Projeto Final
Ciência de Dados

Membros: Ricardo Cavalcante - 377744
Tales Araujo - 374953
Lucas S. Fernandes - 486265
Rômulo Férrer Filho - 385218
João Pedro B. Andrade - 486454

Motivações - O problema

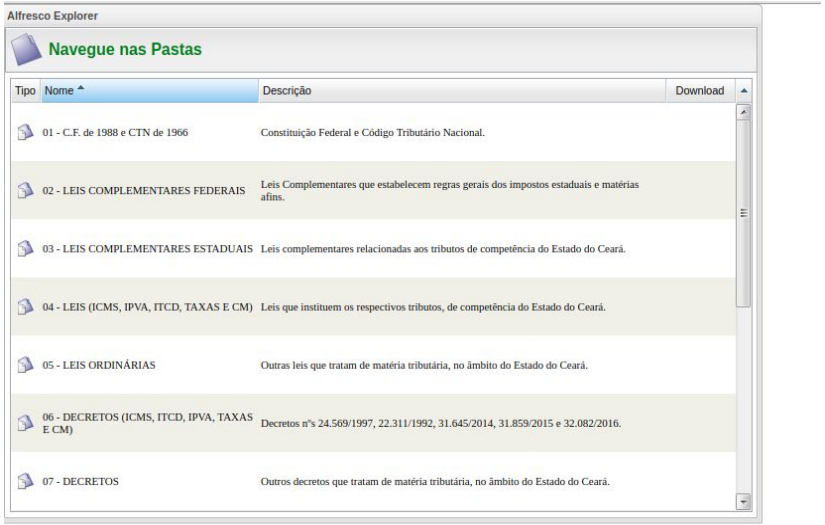
- Classificar documentos de legislação da SEFAZ, disponíveis no portal Sefaz Legis, de acordo como seu tipo: Lei, Decreto,Instrução Normativa, etc.
  - o título do documento contém o tipo documental e deverá ser usado como label;
  - os trabalhos serão avaliados pela acurácia do modelo e pelo tratamento dado aos dados, na solução.

Aquisição dos dados

- A fonte dos dados é feita com uma ferramenta chamada alfresco, que dificultou a navegação automatizada.
- 160 pastas 14 externas e 146 internas
- Script para recuperar itens das pastas:

```
let links = document.getElementsByTagName('a');

for(let i=0; i < links.length; i++) {
    links[i].dispatchEvent(new MouseEvent("click"))
}
```



- **3612 arquivos**

Pré-processamento

- Extração do conteúdo dos PDFs
  - Py2PDF
  - Slate
  - **PdfPlumber**
- Erros de Encoding
- Filtros de classe:
  - Ato Declaratório
  - Decreto
  - Instrução Normativa
  - Lei
  - Norma de Execução
  - Nota Explicativa

ESTADO DO CEARÁ
SECRETARIA DA FAZENDA

DECRETO Nº 26.527, DE 7 DE MARÇO DE 2002.

07/03/2002
* Publicado no DOE em 08/03/2002.

Prorroga o prazo de pagamento do ICMS de que trata o art. 2º, inciso I, alínea *a*, do Decreto nº 26.483, de 28/12/2001.

O GOVERNADOR DO ESTADO DO CEARÁ, no uso das atribuições que lhe conferem os incisos IV e VI do art. 88 da Constituição Estadual, e

Considerando a necessidade de prorrogar o prazo de pagamento do ICMS, tendo em vista problema decorrente do sistema de informática da rede bancária estadual,

DECRETA:

Art. 1º O prazo de recolhimento do ICMS de que trata o art. 2º, inciso I, alínea *a*, do Decreto nº 26.483, de 28/12/2001, com vencimento previsto para o dia 28 de fevereiro de 2002, fica, excepcionalmente, prorrogado para o dia 1º de março de 2002.

Art. 2º Este Decreto entra em vigor na data de sua publicação.

PALÁCIO DO GOVERNO DO ESTADO DO CEARÁ, em Fortaleza, aos 7 de março de 2002.

