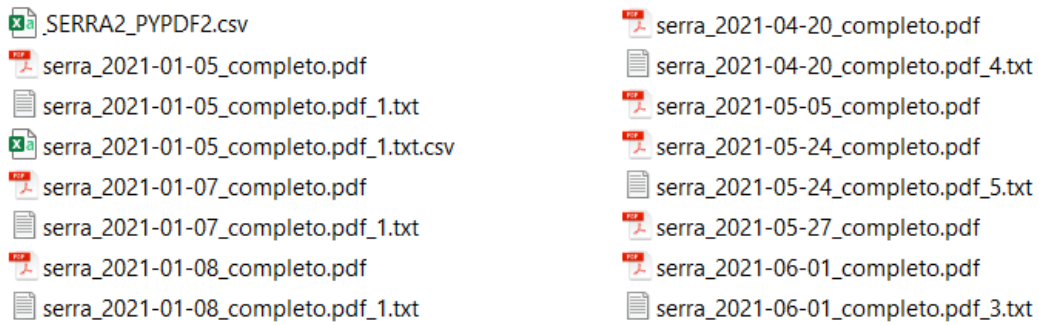


Figura 13. Pasta de uma cidade (Serra - ES) após criação de todos os arquivos

3. Visualização dos dados

Como visto nas figuras anteriores, mas se formalizando o modelo de dados, segue estrutura da tabela de nomeações:

Tabela 1. Atributos da tabela de nomeações

posicao (INT)	posição na qual a pessoa se classificou, zero se não for disponível
nome (VARCHAR)	nome do nomeado
cargo (VARCHAR)	cargo para o qual foi nomeado

data (DATE)	data do diário (pode não corresponder à data exata da nomeação
cidade (VARCHAR)	Cidade do diário, onde o cargo será exercido
estado (VARCHAR)	Estado do diário, onde o cargo será exercido

Como se pode ver, a tabela no Power BI é UNION de todas as fontes de dados. À medida que pessoas criarem os scripts para outras localidades, basta atualizar o UNION para incorporá-los, desde que os csvs estejam no mesmo formato.

```
1  Tabela 2 = UNION(ES_FITZ,ES_PYPDF2,SERRA_PYPDF2,SERRA2_PYPDF2,VV_PYPDF2,VIX_PYPDF2)
```

nome	cargo	data	cidade	estado
ARLETELÃ□ZARO	emcomissÃ£o de Assistente deGerencia, Ref'	*01/04/2003 00:00:00*	Serra	ES
MARCIADOS SANTOS SOBRAL	de SupervisorI, QC-01, da Secretaria deEstado do Planejan	*03/04/2003 00:00:00*	Serra	ES
DAVID DEPAULA LUIZ	deprovimento em comissÃ£o deAuxiliar TÃ©cnico, Ref'	*03/04/2003 00:00:00*	Serra	ES
VERA LUCIA GOMES FARIA	em comissÃ£o deSupervisor de SeguranÃ§a , Ref'	*07/04/2003 00:00:00*	Serra	ES
LIDIA CELINA DOS SANTOS	em comissÃ£ode Gerente, Ref'	*15/04/2003 00:00:00*	Serra	ES
MARIZALAGE DA SILVA	em comissÃ£ode ASSISTENTE DE GABINETE,QC'	*02/05/2003 00:00:00*	Serra	ES
CARMELITA DA PENHAMIRANDA	em comissÃ£o deSUPERVISOR DE AREAFAZENDARIA, QC'	*02/05/2003 00:00:00*	Serra	ES
ROSENIFABRES COELHO	em comissÃ£ode SUBGERENTE DE EXECUÃ‡ÃƒOFINANCEIF	*02/05/2003 00:00:00*	Serra	ES
ANGELACELINA HOTT GOMES	emcomissÃ£o de SUBGERENTE DADÃ□VIDA PÃšBLICA, QC'	*02/05/2003 00:00:00*	Serra	ES
GLENIR GONÃ‡ALVESLOPES	emcomissÃ£o de Agente de ServiÃ§osII, Ref'	*02/05/2003 00:00:00*	Serra	ES
MARLIDINIZ ALVES	em comissÃ£o de Diretorde Unidade, Ref'	*14/05/2003 00:00:00*	Serra	ES
SIMONECARVALHO TRANCOSO MODOLO	de provimentoem comissÃ£o de Assessor EspecialNÃvel IV	*19/05/2003 00:00:00*	Serra	ES
ERIKA SEIBEL PINTO	em comissÃ£o de Chefe deCentro, Ref'	*21/05/2003 00:00:00*	Serra	ES
ELIZETE DE ARAUJO BELO	em comissÃ£o deAssistente de CoordenaÃ§Ã£o ÂŒ QC-05	*03/06/2003 00:00:00*	Serra	ES
BRUNA MARIAZAMPROGNO MADEIRA	em comissÃ£o deCoordenador de DireitoConstitucional eA	*03/06/2003 00:00:00*	Serra	ES
CRISTIANOLEONARDO SANTOS	em comissÃ£o deAgente de ServiÃ§o II, Ref'	*11/06/2003 00:00:00*	Serra	ES
IRENE LÃ‰IA BOSSOIS	de provimento emcomissÃ£o de Assessor Especial NÃvelII	*12/06/2003 00:00:00*	Serra	ES

