Instituto Brasiliense de Direito Público
Escola de Administração de Brasília

Agrupamento Automático de Documentos Jurídicos com uso de Inteligência Artificial

Amilar Domingos Moreira Martins

Orientador
Prof. Dr. Gilmar Ferreira Mendes.

Brasília
2018

Amilar Domingos Moreira Martins

Agrupamento Automático de Documentos Jurídicos com uso de Inteligência Artificial

Dissertação apresentada como requisito parcial para a conclusão do Mestrado Profissional em Administração Pública.

Orientador
Prof. Dr. Gilmar Ferreira Mendes.

Brasília
2018

Amilar Domingos Moreira Martins

Agrupamento Automático de Documentos Jurídicos com uso de Inteligência Artificial

Dissertação apresentada como requisito parcial para a conclusão do Mestrado Profissional em Administração Pública.

Aprovado em: ___/___/____

BANCA EXAMINADORA

Prof. Doutor André Telles Campos
Examinador Externo

Prof. Doutor Jefferson Carlos Carús Guedes
Examinador Externo

Profª. Doutora Luciana Silva Garcia
Examinadora Interna

Prof. Doutor Gilmar Ferreira Mendes.
Orientador

Para Fabiana, Júlio e Caio.

You´re always on my mind.

AGRADECIMENTOS

Aos meus pais, Geraldo (*in memoriam)* e Jacira, por me ensinarem sobre o poder transformador do conhecimento.

Ao colega do STJ, Luiz Anísio, pelas enriquecedoras trocas de conhecimento.

Aos amigos do mestrado, Mauro Kaiser, Alexandre Nunes, Alexandre Cabral, Gisele e Adonai, pelos quais a vida me ensinou a ter respeito e admiração.

Ao companheiro Montgomery Muniz, pelo incentivo ao início desta jornada.

RESUMO

O Poder Judiciário brasileiro vem, ano após ano, sendo inundado por um volume crescente de demandas que em muito excedem sua capacidade. No Superior Tribunal de Justiça isso não é diferente. Os números apontam um crescimento anual constante nos recursos que aportam no Tribunal da Cidadania, fato que tem consumido grande energia por parte dos Ministros e servidores da Corte. Essa escalada de demandas exige que sejam implementadas soluções inovadoras a fim de garantir uma justiça célere e uniforme. Dentre as diversas atividades repetitivas que são executadas dentro da cadeia de valor do tribunal, observa-se que o agrupamento de feitos semelhantes é uma das mais desgastantes, forçando o analista a realizar a leitura de inúmeros processos a fim de selecionar aqueles que estão dentro de seu espectro de atuação. Nesta realidade, o presente estudo realiza um experimento de aplicação de técnicas de Inteligência Artificial, notadamente de Processamento de Linguagem Natural, no agrupamento de documentos jurídicos de forma a constituir um acelerador da atividade de confecção de minutas de julgamento ao realizar automaticamente a identificação de documentos semelhantes, possibilitando que o operador volte seu esforço na confecção das minutas de julgamento. Para alcançar esse objetivo foi utilizado um corpus de documentos extraídos da jurisprudência indexada do Tribunal, esse corpus foi submetido ao algoritmo *Paragraph Vector*, gerando um modelo de Inteligência Artificial capaz de inferir uma representação vetorial de um documento jurídico. Os resultados da submissão dos documentos de teste ao modelo produzido permitem concluir pela utilidade da técnica nas atividades vinculadas à triagem baseada em documentos. Ao final, são propostas melhorias, novos experimentos e avaliações da aplicação do *Paragraph Vector* em outras atividades repetitivas.

Palavras chave: Superior Tribunal de Justiça. Inteligência Artificial. Processamento de Linguagem Natural. Documentos jurídicos. Agrupamento. Vetor de Parágrafo.

ABSTRACT

The Brazilian Judiciary comes, year after year, being flooded by a growing volume of demands that far exceed its capacity. In the Superior Court of Justice this is no different. The figures point to a constant annual growth in the resources that they bring to the Citizens Tribunal, a fact that has consumed great energy on the part of the Ministers and servants of the Court. This escalation of demands requires that innovative solutions be implemented to ensure swift and uniform justice. Among the many repetitive activities that are performed within the value chain of the court, it is observed that the grouping of similar deeds is one of the most exhausting, forcing the analyst to perform the reading of numerous processes in order to select those that are within its spectrum of performance. In this reality, the present study carries out an experiment of application of techniques of Artificial Intelligence, especially of Natural Language Processing, in the grouping of legal documents in order to constitute an accelerator of the activity of making drafts of judgment by automatically performing the identification of similar documents, allowing the operator to return his effort in the making of the minutes of judgment. In order to achieve this objective, a corpus of documents extracted from the indexed jurisprudence of the Court was used. This corpus was submitted to the Paragraph Vector algorithm, generating an Artificial Intelligence model capable of inferring a vector representation of a legal document. The results of the submission of the test documents to the model produced allow us to conclude that the technique is useful in activities related to document - based sorting. At the end, improvements are proposed, new experiments and evaluations of the application of the Paragraph Vector in other repetitive activities

Keywords: Superior Court of Justice. Artificial intelligence. Natural Language Processing. Legal documents. Grouping. Paragraph Vector.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

AB2L	Associação Brasileira de LawTechs e LegalTechs
AREsp	Agravo em Recurso Especial
BoW	Bag of Words (saco de palavras)
CF/88	Constituição Federal de 1988
CNJ	Conselho Nacional de Justiça
DBSCAN	Density Based Spatial Clustering of Application with Noise (Agrupamento Espacial Baseado em Densidade de Aplicação com Ruído)
IA	Inteligência Artificial
OCR	Optical Character Recognition (Reconhecimento Ótico de Caracteres)
PCA	Principal Component Analisis (Análise de Componente Principal)
PLN	Processamento de Linguagem Natural
PV-DM	Distributed Memory Model of Paragraph Vectors
REsp	Recurso Especial
RHC	Recurso em Habeas Corpus
SIAJ	Sistema Integrado de Atividade Judiciária
SJR	Secretaria de Jurisprudência do Superior Tribunal de Justiça
STI	Secretaria de Tecnologia da Informação e Comunicação
STJ	Superior Tribunal de Justiça
TRF3	Tribunal Regional Federal da 3ª Região

SUMÁRIO

Capítulo 1 - Introdução

Este capítulo descreve o contexto em que o experimento proposto se torna útil, seus objetivos e motivações.

1.1 - Contexto Geral

De acordo com o relatório Justiça em Números 2017[1], do Conselho Nacional de Justiça (CNJ), em 2016 entraram no judiciário brasileiro 29,4 milhões de ações.

O Relatório informa que para cada grupo de 100 mil habitantes, há 12,907 mil ações em estoque. O gráfico abaixo apresenta a evolução da entrada e do estoque de processos.

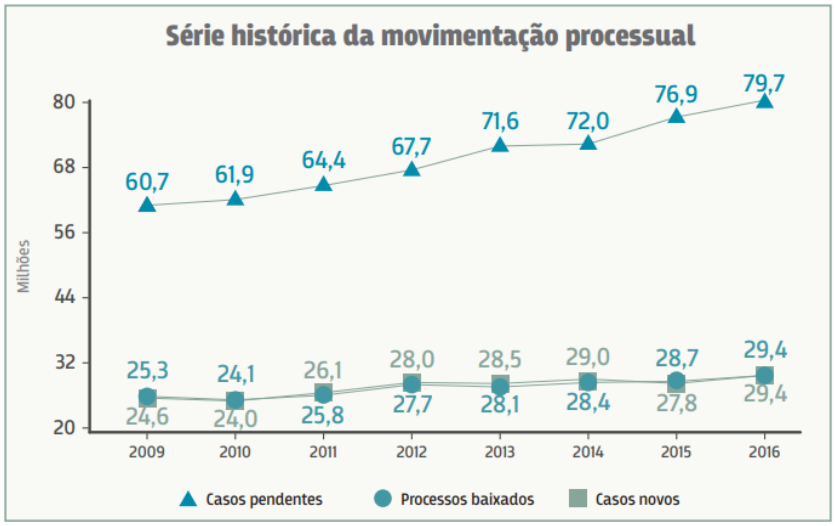

Gráfico 1. Série histórica da movimentação processual no Brasil.

Ainda de acordo com o Relatório, em 2016, cada juiz brasileiro solucionou 1,749 mil processos, o que equivale a mais de sete por dia. Embora a produtividade dos julgadores possa ser considerada elevada, ela foi suficiente apenas para manter o atendimento à demanda de entrada, fazendo com que o estoque apresentasse crescimento de 2,8 milhões de feitos.

Conclui-se que, ainda que a produtividade dos magistrados seja mantida nos altos patamares atuais, o número de processos pendentes tornará inviável o funcionamento do poder judiciário a médio prazo, quando observada a tendência do estoque.

Importante registrar também que 70,1% das demandas submetidas ao judiciário vieram por meio eletrônico. Esse número vem apresentando evolução crescente, conforme observamos abaixo:

[1]Conselho Nacional de Justiça. Justiça em Números. 2017. Disponível em http://www.cnj.jus.br/programas-e-acoes/pj-justica-em-numeros. Último acesso em 16/08/2018.