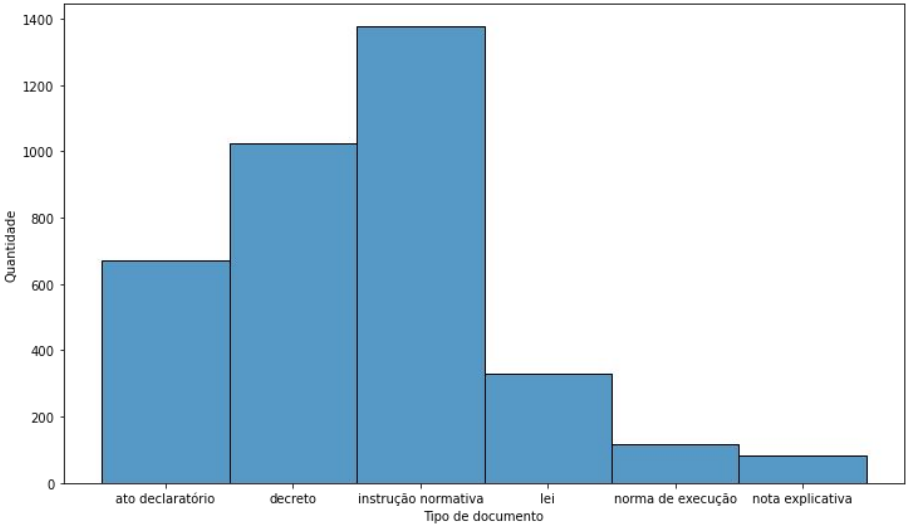
Tasso Ribeiro Jereissati
GOVERNADOR DO ESTADO

Ednilton Gomes de Soárez
SECRETÁRIO DA FAZENDA

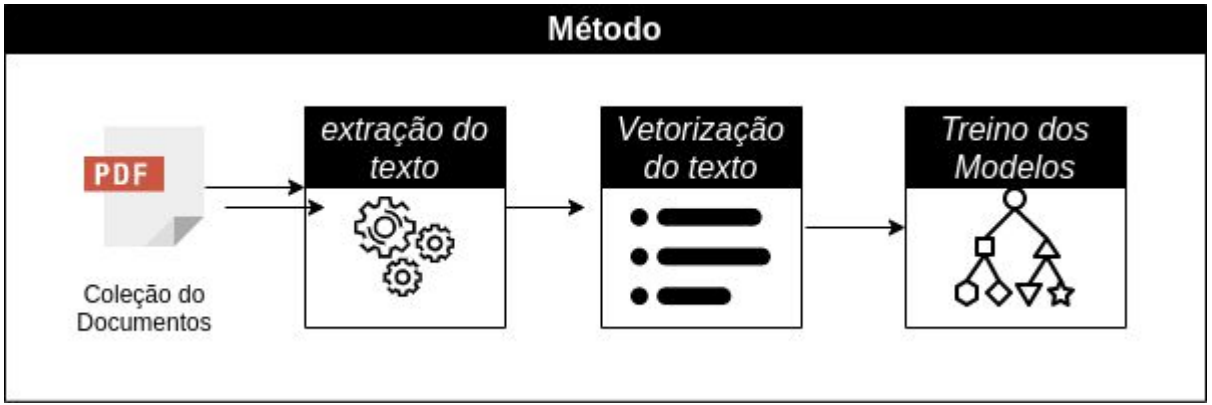
Análise dos dados

Questões encontradas:

- Dados faltantes
- Ambiguidades nas *labels*
- Conjuntos não-disjuntos de classes
- Seleção de classes para construção do modelo



Solução



O conjunto de métodos abaixo foi aplicado para o dataset e os em destaque obtiveram melhor resultado, portanto foram escolhidos como solução principal:

- **Count Vectorizer**
- TF-IDF Vectorizer

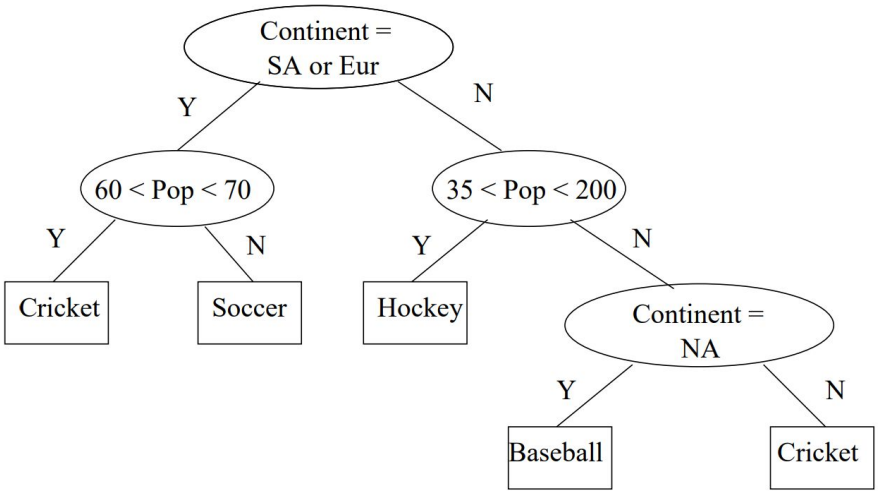
- **Decision Tree**
- Naive Bayes
- Random Forest

Count-Vectorizer

- Tokenização de palavras via *Bag-of-Words* em coleção de documentos
- Simplificação de sentenças por remoção de termos prolixos
- Retorna vetor codificado contendo:
  - Comprimento do Vocabulário
  - Contagem de cada palavra

Decision Tree

- Utiliza propriedades do vetor de atributos para decidir a classe da entrada.
- Representada em forma de árvore, onde cada nó é um teste e as folhas são as decisões.