Figura 14. Fonte de dados consolidada no Power BI, após junção (UNION) de cada cidade.

O dashboard principal no Power BI exibe as quantias de funcionários nomeados por ano e por cidade, cargos distintos com nomeação e cargos com maior quantidade de nomeações e visualização por cidade (mapa). Há filtros por trimestre e cidade e possibilidade de delimitar um período de tempo.

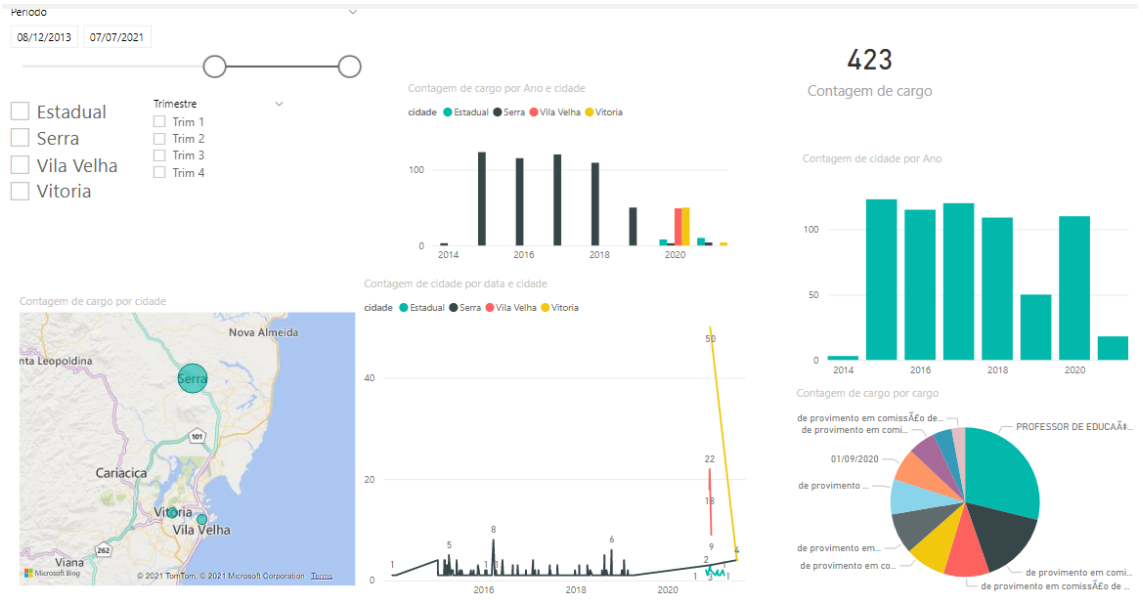


Figura 15. Dashboard principal

No dashboard Timeline (figura X), têm-se uma melhor idéia do acréscimo ou decréscimo a cada ano.