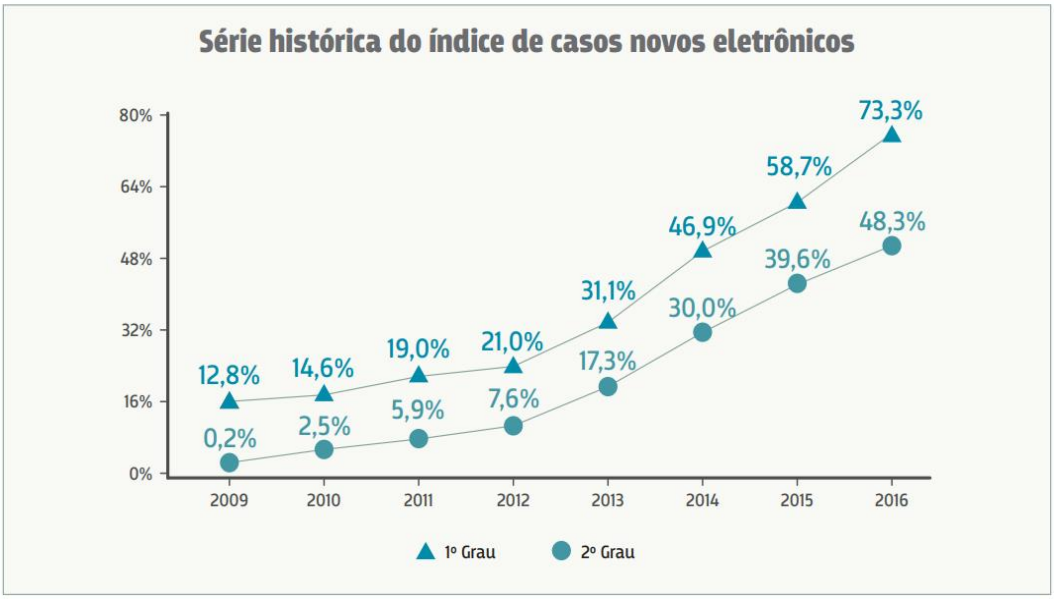

Gráfico 2. Série histórica do índice de casos novos eletrônicos no judiciário nacional.

Esse volume de processos eletrônicos, notadamente quando seus documentos são remetidos em formato texto, pode transformar o desafio do excesso de feitos em uma oportunidade, quando buscamos aproveitar o turbilhão de informações que eles trazem.

Em especial, a aplicação de técnicas de Inteligência Artificial na gestão e aceleração da prestação jurisdicional pode refletir o que já está ocorrendo no mercado jurídico. As chamadas *Lawtechs* ou *Legaltechs* têm investido na IA para solucionar os gargalos jurídicos enfrentados pelos operadores. A Associação Brasileira de LawTechs e LegalTechs (AB2L) publica periodicamente um gráfico chamado Radar, que descreve as empresas atuando nas principais áreas vinculadas à IA no setor jurídico.

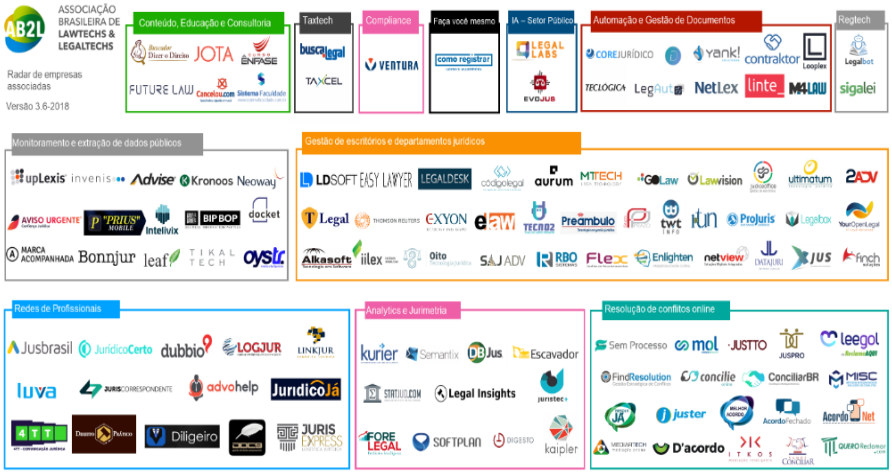

Figura 1. Radar de lawtechs e legaltechs.

Entretanto, o Poder Judiciário brasileiro não tem demonstrado o mesmo interesse no desenvolvimento de soluções baseadas em inteligência artificial. Em pesquisa realizada junto a 55 tribunais do país, das justiças federal, comum e trabalhista, apenas 4 informaram ter iniciativas voltadas à aplicação de IA.

Mesmo entre essas iniciativas, não há informações de um planejamento de convergência das soluções, ou seja, estão sendo desenvolvidas soluções concorrentes, em um momento que demanda uma confluência de iniciativas de forma a gerar resultados melhores, mais rapidamente e a um custo menor.

1.2 – Contexto Específico – O Superior Tribunal de Justiça

O Superior Tribunal de Justiça (STJ) é órgão do Poder Judiciário, nos termos do art. 92, II da Constituição Federal de 1988 (CF/88), composto de no mínimo 33 ministros, nomeados pelo Presidente da República dentre brasileiros com mais de trinta e cinco e menos de sessenta e cinco anos, de notório saber jurídico e reputação ilibada. A função precípua do Tribunal é uniformizar a interpretação da legislação federal[2].

[2] CF/1988. Art. 105. Compete ao Superior Tribunal de Justiça:
(...) III - julgar, em recurso especial, as causas decididas, em única ou última instância, pelos Tribunais Regionais Federais ou pelos tribunais dos Estados, do Distrito Federal e Territórios, quando a decisão recorrida:
a) contrariar tratado ou lei federal, ou negar-lhes vigência;
b) julgar válido ato de governo local contestado em face de lei federal;
c) der a lei federal interpretação divergente da que lhe haja atribuído outro tribunal.

Embora sua função constitucional seja manter a uniformidade da interpretação da legislação federal, o STJ tem, ao longo dos anos, sido submetido a uma situação em que o uso predatório do sistema judicial absorve toda sua energia em um turbilhão de recursos que, respeitada a designação estabelecida pela CF/88, não deveriam chegar à corte.

O quadro abaixo demonstra a evolução no número de feitos recebidos na Corte na última década.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
	271.521	292.103	228.981	290.901	289.524	309.677	314.316	332.905	335.779	332.284
Variação	-	7,58%	-21,61%	27,04%	-0,47%	6,96%	1,50%	5,91%	0,86%	-1,04%

Tabela 1. Evolução da entrada de feitos no Superior Tribunal de Justiça.

Transpostos os dados para o gráfico, podemos observar de forma mais clara a tendência de aumento constante do número de feitos que chegaram ao STJ nos últimos 10 anos:

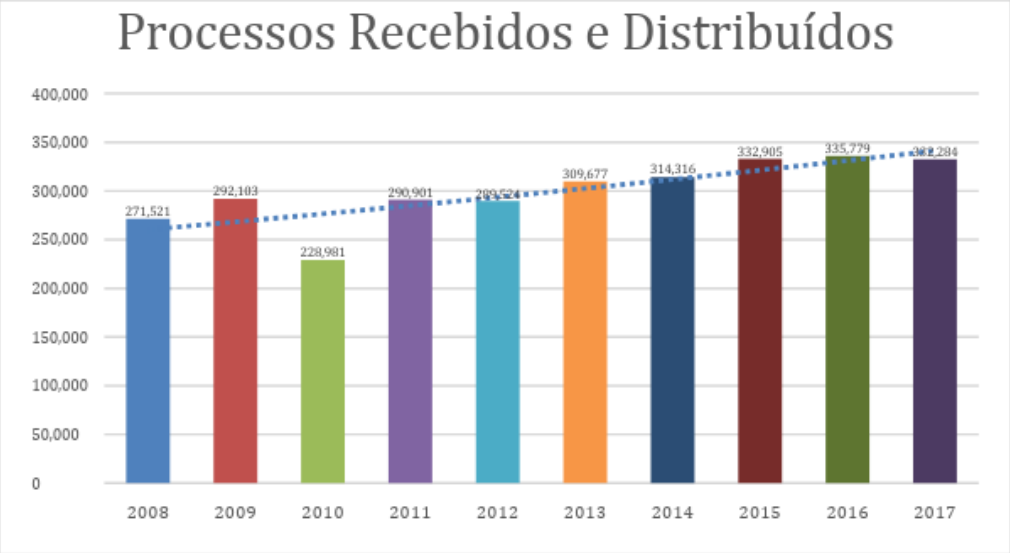

Gráfico 3. Série histórica de processos recebidos e distribuídos no STJ.

Observa-se que a entrada de feitos partiu de 271.521 em 2008 e alcançou, em 2017, um total de 332.284[3], perfazendo neste ano uma média de 10.069 processos para cada um de seus 33 ministros, consistindo tal volume em carga de trabalho incompatível não só com a missão, mas também com o número de ministros da Corte.

