

Classificação de Documentos da SEFAZ

Integrantes:

- ALEXEI ALVES DE SOUZA - 398611
- GUSTAVO BEZERRA FECHINE - 397269
- NATANAEL MOREIRA DE LEMOS - 398447
- RAMIRO CAMPOS DE CASTRO - 400723




Problema a ser resolvido:








Classificar os documentos da SEFAZ , de acordo com o seu tipo.

O tipo do documento é de acordo com seu nome de arquivo, podendo ser uma Lei, um Ato Declaratório, um Decreto, ou outro tipo.

O Problema Inicial - ALFRESCO

Alfresco Explorer

 **Navegue nas Pastas**

Tipo	Nome ▲	Descrição	Download
	01 - C.F. de 1988 e CTN de 1966	Constituição Federal e Código Tributário Nacional.	
	02 - LEIS COMPLEMENTARES FEDERAIS	Leis Complementares que estabelecem regras gerais dos impostos estaduais e matérias afins.	
	03 - LEIS COMPLEMENTARES ESTADUAIS	Leis complementares relacionadas aos tributos de competência do Estado do Ceará.	
	04 - LEIS (ICMS, IPVA, ITCD, TAXAS E CM)	Leis que instituem os respectivos tributos, de competência do Estado do Ceará.	
	05 - LEIS ORDINÁRIAS	Outras leis que tratam de matéria tributária, no âmbito do Estado do Ceará.	
	06 - DECRETOS (ICMS, ITCD, IPVA, TAXAS E CM)	Decretos nºs 24.569/1997, 22.311/1992, 31.645/2014, 31.859/2015 e 32.082/2016.	
	07 - DECRETOS	Outros decretos que tratam de matéria tributária, no âmbito do Estado do Ceará.	



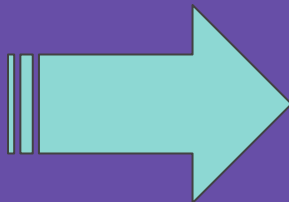
Web Scraping

- Selenium
 - Acessar as pastas
 - Gerar arquivos com os links de download
- Wget
 - Downloads

