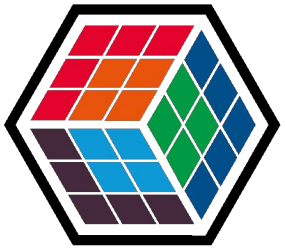


THE DEVELOPER'S CONFERENCE

ML - Aplicações de Machine Learning na área Jurídica, um exemplo com classificação de textos

Juliano Pacheco

Arquiteto de Software

Sobre

THE DEVELOPER'S CONFERENCE

- Arquiteto de Software (Cast Group)
- Java / Python
- Nerd / Games
- Time de inovação (TJRS)
- Curioso da área de ML

Tópicos

THE DEVELOPER'S CONFERENCE

- Contexto Jurídico
- Projetos
- Ementas (exemplo)
- Possibilidades de aplicação

Contexto Juridico BR

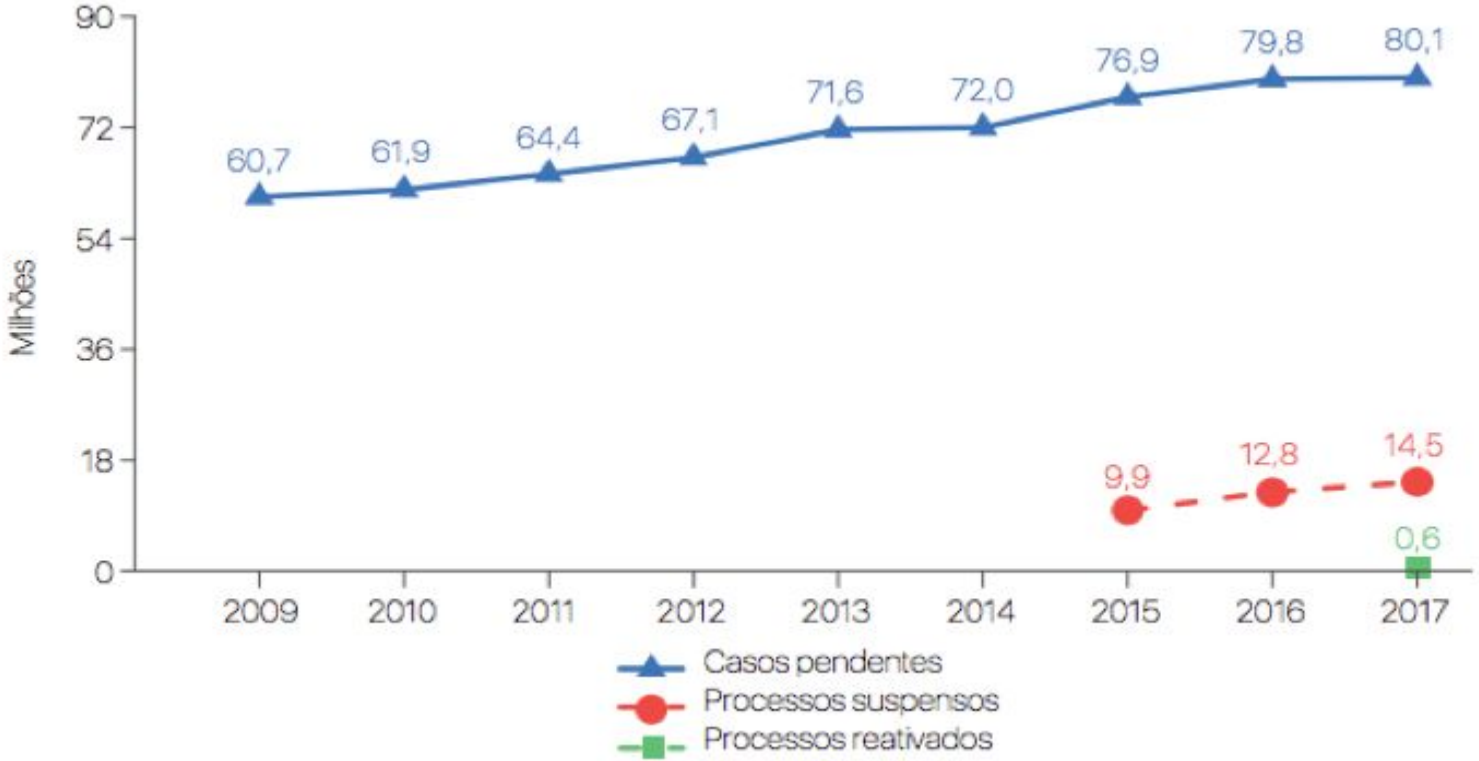
THE DEVELOPER'S CONFERENCE

Em 2017, cada juiz brasileiro julgou em média, 1819 processos, o que equivale a 7,2 casos por dia útil.