[3] Superior Tribunal de Justiça - Relatório Estatístico 2017. disponível em http://www.stj.jus.br/webstj/Processo/Boletim/verpagina.asp?vPag=0&vSeq=301, último acesso em 14/08/2018;

Na tentativa de diminuir a sobrecarga aos demais ministros, evitando que processos que trazem vícios sejam àqueles distribuídos, as presidências do tribunal vêm implementando há vários mandatos consecutivos filtros que são aplicados aos feitos na entrada, permitindo que sejam tratados rapidamente processos cujos vícios inviabilizam o conhecimento da causa ou cuja matéria já esteja tratada sob o rito dos recursos repetitivos[4].

Para a identificação dos processos e aplicação destes filtros, a quase totalidade dos feitos que ingressam no Tribunal da Cidadania passa por uma série de atividades de análise e transcrição de informações. No desempenho dessas atividades, faz-se necessária uma aplicação massiva de mão-de-obra em tarefas excessivamente repetitivas, ao mesmo passo, essa mão de obra torna-se mais escassa a cada dia, em parte por conta de restrições orçamentárias e em outra parte por conta de políticas de lotação de pessoal.

Nesse contexto de escassez de recursos humanos, a aplicação de técnicas de Inteligência Artificial (IA)[5] para o apoio a tarefas repetitivas pode ajudar a dinamizar as atividades de toda a cadeia de valor do STJ, mesmo em um ambiente visivelmente desfavorável.

Por outro lado, conforme detalha Muniz (2018), pode-se afirmar que quase a totalidade dos feitos que aportam ao STJ trata de matérias cuja jurisprudência já é estável, algo que torna o Superior Tribunal de Justiça uma espécie de “terceira instância” no sistema judiciário brasileiro. Afirma o autor que 94,6% dos processos julgados no STJ em 2016 podem ser considerados “casos fáceis”, ou seja, aqueles “onde apenas se repetem decisões anteriores”.

Partindo da premissa que a grande maioria dos feitos que aportam no tribunal tratam de matérias já debatidas e com entendimento firmado, caso a Corte esteja munida de uma ferramenta que permita indicar processos semanticamente semelhantes àquele sob análise, seria possível aplicar a este a mesma solução aplicada ao anterior.

---

[4]Sobre Recursos Repetitivos, Núcleo de Gerenciamento de Precedentes – Nugep – Gabinete da Presidência – Superior Tribunal de Justiça – Disponível em http://www.stj.jus.br/sites/STJ/default/pt_BR/Processos/Repetitivos-e-IAC/Saiba-mais/Sobre-Recursos-Repetitivos, último acesso em 14/10/2017;

[5] Segundo Atheniense, Alexandre R.; Resende, Tatiana C., “Inteligência artificial (*artificial intelligence - AI*) é um termo criado pela primeira vez em 1956, por um cientista da computação norte-americano chamado John McCarthy, para designar o desempenho de máquinas capazes de realizar tarefas características da inteligência humana.”

Neste sentido, é importante firmar que o conceito de feito semelhante é, na verdade, o conceito de peça processual semelhante, explica-se: em um Agravo em Recurso Especial (AREsp), a peça inicialmente útil ao agrupamento é o Agravo. Para um Recurso Especial (REsp) a peça útil ao agrupamento é o Recurso Especial, para um Recurso em Habeas Corpus (RHC), essa peça é o Recurso Ordinário.

Em outras palavras: o agrupamento de feitos semelhantes utiliza como parâmetro a peça relevante naquele momento processual, e não todo o processo.

Registre-se ainda que a semelhança entre duas peças não pode ser medida pelo número de termos iguais, mas sim pela semelhança semântica entre essas peças. É essa semelhança semântica que esse experimento busca quantificar com o uso de IA.

O problema que se busca resolver é que atualmente boa parte da energia despendida pelos assessores dos Gabinetes de Ministros na confecção de minutas de julgamento é voltada para a triagem e seleção de recursos que estejam dentro de seu do escopo de atuação/especialidade. Uma redução nesse esforço contribuiria necessariamente em aumento de produtividade.

Assim, conclui-se que a utilização de recursos que permitam o agrupamento automático de feitos com peças processuais semelhantes poderia constituir um catalisador da produtividade da corte, otimização do uso de recursos humanos e diminuição do tempo de tramitação dos processos, contribuindo para a celeridade processual. Ao mesmo tempo, o agrupamento, quando realizado incluindo recursos que já foram julgados na corte, pode constituir também um acelerador na identificação de precedentes.

Mais que isso, ao agrupar processos semelhantes e possibilitar que tenham desfechos idênticos, o STJ pode dar uma grande contribuição para a segurança do Sistema Judicial Brasileiro, algo que é basicamente sua atribuição constitucional. Nas palavras de McCormick (2008):

> *... se você deve tratar igualmente casos iguais e diferentemente casos distintos, então novos casos que tenham semelhanças relevantes com decisões anteriores devem* (prima facie*, pelo menos) ser decididos de maneira igual ou análoga aos casos passados*. *Conectada a essa razão está a ideia de um sistema jurídico imparcial que faz a mesma justiça a todos, independentemente de quem forem as partes do caso e de quem está julgando. Num estado moderno, com muitos juízes e muitas cortes, e uma hierarquia estruturada de recursos, as mesmas regras e soluções devem*

> *orientar a decisão independentemente do juiz do caso. Fidelidade ao Estado de Direito requer que se evite qualquer variação frívola no padrão decisório de um juiz ou corte para outro.*

1.3 - Justificativa do Tema

O tema deste estudo é impulsionado por fatores vinculados aos princípios constitucionais da celeridade processual e da segurança jurídica ao mesmo tempo em que é possível vislumbrar a possibilidade de replicação da solução para outros ramos do conhecimento.

Apesar de já existirem iniciativas acadêmicas de agrupamento de documentos jurídicos, pouco se explorou da aplicação dessas técnicas para a triagem processual.

1.3.1 - Celeridade Processual

O modelo de Inteligência Artificial que o experimento visa produzir busca auxiliar no agrupamento de feitos, objetivando a geração de decisões iguais com base em modelos precedentes. A partir da identificação deste precedente semântico, a produção da solução judicial para o caso em análise tende a ser mais rápida do que aquela formulada sem a identificação prévia. Neste contexto, a celeridade processual contribui para decisões a um custo unitário menor, prestigiando também o princípio da economicidade.

1.3.2 - Segurança Jurídica

O agrupamento de feitos semelhantes pode contribuir para que estes recebam desfechos igualmente semelhantes, garantindo a segurança jurídica do sistema e assim auxiliando o STJ no cumprimento de sua missão constitucional. Ao mesmo tempo, a insegurança jurídica, mais ainda quando derivada da atuação de um tribunal superior, gera efeitos econômicos extremamente negativos. Segundo Oliveira (2011) quase todos os entraves para o desenvolvimento, que têm origem em incertezas, nascem da insegurança jurídica, que se apresenta, por conseguinte, como o principal e o maior obstáculo para o povo brasileiro se desenvolver economicamente, mas também culturalmente em qualquer outro aspecto de sua existência.

1.3.3 - Possibilidade de Replicação

A trilha do experimento pode apoiar outras áreas do conhecimento que buscam valer-se do Processamento de Linguagem Natural, como a linguística, a literatura, o jornalismo e a jurimetria.

1.4 – Hipótese de Pesquisa:

O presente estudo pretende verificar a hipótese de que a utilização de técnicas de Inteligência Artificial, notadamente aquelas relativas ao Processamento de Linguagem Natural com uso de vetores de parágrafos, para agrupamento de peças processuais semanticamente semelhantes pode auxiliar na celeridade processual ao incrementar o potencial produtivo dos operadores e na segurança jurídica ao facilitar o desfecho idêntico de feitos iguais.

1.5 - Objetivos:

O objetivo geral do estudo é avaliar a efetividade da aplicação de técnicas de Inteligência Artificial no agrupamento de documentos jurídicos no âmbito do Superior Tribunal de Justiça.

Esse objetivo será alcançado com a produção de um modelo de inteligência artificial capaz de agrupar documentos jurídicos semanticamente semelhantes.

Para esse fim, deverão ser alcançados os seguintes objetivos específicos:

1.5.1 – Criar um modelo de Inteligência Artificial baseado em documentos jurídicos.

Um modelo de inteligência artificial será treinado com o uso de um corpus constituído pelos acórdãos indexados pela Secretaria de Jurisprudência do STJ (SJR) entre os anos de 2015 e 2017 num total de 328.732 documentos. Os documentos foram fornecidos pela Secretaria de Tecnologia da Informação e Comunicação do STJ (STI) e seu lapso temporal determinado apenas pela disponibilidade imediata do acesso. O objetivo específico será alcançado com a

produção de um modelo de IA capaz de inferir o vetor de um documento jurídico em um plano multidimensional para o contexto jurídico.

1.5.2 – Submeter um corpus de testes ao modelo

Treinado o modelo, um grupo de documentos jurídicos será submetido a ele de forma a gerar grupos de recursos que, embora não conexos, sejam semanticamente semelhantes. Após essa fase, os grupos serão analisados com base em outros dados lançados no Sistema Integrado de Atividade Judiciária (SIAJ) do STJ e por especialista humano que determinará o grau de eficiência do experimento.