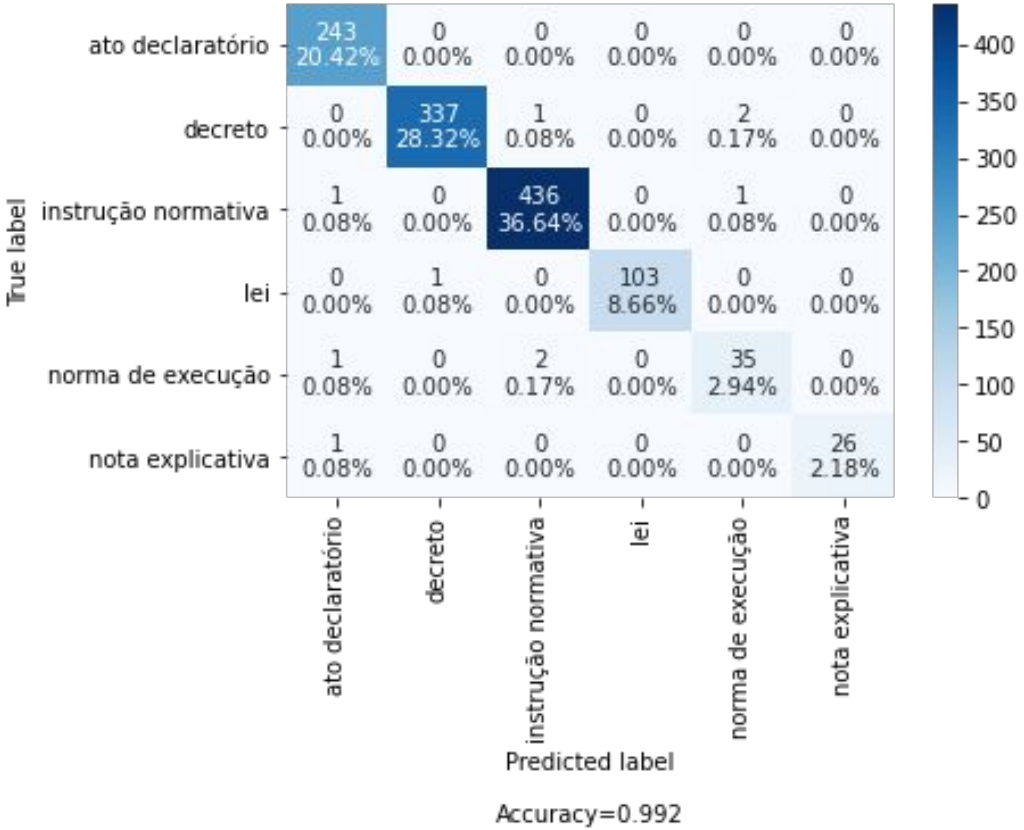
Exemplo de DT para esporte favorito de um país. Fonte: MMDS (Leskovec *et. al.*)

Resultados

	precision	recall	f1-score	support
ato declaratório	0.99	1.00	0.99	243
decreto	1.00	0.99	0.99	340
instrução normativa	0.99	1.00	0.99	438
lei	1.00	0.99	1.00	104
norma de execução	0.92	0.92	0.92	38
nota explicativa	1.00	0.96	0.98	27

accuracy			0.99	1190
macro avg	0.98	0.98	0.98	1190
weighted avg	0.99	0.99	0.99	1190

Resultados


ato declaratório
decreto
instrução normativa
lei
norma de execução
nota explicativa
True label
243 20.42%
0 0.00%
0 0.00%
0 0.00%
0 0.00%
0 0.00%
0 0.00%
337 28.32%
1 0.08%
0 0.00%
2 0.17%
0 0.00%
1 0.08%
0 0.00%
436 36.64%
0 0.00%
1 0.08%
0 0.00%
0 0.00%
1 0.08%
0 0.00%
103 8.66%
0 0.00%
0 0.00%
1 0.08%
0 0.00%
2 0.17%
0 0.00%
35 2.94%
0 0.00%
1 0.08%
0 0.00%
0 0.00%
0 0.00%
0 0.00%
26 2.18%
ato declaratório
decreto
instrução normativa
lei
norma de execução
nota explicativa
Predicted label
Accuracy=0.992
400
350
300
250
200
150
100
50
0

Conclusão

- A combinação ***CountVectorizer*** e ***DecisionTree*** demonstrou excelentes resultados
- Possíveis próximo passos:
  - Validação cruzada
  - Tratar o desbalanceamento dos dados (resampling?)
  - Testar outras outras técnicas de vetorização de texto e classificadores

Dúvidas?

Obrigado pela atenção!

Repositório do Trabalho:

Github - Trabalho Sefaz