O ano de 2017 teve um crescimento de estoque de 80,1 milhões de processos que aguardam uma solução definitiva, o que significa um aumento de 244 mil casos pendentes em relação a 2016

Contexto Juridico BR

THE DEVELOPER'S CONFERENCE



Contexto Jurídico BR

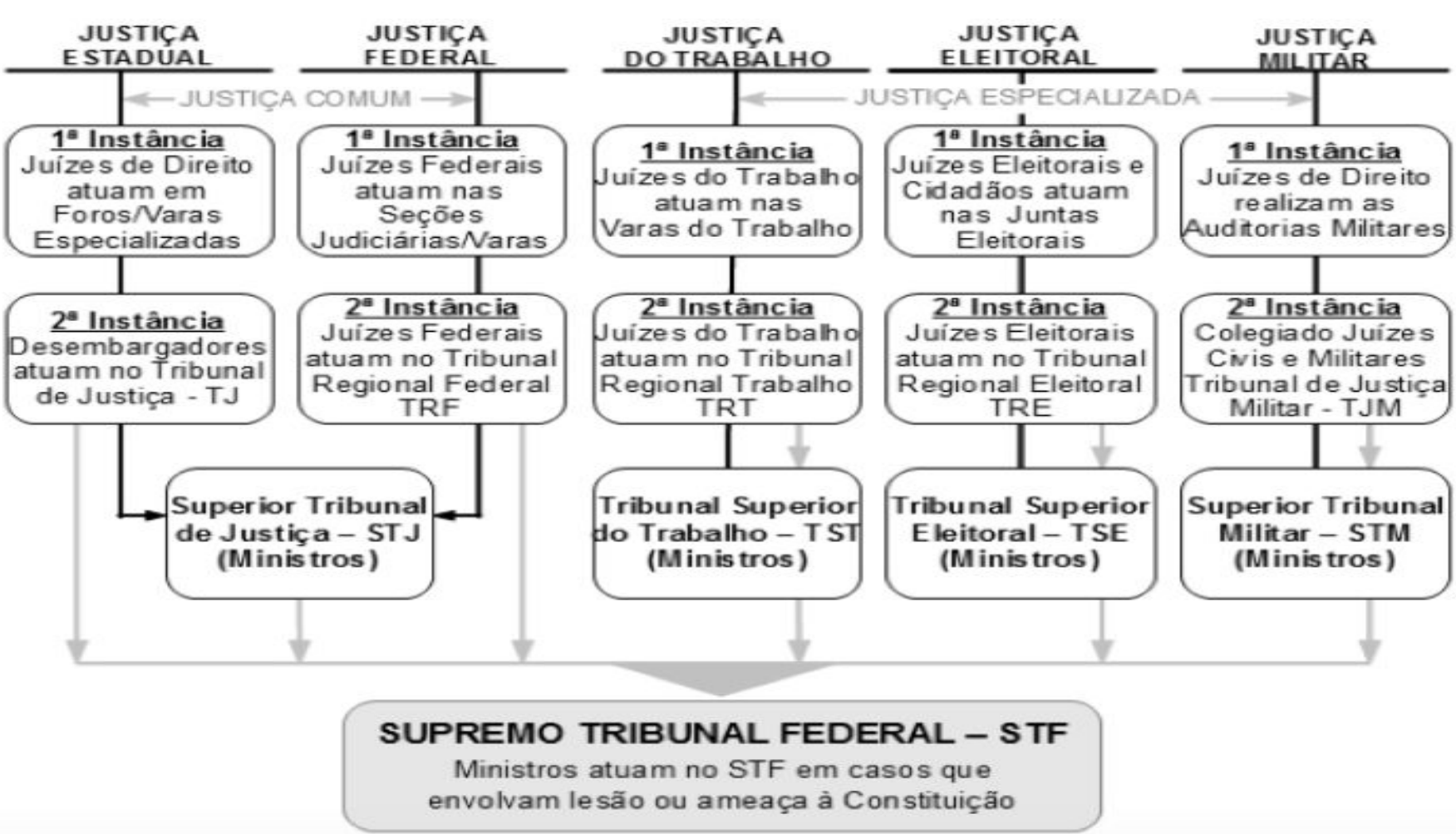
THE DEVELOPER'S CONFERENCE

A justiça estadual concentra a maior parte do estoque de processos, 63.482 milhões, o equivalente a 79% dos processos pendentes. A justiça federal concentra 12,79% dos processos, e a justiça trabalhista, 6, 9%. Os demais segmentos da justiça, acumulam 1% dos casos pendentes.