O objetivo específico será alcançado com o agrupamento automático dos documentos jurídicos integrantes da amostra e a validação destes grupos por avaliador humano.

1.6 - O desafio do estudo

O desafio do estudo é garantir que, com a aplicação do modelo produzido, os integrantes de cada um dos grupos estimados estejam semanticamente mais próximos dos demais integrantes do grupo do que de qualquer integrante de outro grupo, independentemente do vocabulário específico utilizado no documento.

1.7 – Limites do Estudo

Não é objeto deste experimento o desenvolvimento de novos algoritmos de IA, bem como a realização em ajustes nos algoritmos já disponíveis no mercado. Também não se destina este trabalho à realização de comparações de desempenho entre abordagens de Inteligência Artificial. O experimento visa aplicação do algoritmo *Paragraph Vector* e a avaliação dessa aplicação em um contexto jurídico.

1.8 - Estrutura deste Documento

Este documento está dividido em cinco capítulos.

Este Capítulo 1 apresenta a contextualização do problema, os objetivos gerais e específicos, as motivações e a estrutura do restante do documento.

O Capítulo 2 expõe os conceitos essenciais da área de Inteligência Artificial que são aplicados ao experimento.

O Capítulo 3 descreve a metodologia da pesquisa, aplicada nas fases de seleção, detalhamento e preparação dos corpora de documentos, detalhando suas estruturas e verificando sua representatividade dentro do contexto jurídico. Apresentará também a técnica empregada na construção do modelo de inteligência artificial, seus desafios, conceitos e justificativas.

O Capítulo 4 demonstra o resultado da aplicação do modelo sobre um *set* de documentos jurídicos que não tenha composto seu treinamento, bem como os resultados da validação realizada por operador humano.

O Capítulo 5 expõe as conclusões do experimento, limitações e potencialidades da solução bem como possibilidades futuras.

Capítulo 2 - Inteligência Artificial

Este capítulo expõe os conceitos essenciais da área de Inteligência Artificial que são aplicados ao experimento.

2.1 - Conceitualização

> Inteligência Artificial (IA) é o mantra da era atual. A frase é entoada por tecnólogos, acadêmicos, jornalistas e capitalistas de risco. Tal como acontece com muitas frases que passam dos campos acadêmicos técnicos para a circulação geral, há um mal-entendido significativo acompanhando o uso da frase. Mas este não é o caso clássico do público que não entende os cientistas - aqui os cientistas ficam tão confusos quanto o público. A ideia de que nossa era está de alguma forma vendo o surgimento de uma inteligência em silício que rivaliza com a nossa própria nos entretém - nos fascinando e nos assustando em igual medida. E, infelizmente, nos distrai.
>
> Michael I. Jordan[6]

É fundamental, para melhor entender o objetivo do experimento, trazer alguns conceitos relativos à Inteligência Artificial. O objetivo aqui não é abranger todos os temas inerentes a esta ciência, mas apontar uma base conceitual mínima aplicável ao experimento que será proposto.

Segundo Minsky (1968), inteligência artificial é a ciência de fazer máquinas executarem tarefas que exigiriam inteligência se fossem feitas por homens. Para alcançar o objetivo dessa ciência, o autor propôs a seguinte questão: “Como alguém pode fazer as máquinas entenderem as coisas?”. Essa pergunta, feita em 1968, é a síntese de um tema que a ciência se esforça para tratar há mais de 50 anos.

Desde o seu surgimento, várias correntes de pensamento e consequentemente de pesquisa se formaram nesta nova ciência. Diversas abordagens foram aplicadas na busca de soluções que possibilitassem à máquina ajustar seu comportamento com base na experiência.

2.1.1 - Aprendizagem de Máquina (Machine Learning)

A aprendizagem de máquina, segundo Mitchell (1997), está preocupada com a questão de como construir programas de computador que melhoram automaticamente com a experiência.

O autor definiu um modelo para caracterizar um programa capaz de aprender como sendo “... um programa de computador aprende com a experiência *E* com respeito a alguma classe de tarefas *T* e medida de desempenho *P*, se o seu

---

[6] JORDAN, Michael I. Artificial Intelligence—The Revolution Hasn’t Happened Yet. 2018, disponível em <https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7>. Acesso em 16/08/2018.

desempenho em tarefas em *T*, conforme medido por *P*, melhora com a experiência *E*.”.

A aprendizagem de máquina é normalmente dividida em duas principais subáreas:

2.1.1.1 - Aprendizado Supervisionado

Subárea da aprendizagem de máquina, aprendizado supervisionado consiste em, pela análise de um conjunto de exemplos portadores de atributos, extrair um padrão que permita a inferência de um atributo alvo ausente quando apresentado um exemplo novo.

Especificamente quanto ao Processamento de Linguagem Natural (PLN), a ideia dos métodos baseados em aprendizado supervisionado consiste em extrair de uma coleção de treinamento as características necessárias para a correta classificação das entidades nomeadas mencionadas em textos desconhecidos (SILVA, 2012).

Frutuoso (2014) afirma que, ainda que consideradas as vantagens do aprendizado supervisionado, é fundamental que a base de treinamento esteja correta e compreenda a maior parte possível dos contextos. Isso torna a criação da base de treinamento “cansativa e custosa, visto que requer mão de obra especializada e um grande esforço de tempo”.

Seguindo este entendimento, pode-se acreditar que o uso de um grupo de documentos extraídos da jurisprudência do STJ abarque suficientemente o contexto jurídico, tendo em vista a ampla jurisdição da Corte.

2.1.1.2 - Aprendizado Não-Supervisionado – Clustering

Ao contrário do aprendizado supervisionado, neste não há uma saída ou um atributo esperado. A técnica de *clustering* consiste em observar os exemplos em um espaço multidimensional buscando agrupá-los.

Segundo Tan, Steinbach e Kumar (2006), a análise de *cluster* divide os dados em grupos (*clusters*) que são significativos, úteis ou ambos. Se grupos significativos são o objetivo, então os *clusters* devem capturar a estrutura natural dos dados. Em alguns casos, no entanto, a análise de cluster é apenas um ponto de

partida útil para outros fins, como a sumarização de dados. De qualquer forma, para compreensão ou utilidade, a análise de *cluster* tem desempenhado um importante papel em uma ampla variedade de campos: psicologia e outras ciências sociais, biologia, estatísticas, reconhecimento de padrões, recuperação de informações, aprendizado de máquina e mineração de dados.

Algoritmos de *cluster* agrupam um conjunto de documentos em subconjuntos ou *clusters*. O objetivo dos algoritmos é criar grupos que sejam coerentes internamente, mas claramente diferentes de outros. Em outras palavras, os documentos dentro de um *cluster* devem ser mais semelhantes quanto possível; e documentos em um *cluster* devem ser tão diferentes quanto possível de documentos em outros *clusters*. (MANNING et al, 2009).

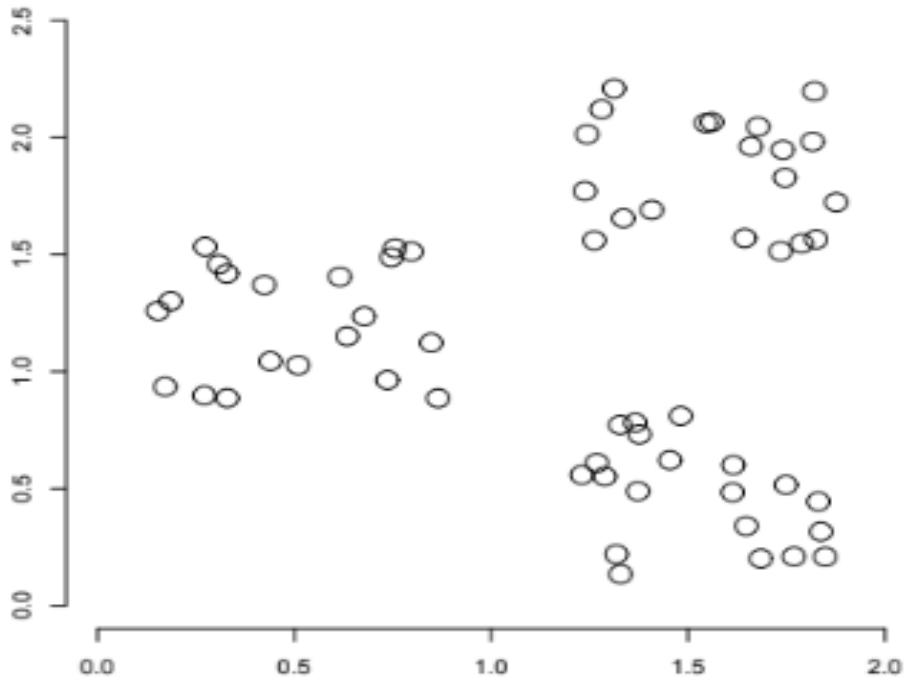

Gráfico 4. Exemplo de um conjunto de dados com estrutura de cluster. (MANNING et al, 2009).