Agrupamento dos Documentos

Pastas da SEFAZ

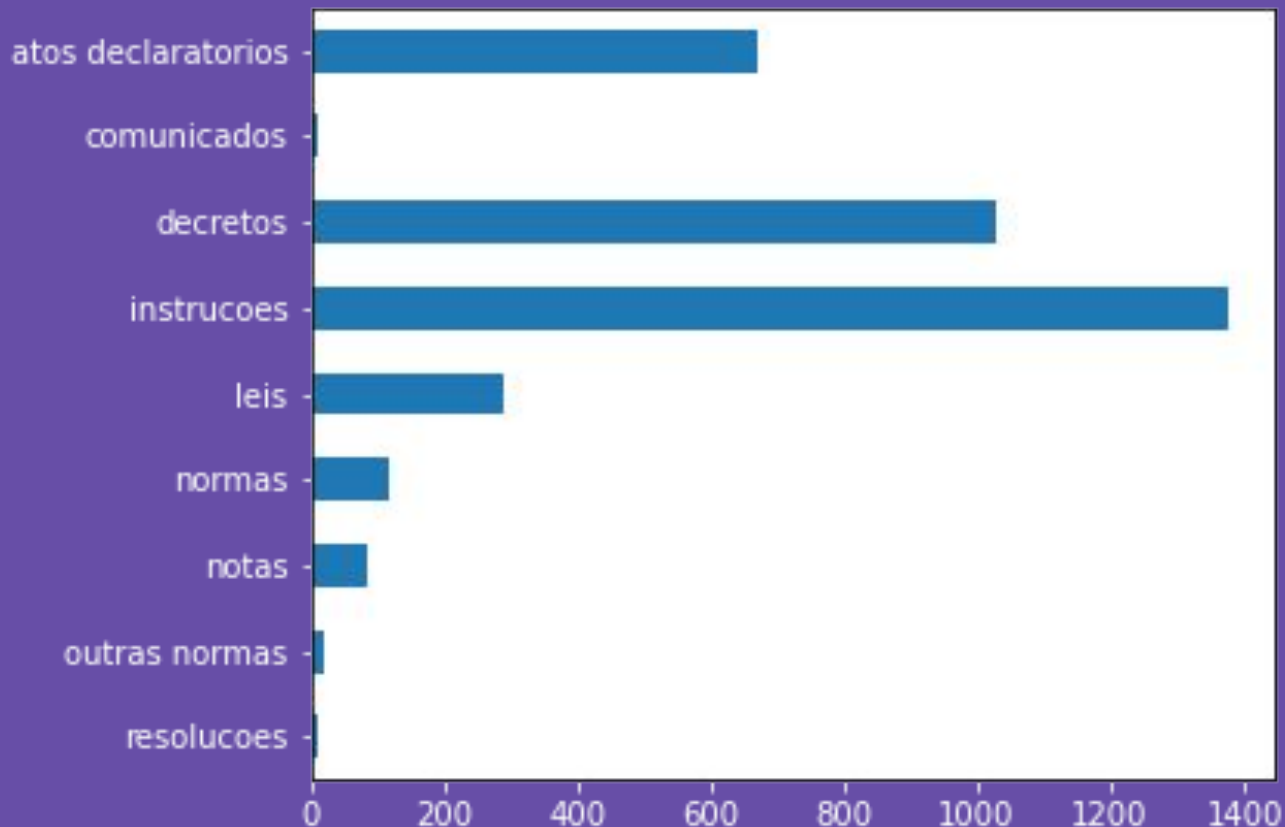
- 01 - C.F. de 1988 e CTN de 1966
- 02 - LEIS COMPLEMENTARES FEDERAIS
- 03 - LEIS COMPLEMENTARES ESTADUAIS
- 04 - LEIS (ICMS, IPVA, ITCD, TAXAS E CM)
- 05 - LEIS ORDINÁRIAS
- 06 - DECRETOS (ICMS, ITCD, IPVA, TAXAS E CM)
- 07 - DECRETOS
- 08 - INSTRUÇÕES NORMATIVAS
- 09 - NOTAS EXPLICATIVAS
- 10 - NORMAS DE EXECUÇÃO
- 11 - ATOS DECLARATÓRIOS
- 12 - RESOLUÇÕES
- 13 - OUTRAS NORMAS
- 14 - COMUNICADOS PÚBLICOS



Nosso Agrupamento

- Atos Declaratórios
- Comunicados
- Decretos
- Instruções
- Leis
- Normas
- Notas
- Outras Normas
- Resoluções

Distribuição dos Documentos





Pré-Processamento

- pdfPlumber
- nltk
 - stopwords
 - word tokenize
- string
- regex
- multiprocessing

Pré-Processamento

\n* Publicado no DOE em 01/02/2013\nATO DECLARATÓRIO Nº 01 /2013\nO SECRETÁRIO DA FAZENDA DO ESTADO DO CEARÁ, no uso de suas atribuições \nlegais.\nRESOLVE:\n1. Revogar, a pedido do contribuinte, o Termo de Acordo FDI/PCDM nº 464, de 14 de \nagosto de 2008, celebrado com a empresa TECNO INDÚSTRIA E COMÉRCIO DE \nCOMPUTADORES LTDA, inscrita no CGF sob o nº 06.358.939-7.\n2. Este Ato Declaratório gera efeito a partir de 1º de fevereiro de 2013\n3. Publique-se. Cumpra-se. Dê-se ciência à interessada.\nSECRETARIO DA FAZENDA DO ESTADO DO CEARÁ, aos 28 de janeiro de 2013\nCarlos Mauro Benevides Filho\n SECRETÁRIO DA FAZENDA



publicado doe ato declaratório secretário fazenda estado ceará uso atribuições legais resolve revogar pedido
contribuinte termo acordo fdi pcdm outubro celebrado empresa tecno indústria comércio computadores ltda inscrita cgf
sob ato declaratório gera efeito partir fevereiro publique-se cumpra-se dê-se ciência interessada secretario fazenda
estado ceará janeiro carlos mauro benevides filho secretário fazenda