Resposta Poder Judiciário

THE DEVELOPER'S CONFERENCE

- Investimentos em tecnologia
- Informatização e automatização de procedimentos manuais
- Investimento em aplicações de IA
- Lab de inovação e IA aplicada ao PJe (principal sistema da justiça), através da portaria n 25/2019


JUSTIÇA ESTADUAL
JUSTIÇA FEDERAL
JUSTIÇA DO TRABALHO
JUSTIÇA ELEITORAL
JUSTIÇA MILITAR
JUSTIÇA COMUM
JUSTIÇA ESPECIALIZADA
1ª Instância
Juízes de Direito atuam em Foros/Varas Especializadas
1ª Instância
Juízes Federais atuam nas Seções Judiciárias/Varas
1ª Instância
Juízes do Trabalho atuam nas Varas do Trabalho
1ª Instância
Juízes Eleitorais e Cidadãos atuam nas Juntas Eleitorais
1ª Instância
Juízes de Direito realizam as Auditorias Militares
2ª Instância
Desembargadores atuam no Tribunal de Justiça - TJ
2ª Instância
Juízes Federais atuam no Tribunal Regional Federal TRF
2ª Instância
Juízes do Trabalho atuam no Tribunal Regional Trabalho TRT
2ª Instância
Juízes Eleitorais atuam no Tribunal Regional Eleitoral TRE
2ª Instância
Colegiado Juízes Civis e Militares Tribunal de Justiça Militar - TJM
Superior Tribunal de Justiça – STJ (Ministros)
Tribunal Superior do Trabalho – TST (Ministros)
Tribunal Superior Eleitoral – TSE (Ministros)
Superior Tribunal Militar – STM (Ministros)
SUPREMO TRIBUNAL FEDERAL – STF
Ministros atuam no STF em casos que envolvam lesão ou ameaça à Constituição

TJRS (em 2019)

THE DEVELOPER'S CONFERENCE

- 788 magistrados
- 7948 servidores
- 4393 estagiários
- total de 13129 colabores

TJRS

THE DEVELOPER'S CONFERENCE

Atualmente em levantamento efetuado pelo TJRS, existem 3.665.371 processos em andamento, sendo julgados.

Cerca de 1.639 processos sendo julgados por magistrados de 1º grau e 3.698 processos por desembargadores.

Projeto Victor (STF / UnB)

THE DEVELOPER'S CONFERENCE

- Separar e classificar as peças processuais mais utilizadas nas atividades do STF
  - Realiza esta atividade em 5 segundos, antes os servidores precisavam de 30 minutos.
- Avaliação do enquadramento dos recursos em relação aos principais temas de repercussão geral fixados pelo tribunal.

Projeto Elis (TJPE)

- Analisar executivos fiscais do município de Recife
- Identifica divergências cadastrais, prescrições e competências.
- Em próxima etapa, pretende-se inserir minutas automaticamente e assinar despachos
- Documentos analisados: certidão de dívida ativa e petição inicial

Projeto Elis (TJPE)

THE DEVELOPER'S CONFERENCE

Avaliou em apenas 3 dias (ambiente de homologação) 5.247 processos e conseguiu classificar com precisão a competência das ações, divergências cadastrais, erros no cadastro de dívida ativa e casos de prescrição.

Poderá reduzir o trabalho de 18 meses para 15 dias, analisando 80 mil movimentações processuais, com maior acuracia.

Projeto Alice (TCU)

- Caçar fraudes e irregularidades em licitações
- Em média 200 editais por dia

Classificação de textos

- Classificar um conjunto textos (documentos) com o intuito de descobrir uma classe alvo, em nosso caso vamos classificar ementas para descobrir se uma ementa pertence ao tipo de seção civil ou crime.

Grupo IA e TJRS Inovação

- Grupo com estudantes da Unisinos (professores, estudantes graduação/mestrado/doutorado/pos, outros), area de tecnologia e direito
- Setor de Inovação (Juliano Pacheco + 2 colegas)

Corpus (definição)

Conjunto de ementas coletadas no site do TJRS

tjrs.jus.br/site/busca-solr/index.html?aba=jurisprudencia

Site | Diário da Justiça Eletrônico | Publicações Administrativas | Legislação | **Jurisprudência** | Consulta Processual

Palavra-chave

BUSCAR

> Ajuda
> Instruções importantes

○ Inteiro teor ● Ementa

Procurar resultados

Com a expressão:

Com **qualquer uma** das palavras:

Sem as palavras:

Filtrar resultados por: limpar filtros

Tribunal: Todos

Órgão julgador: Todos

Relator/Redator: Todos

Tipo de Processo: Todos

Classe CNJ: Todos

Referência Legislativa:

Comarca de Origem:

Assunto CNJ: Todos

Jurisprudência:

Assunto:

Data de Julgamento:

00/00/0000 até 00/00/0000

Data de Publicação:

00/00/0000 até 00/00/0000

Número do Processo: Seção: ☐ Cível ☐ Crime

Tipo de Decisão:

☐ Acórdão ☐ Monocrática ☐ Admissibilidade

Dúvida de Competência ☐

Resultados **1** a **10** de aproximadamente **8188510** . A pesquisa demorou **1.262** segundos.