2.2 - Processamento de Linguagem Natural (PLN)

Segundo Vieira & Lopes (2010), Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais.

Glavas et al (2012) ensinam que tarefas de Processamento de Linguagem Natural, como classificação e resumo de texto, recuperação de informação e desambiguação de sentido de palavras, dependem de uma medida de semelhança semântica de documentos textuais.

Manning & Schütze (2000) analisam que uma abordagem estatística de PLN procura resolver esses problemas automaticamente aprendendo preferências lexicais e estruturais dos corpora. Em vez de analisar apenas usando categorias sintáticas, como parte de rótulos de fala, reconhecemos que há muita informação nas relações entre as palavras, ou seja, quais palavras tendem a se agrupar entre si. Este conhecimento colocacional pode ser explorado como uma janela para relacionamentos semânticos mais profundos. Em particular, o uso de modelos estatísticos oferece uma boa solução para o problema da ambiguidade: os modelos estatísticos são robustos, generalizam-se bem e comportam-se graciosamente na presença de erros e novos dados. Assim, os métodos de PLN estatístico abriram o caminho para fornecer uma desambiguação bem-sucedida em sistemas de larga escala usando textos que ocorrem naturalmente. Além disso, os parâmetros dos modelos estatísticos de PLN podem frequentemente ser estimados automaticamente a partir de corpora de texto, e essa possibilidade de aprendizado automático não apenas reduz o esforço humano na produção de sistemas de PLN, mas também levanta questões científicas interessantes sobre a aquisição da linguagem humana.

2.3 - Corpus

Para Sinclair (2005), um corpus é uma coleção de peças de linguagem em formato eletrônico, selecionados de acordo com critérios externos para representar, na medida do possível, uma linguagem ou variedade linguística como fonte de dados para a pesquisa linguística.

2.4 – Word Vector:

Os vetores de palavras representam um salto significativo no avanço de nossa capacidade de analisar relacionamentos entre palavras, sentenças e

documentos. Ao fazê-lo, eles avançam a tecnologia fornecendo às máquinas muito mais informações sobre as palavras do que anteriormente era possível usando representações tradicionais. São vetores de palavras que possibilitam tecnologias como reconhecimento de fala e tradução automática.

Abordagens tradicionais da PLN, como os modelos de codificação e de saco de palavras quentes, embora úteis em algumas tarefas de aprendizagens de máquina, não capturam informações sobre o significado ou contexto de uma palavra. Isso significa que relacionamentos potenciais, como a proximidade contextual, não são capturados em coleções de palavras. Por exemplo, uma codificação simples não pode capturar relações simples, como determinar se as palavras “cachorro” e “gato” se referem a animais que são frequentemente discutidos no contexto de animais domésticos. Essas codificações geralmente fornecem linhas de base suficientes para tarefas simples de PLN (por exemplo, classificadores de spam de email), mas não possuem a sofisticação para tarefas mais complexas, como tradução e reconhecimento de fala. Em essência, as abordagens tradicionais da PLN, como as codificações únicas, não capturam relações sintáticas (estrutura) e semânticas (significado) entre coleções de palavras e, portanto, representam a linguagem de uma maneira muito ingênua. (AHIRE, 2018)

O modelo usa os termos que acompanham um termo alvo para inferir seu significado, assumindo que sinônimos são, geralmente, usados no mesmo contexto. A teoria corresponde ao famoso adágio: ‘diz-me com quem tu andas e lhe direi quem tu és’. Assim, se dois termos habitualmente aparecem acompanhados de um grupo de outros termos, é provável que esses dois termos possuam uma relação semântica, desde que observadas em um mesmo contexto. Observemos os termos “areia” e “praia”, sabidamente não são sinônimos, mas é possível observar que as sentenças “descansar na praia” e “descansar na areia” possuem uma grande proximidade semântica.

Goldberg (2017) explica que embora a semelhança de palavras seja difícil de definir e geralmente seja muito dependente de tarefas, as abordagens atuais derivam da hipótese distributiva [Harris, 1954[7]], afirmando que as palavras são semelhantes se aparecerem em contextos semelhantes.

[7] Zellig S. Harris (1954) Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520

2.5 – Paragraph Vector:

O conceito por trás do *Paragraph Vector* é basicamente uma expansão do modelo *Word Vector* para que opere com parágrafos ou documentos inteiros.

Numa definição mais formal, Le & Mikolov (2014) explicam a proposta como uma estrutura não supervisionada que aprende representações contínuas de vetores distribuídos para partes de textos. Os textos podem ser de tamanho variável, variando de sentenças a documentos. O nome *Paragraph Vector* enfatiza o fato de que o método pode ser aplicado a textos de tamanho variável, desde uma frase ou sentença até um documento grande.

O modelo *Paragraph Vector* tem sido utilizado nas mais diversas áreas do processamento de linguagem natural com muito sucesso. Por esse motivo, foi escolhido para este experimento.

2.6 – Análise de Dados

A análise de dados é o processo no qual dados brutos são ordenados e organizados, para serem usados em métodos que ajudam a explicar o passado e prever o futuro. A análise de dados não é sobre os números, é sobre construir/articular perguntas, desenvolver explicações e testar hipóteses. É um campo multidisciplinar, que combina Ciência da Computação, Inteligência Artificial e Aprendizado de Máquina, Estatística e Matemática e Domínio do Conhecimento (CUESTA, 2013).

Capítulo 3 - Construção do Modelo de Inteligência Artificial

Este capítulo descreve a metodologia da pesquisa, aplicada nas fases de seleção, detalhamento e preparação do corpus de documentos, apresentando suas estruturas e verificando sua representatividade em relação ao contexto jurídico. Descreve também a técnica empregada na construção do modelo de inteligência artificial, seus desafios, conceitos e justificativas.

3.1 – O Corpus.

O trabalho de treinamento do modelo será desenvolvido com utilização de um conjunto de 328.732 acórdãos indexados pela SJR entre 2015 e 2017. O volume de documentos parece adequado ao objetivo do experimento, uma vez que pode traduzir a amplitude linguística de um contexto jurídico.

Os documentos somam um total de 643 Mb, englobam um total de 318.122 processos, e geram um dicionário de 49.165 palavras únicas.

3.1.1 – Explorando os Documentos.

Os documentos foram apresentados em arquivos individuais em formato JSON[8] com seu conteúdo dividido em 18 campos, listados no quadro abaixo:

Campo	:	Descrição
SG_CLAS_SUC	:	Sigla da Classe do processo sucessivo.
SEQ_SUCESSIVO	:	Sequencial do processo sucessivo.
ANOPB_S	:	Ano de publicação do processo sucessivo.
ANOPB_P	:	Ano de publicação do processo principal.
COD_DOC_SUCESSIVO	:	Código do documento sucessivo.
NUM_REGISTRO_S	:	Número de registro do processo sucessivo.
NUM_REGISTRO_P	:	Número de registro do processo principal.
EMENTA	:	Ementa do acórdão.
DTPUB_S	:	Data de publicação do processo sucessivo.
DTPUB_P	:	Data de publicação do processo principal.
SEQ_PRINCIPAL	:	Sequencial do processo principal.
COD_DOC_PRINCIPAL	:	Código do documento principal.
COD_ORGAO_JGDR	:	Órgão julgador do processo sucessivo.
NUM_MIN_RELATOR	:	Número do ministro relator do processo sucessivo.
TIPO	:	Tipo do processo.

[8] Vide http://www.json.org/

FOLD	:	Grupo de amostragem.
ACORDAO	:	Parte dispositiva do acórdão.
SG_CLASS_PRINC	:	Sigla da classe do processo principal.

Tabela 2. Dicionário de dados dos documentos do corpus de treinamento.

São de interesse para esse experimento apenas os campos SEQ_SUCESSIVO, NUM_REGISTRO_S e EMENTA.

Abaixo se vê um exemplo do conteúdo do campo EMENTA, que traz o texto que será utilizado para treinamento.