Tratamento dos Documentos

Para processar os documentos, foram utilizados 2 métodos:

- CountVectorizer
- TF-IDF



Modelos para serem testados

Para tentar classificar os documentos adquiridos, tentamos os seguintes modelos de classificação:

Linear:

- Regressão Logística
- SVM Linear

Não-Linear:

- Naive Bayes
- SVM Não-linear
- KNN
- Rede Neural

Resultados achados

Utilizamos a Validação Cruzada com 5-Fold para avaliar a acurácia do Treino dos modelos, e obtivemos os seguintes resultados:

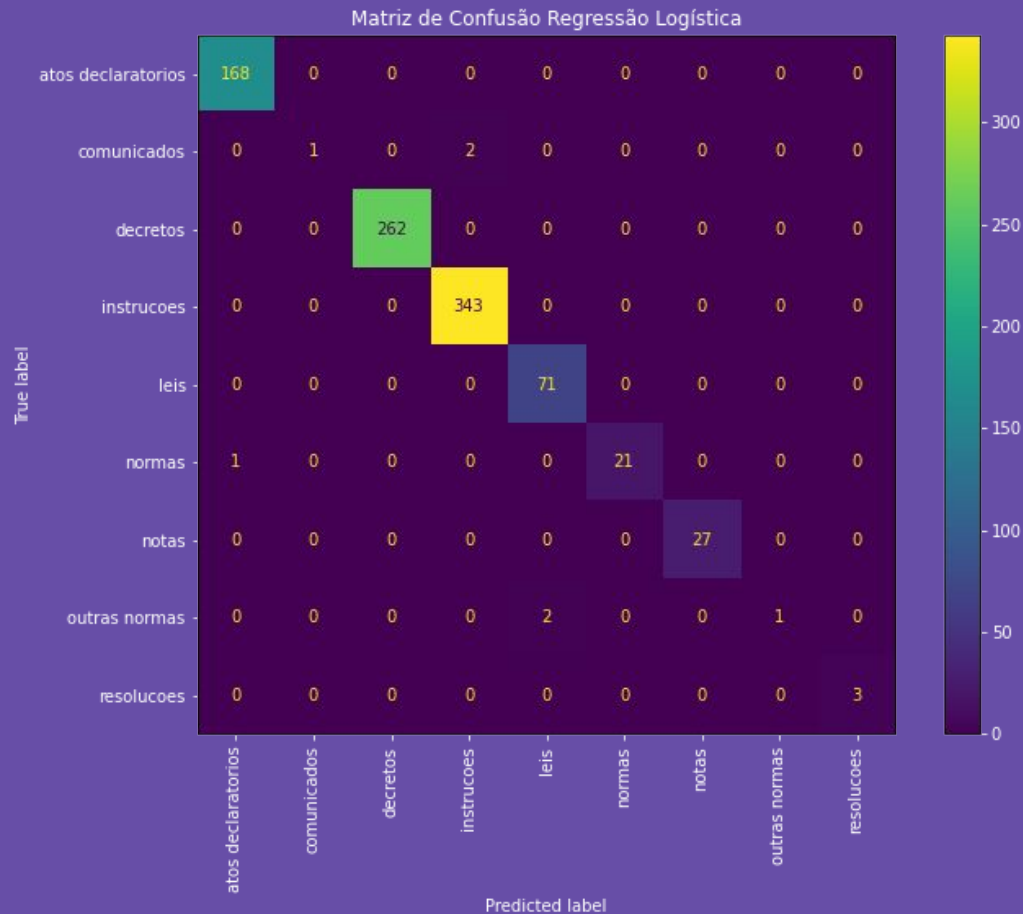
	CountVectorizer	TF-IDF
Naive Bayes	94.3%	90.6%
Regressão Logística	99.0%	98.3%
SVM-Sigmoide	98.2%	97.8%
SVM-Linear	98.9%	99.1%
KNN	90.4%	85.7%
Rede Neural	98.1%	37.7%