Classificar por data decrescente Classificar por data crescente

Filtros mais frequentes

Órgão Julgador

Comarca de Origem

Redator/Relator

Ano de Julgamento

Classe CNJ

Assunto CNJ

Tribunal

Tipo Processo

Data Publicação

1. Núm.:70083102335 **Inteiro teor:** doc html

Tipo de processo: Agravo de Instrumento

Tribunal: Tribunal de Justiça do RS

Classe CNJ: Agravo de Instrumento

Relator: Laura Louzada Jaccottet

Órgão Julgador: Segunda Câmara Cível

Comarca de Origem: ESTEIO

Seção: CIVEL

Assunto CNJ: ISS/ Imposto sobre Serviços

Decisão: Monocratica

Ementa: AGRAVO DE INSTRUMENTO. DIREITO TRIBUTÁRIO. EXECUÇÃO FISCAL. CITAÇÃO POR EDITAL. NECESSIDADE DE ESGOTAR AS POSSIBILIDADES DE CITAÇÃO EM TODOS OS ENDEREÇOS CONSTANTES DOS AUTOS. 1. O Superior Tribunal de Justiça, em recurso representativo de controvérsia, pacificou o entendimento de que, segundo o art. 8º da Lei 6.830/80, a citação por edital, na execução fiscal, somente é cabível quando não exitosas as outras modalidades ali previstas, a saber, a citação via Correios e por Oficial de Justiça. Inteligência da Súmula n. 414. Demais, conforme entendimento do Tribunal da Cidadania, para que se efetue a citação por edital, é prescindível o esgotamento de meios extrajudiciais disponíveis para a localização do endereço do devedor, pois a norma legal exige tão somente as tentativas frustradas de citação pelos Correios e via mandado. Neste órgão fracionário o entendimento se coaduna com a jurisprudência da Corte Superior. 2. No caso concreto, porém, embora realizada tentativa de citação por AR e mandado em dois dos endereços constantes dos autos, ausente tentativa em um terceiro endereço também constante do processo. Assim, descabe, por ora, a ordem de citação por edital, sendo necessária a tentativa no endereço faltante sob pena de cerceamento de defesa. AGRAVO DE INSTRUMENTO DESPROVIDO, EM DECISÃO MONOCRÁTICA.(Agravo de Instrumento, Nº 70083102335, Segunda Câmara Cível, Tribunal de Justiça do RS, Relator: Laura Louzada Jaccottet, Julgado em: 25-11-2019)[0]

Data de Julgamento: 26-11-2019

Versão para impressão

Corpus (entenda o dado)

THE DEVELOPER'S CONFERENCE

- O que é uma ementa (texto, de exemplo, verbetação e dispositivo)

“APELAÇÃO CÍVEL. PROMESSA DE COMPRA E VENDA. EMBARGOS À PENHORA. NULIDADE DO AUTO DE PENHORA. INOCORRÊNCIA. NULIDADE POR AUSÊNCIA DE INTIMAÇÃO DO CÔNJUGE…"

"Nulidade do auto de penhora. Inobstante o disposto no art. 665, IV, do CPC, atual art. 838, a falta de indicação de do local da penhora e os nomes do credor e do devedor, não invalida a penhora, pois devem ser aplicados, no caso, os princípios da instrumentalidade das formas, do princípio da finalidade..."