> ADMINISTRATIVO E PROCESSUAL. AGRAVO REGIMENTAL. MANDADO DE SEGURANÇA. PROCESSO ADMINISTRATIVO DISCIPLINAR. PENA DE DEMISSÃO EM RAZÃO DE IMPROBIDADE. OPERAÇÃO EUTERPE. ALEGAÇÃO DE VIOLAÇÃO DO DEVIDO PROCESSO LEGAL. PARCIALIDADE (SUSPEIÇÃO) NÃO COMPROVADA. LEGÍTIMA UTILIZAÇÃO DA PROVA EMPRESTADA. AUSÊNCIA DE REFORMATIO IN PEJUS. INDEFERIMENTO LIMINAR MANTIDO. SEGURANÇA DENEGADA. Histórico da demanda 1. Trata-se de procedimento Administrativo Disciplinar que resultou em demissão no Ibama em razão de improbidade administrativa. Conforme consta de Relatório Final da Comissão Processante, 'Desmantelou a Polícia Federal na Operação Euterpe, o mundo criminoso instalado no âmbito do meio ambiente, atingindo o cerne da quadrilha, o núcleo interno, formado por vários segmentos de servidores do Ibama/RJ, precipuamente fiscais e técnicos e o externo, que se valia do poder financeiro para proteger seus interesses ilegais'. 2. Consta do Relatório Final da Comissão Processante a descrição das condutas investigadas que deram origem à presente impetração, que 'O investigado Leonardo mantém uma forte relação extra-funcional com (..), empresário na área de construção e de consultoria na área de meio ambiente. Na maioria das conversas existentes e degravadas neste relatório, identifica-se práticas criminais ocorridas entre ambos. (..) Verifica-se a ocorrência de fiscalização por parte do Ibama em obra do interlocutor de Maia. Inclusive nestas ligações, as conversas giram entorno de quanto, em valores, pode-se pagar. Posteriormente, Maia, em conversa com Leonardo, na data de (..) informa a este que a pessoa bateu o pé em valor de dois mil reais, que foram depositados na conta de Maia através de cheque conforme conversa (..) entre Maia e Isidoro. Leonardo também serve de intermediário entre Maia e os outros fiscais do Ibama. Quando alguma obra é fiscalizada pelo pessoal do Ibama, os empresários ligam para Maia, este descobre quem foi o fiscal, em seguida liga para Leonardo. Efetiva o contato entre as partes (..). Leonardo também indica os serviços de Marcos Maia para empresas que ele mesmo fiscaliza, é o caso da Construtora Ontra, a qual Leonardo liga na data de (..) para maia e informa que o pessoal da Ontra vai procurá-lo em nome dele. Restou comprovado do apurado que o acusado Leonardo Edward Rose não respeitou as normas legais atuando de acordo com seus interesses, sendo desleal à instituição, arredio às atribuições de seu cargo (..). Resultou demonstrado no apurado, forte e consistente comprovação de que o acusado Leonardo Edward Rose, associado às condutas de (..), obteve vantagem pecuniária ilícita, em detrimento da dignidade da função pública e se utilizando da condição de servidor público (..)'. Precedentes da Terceira Seção e julgamento em curso na Primeira 3. A Terceira Seção julgou dois Mandados de Segurança bastante semelhantes, concedendo a segurança. Trata-se dos MS 14.958, da relatoria do Min. Napoleão Nunes Maia Filho, e 14.959, da relatoria do Min.

Haroldo Rodrigues. Na Primeira Seção, por sua vez está em curso o julgamento do MS 15.321, sob a relatoria do Min. Mauro Campbell Marques. Neste, após o voto do relator, denegando a segurança, o Min. Napoleão Nunes Maia Filho abriu divergência, no que foi acompanhado pela Min. Regina Helena Costa, tendo o Min. Benedito Gonçalves pedido vista. Respeitosamente, divirjo do entendimento adotado pela Terceira Seção, comungando do pensamento expressado pelo Min. Mauro Campbell Marques. Ausência de influência de Carlos Minc sobre o processo 4. Não há prova de influência externa sobre o processo. Carlos Minc efetivamente foi responsável pela denúncia de casos de corrupção como deputado estadual. Porém, ele: a) não era Ministro de Estado quando do início do processo administrativo, b) não nomeou a primeira Comissão Processante, c) não era competente para anular o primeiro PAD após o Relatório da Comissão Processante, ato que competia e foi levado a efeito pelo Presidente do Ibama com fundamento em parecer consultivo de Procurador Federal e do Presidente nomeado de Comissão Processante; d) não exercia influência perniciosa, dado que os Presidentes das Comissões Processantes sempre foram Procuradores Federais submetidos ao Advogado Geral da União; e) não assinou a Portaria de demissão do agravante. Atuação do Procurador Federal Elielson Ayres de Souza 5. Inexistente vício na nomeação do Procurador Federal Elielson Ayres de Souza como Presidente de Comissão Processante. Por sugestão de parecer consultivo da Procuradoria Federal, José Batista Lima, Presidente do Ibama, determinou a nulidade parcial do Processo e, nos termos da Lei 8.112/1990, art. 169, constituiu nova Comissão Processante e nomeou o referido Procurador Federal como Presidente. Na análise exclusiva dos vícios processuais insanáveis, recomendou a nulidade total e a fragmentação das Portarias de instauração de PADs, o que foi acolhido. Ausentes irregularidades no procedimento. Imparcialidade da Ministra do Meio Ambiente 6. Não se provou a parcialidade da Ministra de Estado do Meio Ambiente, por ter sido supostamente 'seguidora' de Carlos Minc. Sua decisão foi amparada em minucioso Relatório Final de Comissão Processante de 458 laudas, após anos de colheita de material probatório. Prova penal emprestada 7. A doutrina e a jurisprudência são favoráveis à 'prova emprestada', respeitados o contraditório e a ampla defesa no âmbito do processo administrativo disciplinar, cujo traslado da prova penal foi antecedido e devidamente autorizado pelo Juízo Criminal. (Precedentes do STF: Plenário, QO no Inq. 2275, Rel. Min. Carlos Britto, DJe de 26.9.2008; precedentes desta Corte Superior: MS 11.965/DF, Terceira Seção, Rel. Min. Paulo Medina, Rel. p/ Acórdão Min. Arnaldo Esteves Lima, DJ de 18.10.2007; MS 9.212/DF, Terceira Seção, Rel. Min. Gilson Dipp, DJ de 1º.6.2005; MS 7.024/DF, Terceira Seção, Rel. Min. José Arnaldo da Fonseca, DJ de 4.6.2001). 8. Dispensável averiguar se a decisão do Juízo da 5ª Vara Federal de São João de Meriti dando-se incompetente causaria a nulidade da prova produzida e do compartilhamento efetuado. O Tribunal Regional Federal da 2ª Região deu provimento a Recurso em Sentido Estrito do Ministério Público Federal, dando pela competência daquele Juízo, que deferiu e acompanhou as interceptações telefônica, as prisões temporárias e preventivas, as buscas e apreensões, recebeu a denúncia e interrogou todos os réus (RSE 2008.51.1.0004785-0/RJ, rel. Juiz convocado Aluísio Gonçalves de Castro Mendes, Fed. 1ª Turma Especializada do Tribunal Regional Federal da 2ª Região, julg. em 10/3/2010, Dje 9/4/2010). 9. O eminente Min. Néfi Cordeiro negou seguimento aos Recursos Especiais interpostos contra o RSE 2008.51.1.0004785-0/RJ, tendo a Sexta Turma do STJ negado provimento ao Agravo Regimental (AgRg no REsp 1.228.404 , julgado em 15/12/2016, acórdão publicado em 2/2/2016). Ausência de reformatio in pejus 10. Não ocorreu reformatio in pejus. O primeiro Relatório Final produzido não foi ratificado e carece de natureza vinculante. Ao serem destacadas apenas decisões em sentido técnico, houve uma primeira anulação de processo administrativo sem qualquer juízo prévio sobre o ilícito, prolatada pelo Presidente do Ibama, e ulterior decisum pela demissão do agravante,

> proferido pela Ministra de Estado do Meio Ambiente. Do cotejo de ambos não se extrai a alegada nulidade. Conclusão 11. Agravo Regimental não provido.",

Como podemos perceber, o texto do documento traz uma série de redundâncias e especificidades que dificultariam, se utilizado da forma que está, sua utilização para treinamento do modelo. Essa característica sugere a aplicação de atividades de pré-processamento nos documentos.

3.1.2 – Atividades de Pré-processamento do Corpus.

Chamamos de pré-processamento as atividades de ajuste e padronização de um conjunto de dados com o objetivo de extração de informações.

Lavelli, Sebastiani e Zanoli (2004) apontam que diferentes abordagens para representação de documentos podem resultar de diferentes escolhas (i) quanto ao que é um termo e (ii) como os pesos a termo devem ser computados. Uma escolha frequente para (i) é usar palavras isoladas (removidas as *stop words*, ou seja, palavras neutras como artigos e preposições, que geralmente são removidas com antecedência) ou seus troncos (ou seja, suas raízes morfológicas). Dessa forma, para o melhor aproveitamento do corpus, é necessário aplicar sobre ele atividades de pré-processamento que nesse experimento corresponderão à conversão do documento original em outro, limpo, através da aplicação das seguintes atividades:

3.1.2.1 – Conversão para minúsculas:

Todos os caracteres dos documentos que compõem o corpus serão convertidos para minúsculo. Essa padronização garante que não haverá diferenciação entre termos iguais grafados em maiúsculas ou em minúsculas.

3.1.2.2 – Remoção de *StopWords*:

São consideradas *stopwords* termos que, devido ao seu uso em larga escala, não agregam informações úteis para a construção do modelo de inteligência artificial. No experimento foram removidos dos textos advérbios, conjunções, siglas de unidades da federação, algarismos romanos, preposições, pronomes, eventuais caracteres soltos e as conjugações dos verbos de ligação mais comuns (ser, estar, parecer, andar, viver, ficar, tornar, tornar-se, virar, permanecer e continuar).

3.1.2.3 - Remoção de termos entre parênteses:

Da análise inicial dos documentos do corpus, verificou-se que o conteúdo entre parênteses, no maior número de vezes, não agrega conteúdo útil ao modelo de IA. Por conta dessa observação, optou-se por remover todos os textos entre parênteses.