Métricas para avaliação do Modelo

- Precisão
- Recall
- F1-score
- Acurácia

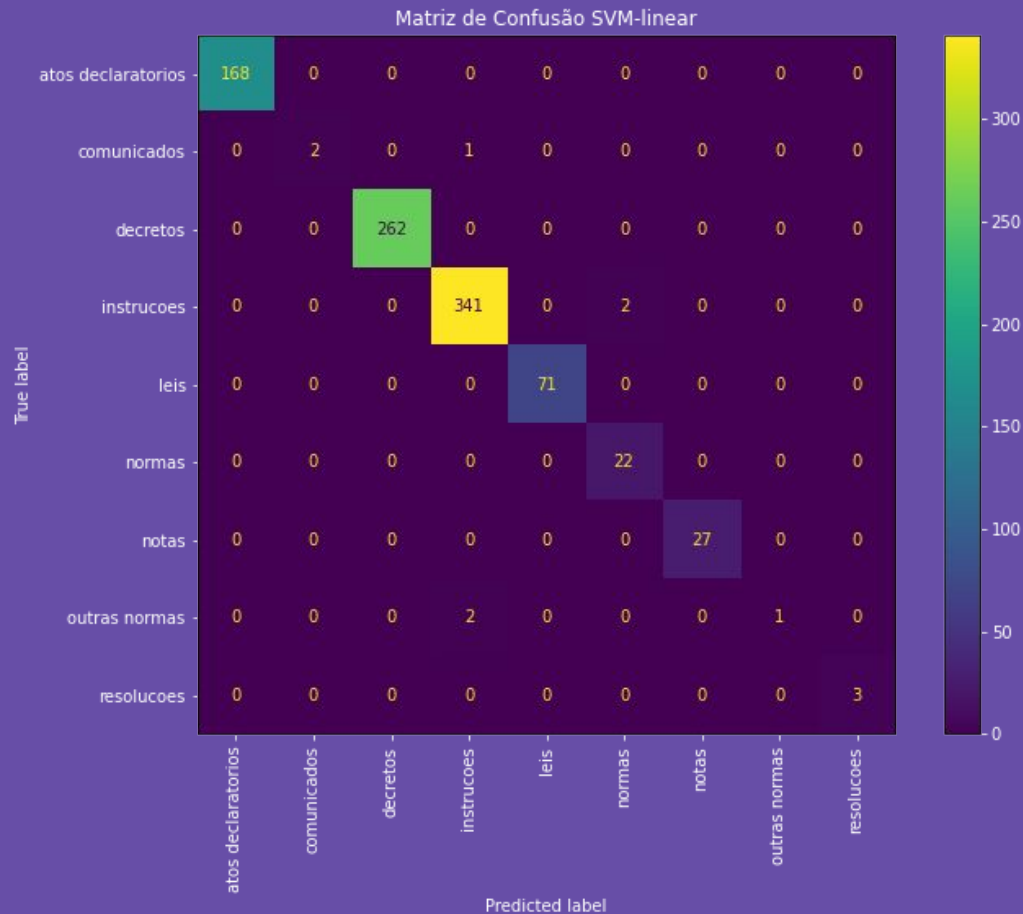
O melhor modelo CountVectorizer em prática



Avaliação da Regressão Logística

	Precisão	Recall	F1-Score	Quantidade
Atos Declaratórios	99%	100%	100%	168
Comunicados	100%	33%	50%	3
Decretos	100%	100%	100%	262
Instruções	99%	100%	100%	343
Leis	97%	100%	99%	71
Normas	100%	95%	98%	22
Notas	100%	100%	100%	27
Outras Normas	100%	33%	50%	3
Resoluções	100%	100%	100%	3