Ambiente / Tecnologia

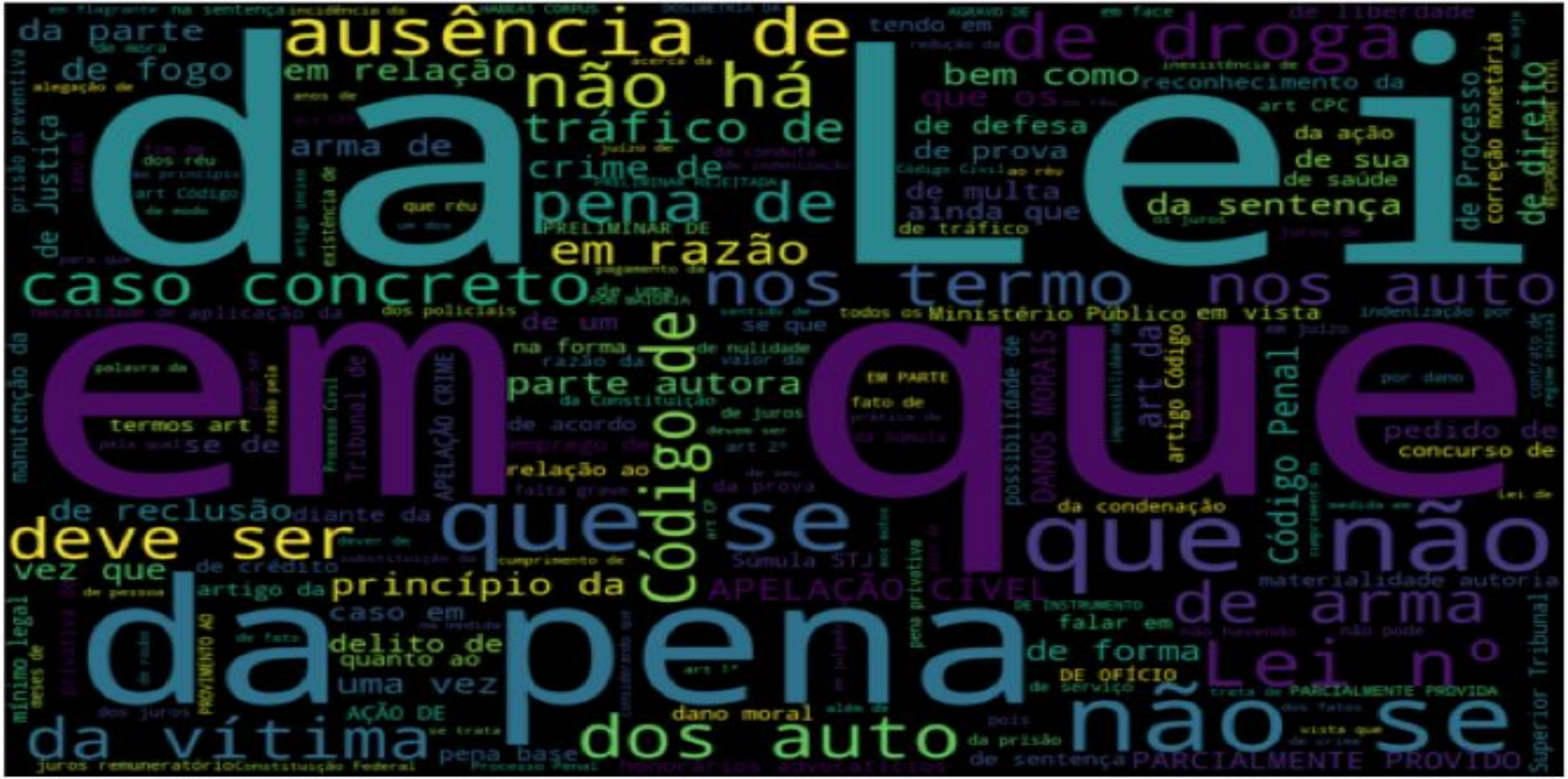
THE DEVELOPER'S CONFERENCE

- Google Colaboratory (Jupyter do google)
- Python
- Scikit-learn (ML)
- NLTK (PLN)

Corpus (analise do dado)

THE DEVELOPER'S CONFERENCE

	EMENTA	CLASSE
5073	RESPONSABILIDADE CIVIL. \rAÇÃO INDENIZATÓRIA. ...	0
3853	HABEAS CORPUS. \r\tESTELIONATO, FALSIDADE IDEO...	1
130	APELAÇÃO CÍVEL. DIREITO PRIVADO NÃO ESPECIFICA...	0
4267	APELAÇÃO CÍVEL. SERVIDOR PÚBLICO. MUNICÍPIO DE...	0
2167	APELAÇÃO CÍVEL. DIREITO PRIVADO NÃO ESPECIFICA...	0


da
Lei
em
que
da
pena
ausência de
de droga
não há
tráfico de
pena de
caso concreto
nos termo
nos auto
que se
que não
Código de
deve ser
de arma
não se
Lei nº
da vítima
dos auto
em razão
crime de
arma de
em relação
de fogo
da parte
bem como
reconhecimento da
de defesa
de prova
de multa
ainda que
de tráfico
da sentença
da ação
de sua
de saúde
art CPC
de Processo
correção monetária
de direito
parte autora
princípio da
de reclusão
vez que
caso em
delito de
quanto ao
uma vez
dano moral
de forma
DE OFÍCIO
falar em
Código Penal
art da
artigo Código
pedido de
concurso de
materialidade autoria
APELAÇÃO CÍVEL
PARCIALMENTE PROVIDO
PARCIALMENTE PROVIDA
Superior Tribunal
Ministério Público
em vista
na forma
de acordo
relação ao
termos art
se de
artigo da
da condenação
Súmula STJ
de crédito
possibilidade de
DANOS MORAIS
Tribunal de
APELAÇÃO CRIME
de Justiça
manutenção da
mínimo legal
prisão preventiva
juros remuneratório
Constituição Federal
pena base
de sentença
da prisão
dos policiais
todos os
se que
de um
PRELIMINAR DE
da Constituição
de juros
fato de
EM PARTE
AÇÃO DE
diante da
da prova
tendo em
de liberdade
em face
dos réu
art Código
que réu