3.1.2.4 - Remoção de Não-letras:

Todos os caracteres que não correspondem a letras foram removidos, isso inclui algarismos arábicos, hifens, parênteses, aspas, barras e traços.

3.1.2.5 – Singularização de Verbetes:

Um algoritmo de conversão de plurais em singulares foi aplicado sobre o texto. Isso tende a possibilitar uma melhora no cálculo da frequência dos termos ao igualar singulares e plurais (p. ex.: réu preso = réus presos).

3.1.2.6 – Conversão de caracteres especiais:

Todos os caracteres acentuados e a cedilha foram convertidos para suas versões básicas. Essa atividade garante que eventuais erros de grafia sejam ignorados na construção do modelo.

Ao final do pré-processamento, o texto exemplo assumiu a seguinte feição:

> administrativo processual agravo regimental mandado seguranca processo administrativo disciplinar pena demissao razao improbidade operacao euterpe alegacao violacao devido processo legal parcialidade comprovada legitima utilizacao prova emprestada ausencia reformatio peju indeferimento liminar mantido seguranca denegada historico demanda trata procedimento administrativo disciplinar resultou demissao ibama razao improbidade administrativa consta relatorio final comissao processante desmantelou policia federal operacao euterpe mundo criminoso instalado ambito meio ambiente atingindo cerne quadrilha nucleo interno formado vario segmento servidor ibama precipuamente fiscal tecnico externo valia poder financeiro proteger interesse ilegal consta relatorio final comissao processante descricao conduta investigada deram origem presente impetracao investigado leonardo mantem forte relacao extra funcional empresario area construcao consultoria area meio ambiente maioria conversa existente degravada relatorio identifica pratica criminal ocorrida ambos erifica ocorrencia fiscalizacao parte ibama obra interlocutor nesta ligacao conversa giram entorno valor pode pagar conversa leonardo data informa pessoa bateu pe valor doil mil real depositado conta atrave cheque conversa isidoro leonardo serve intermediario fiscal ibama alguma obra fiscalizada pessoal ibama empresario ligam descobre fiscal seguida liga leonardo efetiva contato parte leonardo indica servico empresa fiscaliza caso construtora ontra leonardo liga data informa pessoal ontra vai procura nome restou comprovado apurado acusado leonardo edward rose respeitou norma legal atuando

acordo interesse desleal instituicao arredio atribuicao cargo resultou demonstrado apurado forte consistente comprovacao acusado leonardo edward rose associado conduta obteve vantagem pecuniaria ilicita detrimento dignidade funcao publica utilizando condicao servidor publico precedente julgamento curso primeira julgou doil mandado seguranca semelhante concedendo seguranca trata relatoria relatoria haroldo rodrigue vez curso julgamento relatoria apo voto denegando seguranca abriu divergencia acompanhado tendo pedido vista respeitosamente divirjo entendimento adotado comungando pensamento expressado ausencia influencia carlo minc processo prova influencia externa processo carlo minc efetivamente responsavel denuncia caso corrupcao deputado estadual inicio processo administrativo nomeou primeira comissao processante competente anular pad apo relatorio comissao processante ato competia levado efeito presidente ibama fundamento consultivo federal presidente nomeado comissao processante exercia influencia perniciosa dado presidente comissao processante federal submetido geral uniao assinou portaria demissao atuacao federal elielson ayr inexistente vicio nomeacao federal elielson ayr presidente comissao processante sugestao consultivo procuradoria federal batista presidente ibama determinou nulidade parcial processo termo lei constituiu nova comissao processante nomeou referido federal presidente analise exclusiva vicio processual insanavel recomendou nulidade total fragmentacao portaria instauracao pad acolhido ausente irregularidade procedimento imparcialidade meio ambiente provou parcialidade meio ambiente ter supostamente seguidora carlo minc decisao amparada minucioso relatorio final comissao processante lauda apo ano colheita material probatorio prova penal emprestada doutrina jurisprudencia favoravel prova emprestada respeitado contraditorio ampla defesa ambito processo administrativo disciplinar cujo traslado prova penal antecedido autorizado juizo criminal dispensavel averiguar decisao juizo vara federal meriti dando incompetente causaria nulidade prova produzida compartilhamento efetuado tribunal regional federal regiao deu provimento recurso sentido estrito ministerio publico federal dando competencia juizo deferiu acompanhou interceptacao telefonica prisao temporaria preventiva busca apreensao recebeu denuncia interrogou reu eminente nefi negou seguimento recurso especial interposto rse tendo stj negado provimento agravo regimental ausencia reformatio peju ocorreu reformatio peju relatorio final produzido ratificado carece natureza vinculante destacada apena decisao sentido tecnico primeira anulacao processo administrativo juizo previo ilicito prolatada presidente ibama ulterior decisum demissao proferido meio ambiente cotejo ambos extrai alegada nulidade conclusao agravo regimental provido

3.1.3 – Identificação de ngramas:

Ngramas são palavras que atuam mais comumente juntas em determinado contexto do que separadas. A título de exemplo, em um contexto jurídico, a reunião dos termos “Superior”, “Tribunal” e “Justiça” tem um valor semântico diferente daqueles que os termos assumem quando separados. Os ngramas mais comuns são os bigramas, que reúnem dois termos, e os trigramas, que reúnem três. Dentro de um modelo de IA voltado para PLN, os ngramas são importantes ao garantir que termos que apresentam um valor semântico específico quando reunidos sejam tratados como um termo único. Isso contribui para a redução da dispersão do modelo.

Para a identificação de ngramas não basta a seleção de quaisquer grupos de palavras adjacentes. Na frase: “Na análise exclusiva dos vícios processuais insanáveis, recomendou a nulidade total e a fragmentação das Portarias de instauração de PADs, o que foi acolhido.”, o tratamento dos termos “análise exclusiva” ou “nulidade total” como ngramas não agrega conteúdo semântico útil no contexto jurídico. Por outro lado, “vícios processuais insanáveis” é um trigrama que tem peso semântico, isso demonstra que a mera recorrência de termos não é suficiente para determiná-los como úteis.

Neste experimento iremos calcular bigramas e trigramas utilizando a classe phrases do GenSim, baseado no trabalho de Mikolov, et. Al. (2013). Os autores explicam que uma limitação inerente às representações de palavras é sua indiferença à ordem das palavras e sua incapacidade de representar frases idiomáticas. Por exemplo, os significados de “Canadá” e “Air” não podem ser facilmente combinados para obter “Air Canada”. Motivado por este exemplo, os autores apresentaram um método simples para encontrar frases no texto e mostrar que aprender representações vetoriais boas para milhões de frases é possível.

A técnica empregada consiste na aplicação recursiva do algoritmo sobre o corpus de forma a que o primeiro ciclo identifique bigramas, o segundo identifique trigramas/quadrigramas o terceiro identifique quadrigramas/pentagramas e assim sucessivamente.

Ao final do processamento, foram identificados 48.899 bigramas/trigramas no corpus de treinamento. Esses ngramas serão tratados, durante o treinamento do modelo, como um único termo. O quadro abaixo traz uma amostra dos mais recorrentes:

Ngrama	Ocorrências
agravo_interno	252.167
agravo_regimental	245.257
recurso_especial	189.284
agravo_recurso	146.683
embargo_declaracao	138.958
processual_civil	132.195
agravo_recurso_especial	127.220
habeas_corpus	122.068
sumula_stj	114.436
codigo_processo	87.685

agravo_interno_agravo_recurso	61.428
agravo_regimental_agravo_recurso	51.877
incidencia_sumula	47.504
fundamento_decisao	44.505
desta_corte	43.155
codigo_processo_civil	43.110
superior_tribunal	41.723
ausencia_impugnacao	41.305
tribunal_origem	37.695
embargo_declaracao_rejeitado	37.249
sumula_stf	35.681
habeas_corpus_conhecido	35.477
prisao_preventiva	34.607
supremo_tribunal	34.350
tribunal_justica	34.023
recurso_ordinario	30.432
agravo_regimental_provido	28.853
agravo_interno_provido	27.883
agravo_regimental_desprovido	25.946
tribunal_federal	24.968
agravo_regimental_improvido	24.613
especifica_fundamento	23.032
incidencia_sumula_stj	22.781
agravo_regimental_recurso_especial	21.719
fatico_probatorio	21.224
ausencia_impugnacao_especifica_fundamento	21.195
constrangimento_ilegal	21.150
agravo_interno_improvido	21.046
ordem_concedida	21.043
agravo_interno_provimento	20.806
ordem_publica	20.450
substitutivo_recurso	20.274
agravo_regimental_provimento	19.678
processual_penal	19.420
especificamente_fundamento	18.244
habeas_corpus_substitutivo_recurso	17.884
trafico_droga	17.443
superior_tribunal_justica	17.123
jurisprudencia_desta	17.050
agravo_regimental_conhecido	16.948
codigo_penal	16.807
processo_penal	16.324
corte_superior	16.228
recurso_extraordinario	16.113
recurso_ordinario_habeas_corpus	15.897

codigo_processo_penal	15.328
embargo_declaracao_agravo_regimental	15.044
processo_civil	14.742
desconstituir_decisao	14.375
recurso_interposto	14.361
obice_sumula	14.350
grafo_unico	14.127
corte_origem	14.050
repercussao_geral	13.955
instancia_ordinaria	13.656
caso_auto	13.606
decisao_recorrida	13.141
razao_agravo	13.034
tribunal_quo	12.886
mandado_seguranca	12.464
servidor_publico	12.367
garantia_ordem	12.270
embargo_declaracao_agravo_interno	12.214

Tabela 3. Amostra dos ngramas identificados no corpus de treinamento.