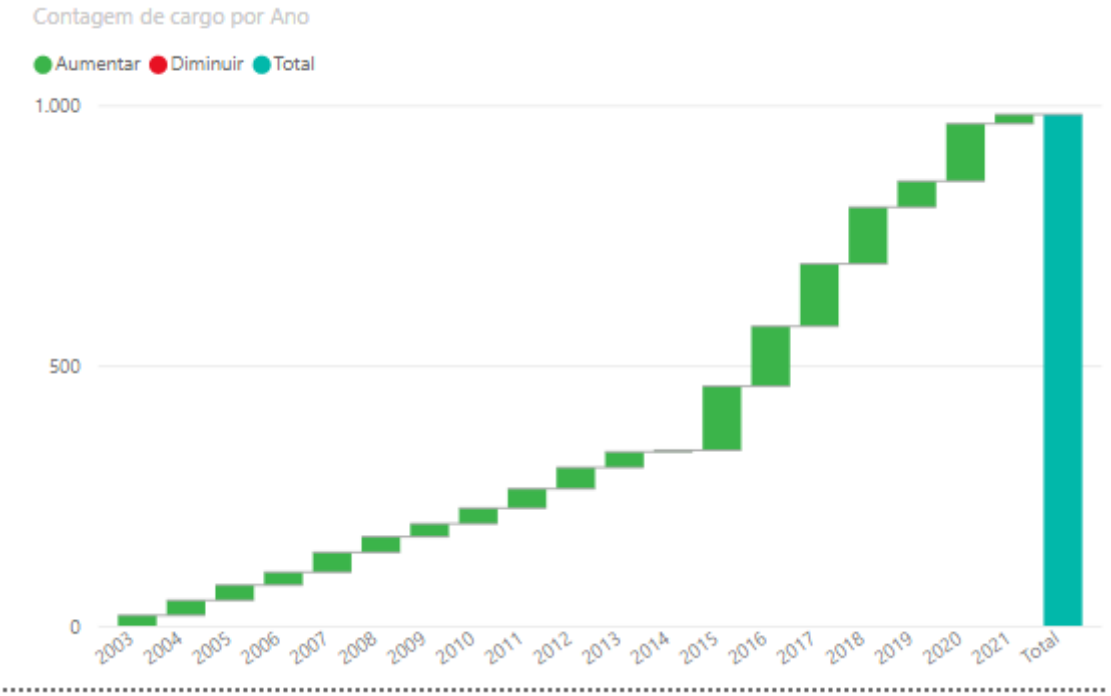


Figura 16. Dashboard Timeline

Com o dashboard Wordcloud - cargos, pode-se ter uma melhor ideia dos termos mais comuns nos cargos com maior número de nomeados. Pode-se fazer uma lista de exclusão.

Figura 17. Dashboard Wordcloud - cargo

4. Conclusão

A estratégia se mostrou funcional, porém a massa de dados utilizada na prova de conceito não é tão grande. O ponto forte da abordagem é que pode ser utilizada em outros domínios. Com a abordagem, o motor de crawling que realiza download dos arquivos e estrutura os dados é a mesma, sendo as expressões regulares e o tratamento de dados final as únicas etapas que possuem peculiaridades em cada fonte.

5. Trabalhos futuros

Entre as oportunidades de melhorias encontradas, estão:

- Atualização incremental de cada fonte, via pentaho, por exemplo. Atualmente está se fazendo um loop com wget para obtenção de todos os arquivos.
- Processo de notificação por email
- Melhorias no tratamento dos nomes e cargos. Foram usadas duas APIs de leitura de pdf, com resultados distintos. Por vezes não era identificado corretamente quebras de linha e espaço em branco, comprometendo parte da massa de dados.
- As expressões regulares podem ser alteradas.
- O volume de dados limitado (mil tuplas) faz que os gráficos por mês, por ex, não sejam tão relevantes, mas pode-se perceber o potencial com maior volume de dados.
- Pode-se estender para fontes de outros tipos, como HTML através do BeautifulSoup, ou documentos do Microsoft Word.

Referências

https://ioes.dio.es.gov.br/diariodaserra

https://www.vilavelha.es.leg.br/institucional/diario-oficial

https://diariooficial.vitoria.es.gov.br

https://ioes.dio.es.gov.br/portal

https://pypi.org/project/fitz

https://pypi.org/project/PyPDF2