Frequências de palavras

THE DEVELOPER'S CONFERENCE

	Palavra	Frequencia
10	art	20922
232	ser	16988
125	caso	15510
158	pena	15053
201	lei	12949
0	reu	12872
139	autos	10681
24	prova	10585
385	artigo	10557
339	parte	10091
3	sentenca	9324
21	fato	8876
93	sendo	8320
181	crime	8074
1025	penal	8053
386	codigo	8016
46	ha	7637
569	vitima	7479
85	forma	7084
122	pois	7030

Pré-processamento

- Remoção do caput
- Limpeza / Normalização
- Tokenização dos dados
- Remoção de stop words (a, o, da, etc)
- Remoção de atributos do contexto (ex.: ECA)
- Aplicação de stemming e / ou lemma
  - stemm (gato, gata, gatas = gat)
  - lemma (gato, gata, gatas = gato)
  - lemma - masculino singular, verbos no infinitivo

Pré-processamento

THE DEVELOPER'S CONFERENCE

	EMENTA	EMENTA_SEM_VERBATIZACAO	CLASSE
5073	RESPONSABILIDADE CIVIL. \rAÇÃO INDENIZATÓRIA. ...	– – 14, § 1°, Caso em que a autora sofreu impo...	0
3853	HABEAS CORPUS. \r\tESTELIONATO, FALSIDADE IDEO...	1. No caso, inexistem as nulidades alegadas pe...	1
130	APELAÇÃO CÍVEL. DIREITO PRIVADO NÃO ESPECIFICA...	Trata-se de recurso de apelação interposto con...	0
4267	APELAÇÃO CÍVEL. SERVIDOR PÚBLICO. MUNICÍPIO DE...	1. Hipótese em que o comprovado agir equivocad...	0
2167	APELAÇÃO CÍVEL. DIREITO PRIVADO NÃO ESPECIFICA...	1. Gratuidade judiciária: deve ser concedido o...	0

Tokenização

“Trata-se de recurso de apelação interposto contra a extinção de ação de obrigação de fazer cumulada com pedido de indenização por danos materiais.”

“['Trata-se', 'de', 'recurso', 'de', 'apelação', 'interposto', 'contra', 'a', 'extinção', 'de', 'ação', 'de', 'obrigação', 'de', 'fazer', 'cumulada', 'com', 'pedido', 'de', 'indenização', 'por', 'danos', 'materiais','.’]”

Tokenização

- Pontuação
- Espaços
- Customizado
- Outros

Word embeddings

THE DEVELOPER'S CONFERENCE

- Modelo Vetorial de Contagem Simples de Ocorrências
- Modelo Vetorial de Frequência de Termos Normalizada
- Modelo Vetorial de Frequência do Termo pelo Inverso da Frequência nos Documentos

Word embeddings

THE DEVELOPER'S CONFERENCE

```
count_vectorizer = CountVectorizer()
```

```
bag_of_words = count_vectorizer.fit_transform(df['DISPOSITIVOS_PROC'])
```

Algoritmos

- Nayve Bayes
- KNN
- Regressão Linear
- Arvore de Decisão
- Floresta aleatória
- SVM (kernel=Linear)

Algoritmos

THE DEVELOPER'S CONFERENCE

```
#MultinomialNB
nb = MultinomialNB()

#SVM (Linear):
#==============
svm = SVC(kernel='linear', probability=True)

#DecisionTreeClassifier (DT-CART)
#========================
dtre = DecisionTreeClassifier(max_depth=5)

#Random Forest:
#==============
rf = RandomForestClassifier(n_estimators=40)

#Logistic Regression:
#========================
#lr = LogisticRegression(solver='lbfgs')
lr = LogisticRegression(solver='lbfgs', multi_class='auto',)

#KNN ( KNeighborsClassifier )
neigh = KNeighborsClassifier(n_neighbors=5)
```

Treinamento / Teste

THE DEVELOPER'S CONFERENCE

- Holdout
- Cross Validation

Holdout

THE DEVELOPER'S CONFERENCE

```
X_treino, X_teste, y_treino, y_teste = train_test_split(bag_of_words,
                                                        df.CLASSE,
                                                        test_size=0.2,
                                                        random_state = random_state_seed)
```

Métricas

THE DEVELOPER'S CONFERENCE

- Acurácia
  - diz quanto, das possíveis previsões, de fato o modelo acertou
- Revocação (recall)
  - responde a pergunta de qual proporção de positivos foi identificados corretamente?
- Precisão
  - busca responder qual a proporção de identificações positivas foi realmente correta?
- F1 Score
  - esta métrica mostra o balanço entre a precisão e o recall do modelo