O melhor modelo TF-IDF em prática



Avaliação do SVM-Linear

	Precisão	Recall	F1-Score	Quantidade
Atos Declaratórios	100%	100%	100%	168
Comunicados	100%	67%	80%	3
Decretos	100%	100%	100%	262
Instruções	99%	100%	100%	343
Leis	99%	100%	99%	71
Normas	100%	100%	100%	22
Notas	100%	93%	96%	27
Outras Normas	50%	33%	40%	3
Resoluções	100%	100%	100%	3

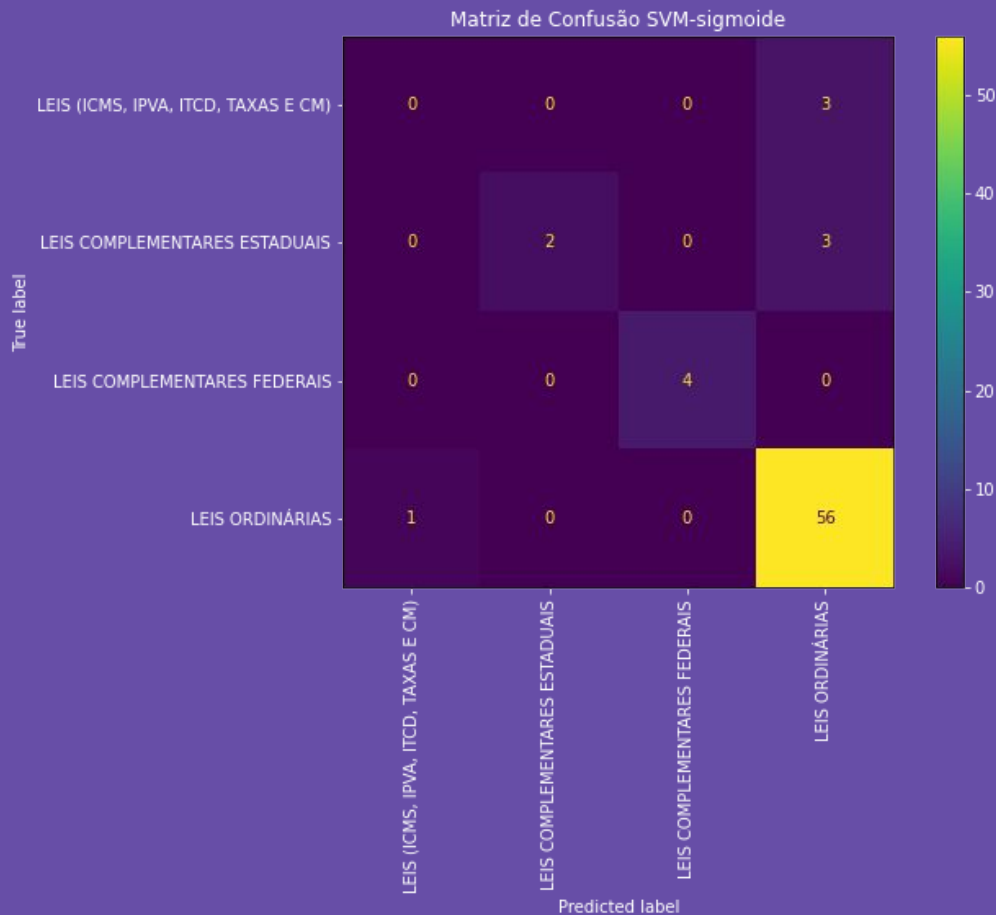


Toques Finais

Tendo classificado com sucesso os dados nos nossos agrupamentos, falta agrupar os dados para as pastas da SEFAZ. Para isso, temos de separar os tipos de Leis, e os Decretos.

Repetimos os mesmos passos anteriores, só que em 2 datasets diferentes, um contendo as Leis e suas pastas da SEFAZ e o outro com os decretos.

O melhor modelo para separação das Leis



Obrigado por assistir!