Resultados Holdout

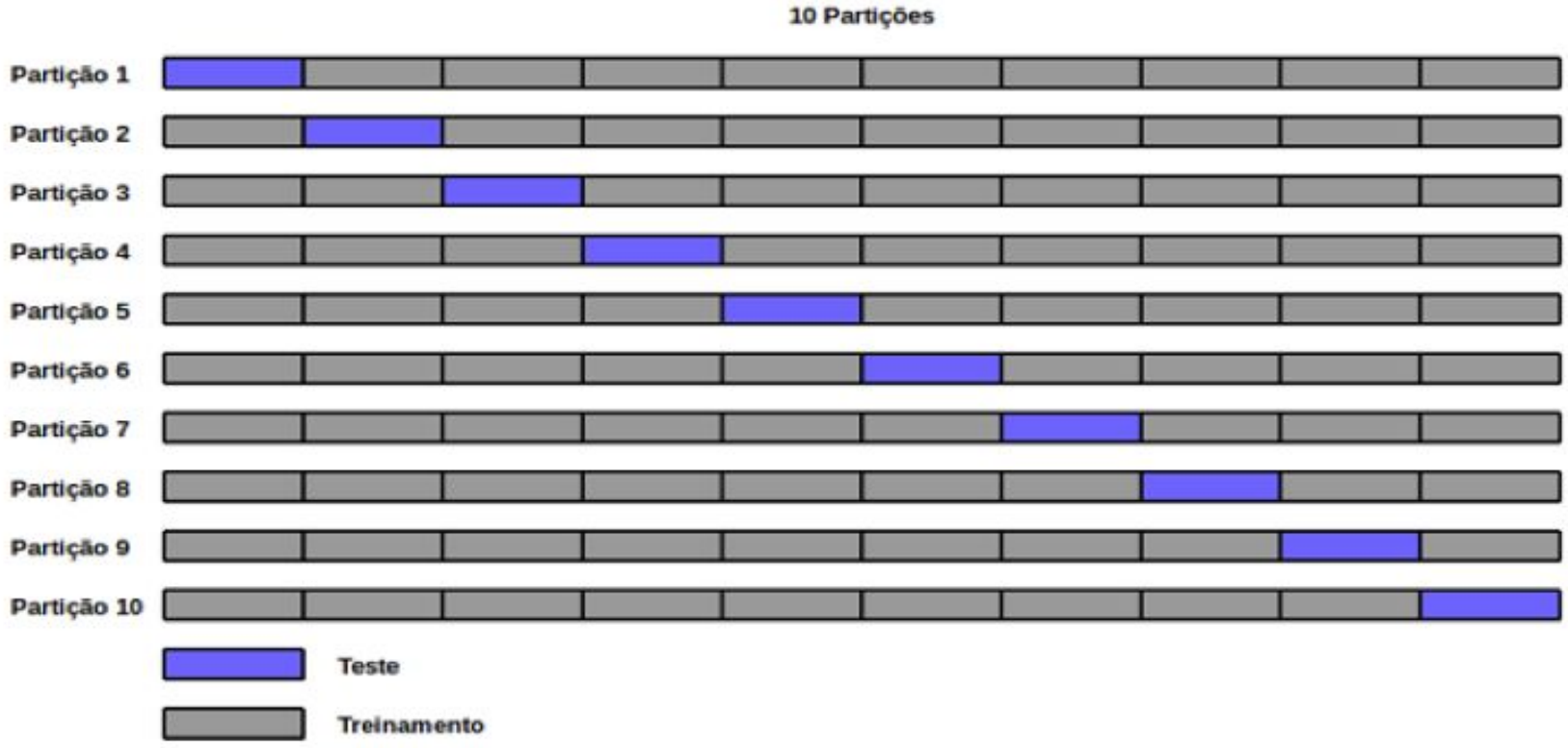
THE DEVELOPER'S CONFERENCE

Algoritmo	Acurácia	F1 Score
Naive Bayes	0.9820833333333333	0.9820810654681816
Arvore de Decisão	0.9645833333333333	0.9645781614989273
Floresta Aleatória	0.98875	0.9887481227336714
KNN	0.8304166666666667	0.8243225817242609
Regressão Logística	0.9945833333333334	0.9945810745069256
SVM - Linear	0.99375	0.9937476021620096

Validação cruzada

- Diminui a aleatoriedade dos dados pois divide os testes em k folds

Validação cruzada



Validação cruzada

THE DEVELOPER'S CONFERENCE

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
cv_results = model_selection.cross_val_score(model, bag_of_words,
                                             y, cv=kfold, scoring='accuracy')
```

Resultados (Cross Val)

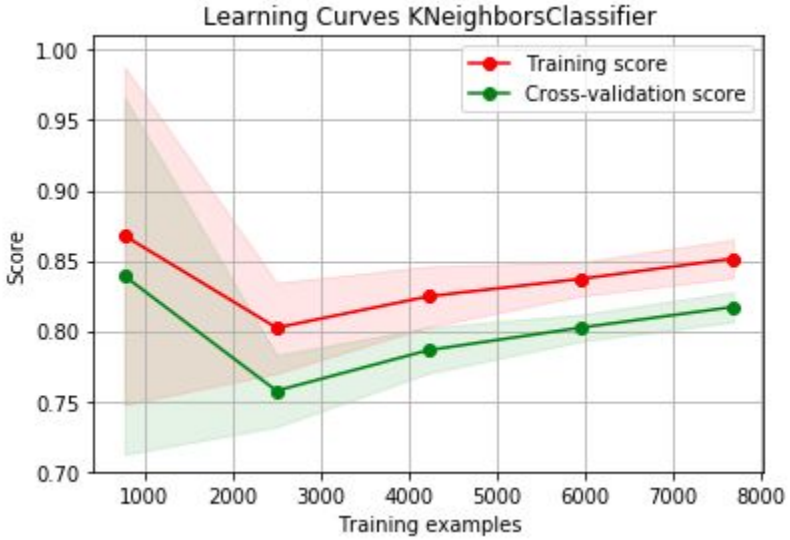
THE DEVELOPER'S CONFERENCE

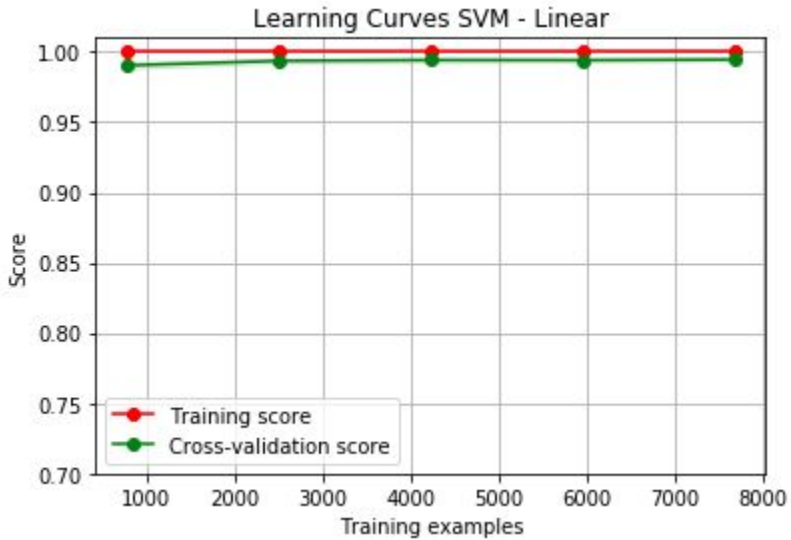
Algoritmo	Acurácia	Desvio Padrão
Naive Bayes	0.987750	0.002740
Arvore de Decisão	0.969833	0.006200
Floresta Aleatória	0.991750	0.003038
KNN	0.837250	0.014439
Regressão Logística	0.996583	0.001417
SVM - Linear	0.994750	0.001239

Gráficos apoio validação

- boxplot
- matriz de confusão
- curva roc
- tabela curva de aprendizagem (treino / validação cruzada)

Gráficos apoio validação





TJRS Inovação (hoje)

THE DEVELOPER'S CONFERENCE

- Executivos fiscais (OCR para extração e classificação dados coletados)
- Chatbot

Possibilidades aplicações

THE DEVELOPER'S CONFERENCE

- Extração e validação de dados de petição inicial
- Classificação de petições iniciais
- Predição, sumarização e sugestão de textos (e modelos de documentos)
- Elaborar ou sugerir despachos e decisões
- Subsidiar o magistrado com teses e jurisprudências sobre a matéria a ser decidida

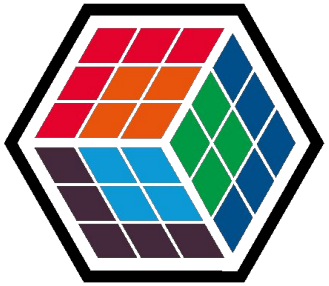
Perguntas??

THE DEVELOPER'S CONFERENCE

Obrigado

THE DEVELOPER'S CONFERENCE

- julianopache@gmail.com
- https://github.com/julianopacheco/
- https://www.twitter.com/julianopacheco_
- https://www.linkedin.com/in/julianopache



THE DEVELOPER'S CONFERENCE