Com a aplicação dos ngramas, o texto do exemplo recebeu a constituição abaixo:

> administrativo_processual agravo_regimental_mandado_seguranca processo_administrativo_disciplinar pena_demissao razao improbidade operacao euterpe alegacao_violacao devido_processo_legal parcialidade comprovada legitima utilizacao_prova emprestada ausencia reformatio_peju indeferimento_liminar mantido seguranca_denegada historico_demanda trata procedimento_administrativo_disciplinar resultou demissao ibama razao improbidade_administrativa consta relatorio_final comissao_processante desmantelou policia_federal operacao euterpe mundo criminoso instalado ambito meio_ambiente atingindo cerne quadrilha nucleo interno formado vario segmento servidor ibama precipuamente fiscal tecnico externo valia poder financeiro proteger interesse ilegal consta relatorio_final comissao_processante descricao_conduta investigada deram origem_presente impetracao investigado leonardo mantem forte relacao extra funcional empresario area construcao consultoria area meio_ambiente maioria conversa existente degravada relatorio identifica pratica criminal ocorrida ambos erifica ocorrencia fiscalizacao parte ibama obra interlocutor nesta ligacao conversa giram entorno valor_pode pagar conversa leonardo data informa pessoa bateu pe valor doil mil_real depositado_conta atrave cheque conversa isidoro leonardo serve intermediario fiscal ibama alguma obra fiscalizada pessoal ibama empresario ligam descobre fiscal seguida liga leonardo efetiva contato parte leonardo indica servico empresa fiscaliza caso construtora ontra leonardo liga data informa pessoal ontra vai procura nome restou_comprovado apurado acusado leonardo edward rose respeitou norma_legal atuando acordo interesse desleal instituicao arredio atribuicao cargo resultou demonstrado apurado forte consistente comprovacao acusado leonardo edward rose associado conduta obteve vantagem_pecuniaria ilicita detrimento dignidade funcao_publica utilizando condicao servidor_publico precedente julgamento curso primeira julgou doil mandado_seguranca semelhante concedendo seguranca trata relatoria relatoria haroldo rodrigue vez curso julgamento_relatoria apo voto denegando seguranca abriu

divergencia acompanhado tendo pedido_vista respeitosamente divirjo entendimento_adotado comungando pensamento expressado ausencia influencia carlo minc processo prova influencia externa processo carlo minc efetivamente responsavel denuncia caso corrupcao deputado estadual inicio processo_administrativo nomeou primeira comissao_processante competente anular pad apo relatorio comissao_processante ato competia levado efeito presidente ibama fundamento consultivo federal presidente nomeado comissao_processante exercia influencia perniciosa dado presidente comissao_processante federal submetido geral_uniao assinou portaria demissao atuacao federal elielson ayr inexistente vicio nomeacao federal elielson ayr presidente comissao_processante sugestao consultivo procuradoria federal batista presidente ibama determinou nulidade parcial processo_termo lei constituiu nova comissao_processante nomeou referido federal presidente analise exclusiva vicio_processual insanavel recomendou nulidade total fragmentacao portaria instauracao_pad acolhido ausente irregularidade procedimento imparcialidade meio_ambiente provou parcialidade meio_ambiente ter supostamente seguidora carlo minc decisao amparada minucioso relatorio_final comissao_processante lauda apo_ano colheita material_probatorio prova penal emprestada doutrina_jurisprudencia favoravel prova_emprestada respeitado contraditorio_ampla_defesa ambito_processo administrativo_disciplinar cujo traslado prova penal antecedido autorizado juizo_criminal dispensavel averiguar decisao_juizo vara_federal meriti dando incompetente causaria nulidade_prova produzida compartilhamento efetuado tribunal_regional_federal_regiao deu_provimento recurso_sentido_estrito ministerio_publico_federal dando competencia_juizo deferiu acompanhou interceptacao_telefonica prisao_temporaria preventiva busca_apreensao recebeu_denuncia interrogou reu eminente nefi negou_seguimento_recurso_especial interposto rse tendo stj negado_provimento agravo_regimental ausencia reformatio_peju ocorreu reformatio_peju relatorio_final produzido ratificado carece natureza vinculante destacada apena_decisao sentido tecnico primeira anulacao_processo administrativo juizo_previo ilicito prolatada presidente ibama ulterior decisum demissao proferido meio_ambiente cotejo ambos extrai alegada_nulidade conclusao agravo_regimental_provido

Todos os documentos do corpus, após o pré-processamento, foram armazenados em um arquivo único. Esse arquivo será utilizado para treinamento do modelo de Inteligência Artificial.

3.2 – A escolha do algoritmo de treinamento

Há na literatura diversas abordagens para construção de modelos de inteligência artificial voltadas ao processamento de linguagem natural. Não é parte do escopo deste experimento a comparação entre eles, uma vez que, de forma geral, os exemplos mais recentes e que apresentam maior acurácia, com objetivo semelhante ao proposto neste experimento, convergem para a aplicação de vetores de palavras ou seus derivados. Por esse motivo, a opção neste experimento foi pelo *Paragraph Vector*.

3.2.1 – O *framework* GenSim[9]:

Não é objetivo do trabalho a implementação de algoritmos, por conta disso foi selecionado o framework Gensim que é voltado ao processamento de linguagem natural. A ferramenta implementa para a linguagem Python[10] os mais populares algoritmos da área, em especial a classe Doc2Vec, que traz uma implementação do *Paragraph Vector* de Le & Mikolov.

3.3 – O treinamento do modelo

Concluída a parte de Pré-processamento e realizada a escolha do algoritmo para criação do modelo, inicia-se a fase de treinamento propriamente dita. O treinamento constitui-se na submissão do corpus pré-processado ao algoritmo selecionado que, obedecendo a uma série de parâmetros, gera o modelo.

3.3.1 – Estabelecimento de Parâmetros:

O *framework* GenSim disponibiliza uma série de parâmetros em sua classe Doc2Vec, que implementa o algoritmo *Paragraph Vector*.

O experimento utilizou o algoritmo *Distributed Memory Model of Paragraph Vectors* (PV-DM). neste modelo, segundo ensinam Cassiano & Cordeiro (2018), cada parágrafo é mapeado para um vetor exclusivo, representado por uma coluna em uma matriz D. Cada palavra também e mapeada para um vetor exclusivo, representado por uma coluna em uma matriz W. A concatenação ou média do vetor de parágrafo com os vetores de palavras são utilizados para prever a próxima palavra em um contexto. O vetor de parágrafo pode ser considerado uma pseudo-palavra e representa as informações que faltam no contexto atual, atuando como uma memória do tópico do parágrafo em questão, que preserva a posição dos termos ao trabalhá-los.

Foi estabelecido um vetor de saída com duzentas dimensões. A experiência indica que quanto maior o número de dimensões, mais genérico é o modelo, entretanto valores muito superiores podem gerar dispersões nos resultados.

---

[9] Vide https://radimrehurek.com/gensim/index.html
[10] Vide https://www.python.org/

O aprendizado foi parametrizado para considerar uma janela de 15 termos, que corresponde à média de termos em uma frase do corpus. Esse valor define que serão considerados para cada termo alvo os 15 termos mais próximos (antecedentes ou seguintes).

O número mínimo de ocorrências de cada termo no corpus para que ele seja considerado durante o treinamento foi fixado em 50, o objetivo disso é apenas descartar termos pouco recorrentes e que não agregam valor ao modelo.

O processamento do modelo demorou 8h33min em um computador com processador I7 de 7ª Geração e 16Gb de memória RAM. Foi gerado um dicionário com 35.120 termos únicos.

Capítulo 4 - Aplicação do Modelo

Este capítulo demonstra o resultado da aplicação do modelo de IA sobre um conjunto de documentos jurídicos. Apresenta também as avaliações dos agrupamentos realizados por um especialista.

4.1 – O corpus de teste

O Corpus de teste é composto por uma amostra de 1.133 acórdãos proferidos pelo Tribunal Regional Federal da 3ª Região (TRF3) cujos recursos chegaram ao STJ no primeiro semestre de 2018. Cada documento foi identificado pelo seu respectivo número de registro no STJ.

A seleção da amostra baseou-se em dois critérios:

a) A amostra deve ser composta por documentos que tenham variabilidade suficiente para garantir que não tratem de controvérsia idêntica;
b) Entre alguns dos documentos exista proximidade semântica suficiente para classificá-los em um mesmo grupo.

4.1.1 – O pré-processamento:

Os documentos do corpus foram submetidos às mesmas atividades de pré-processamento descritos nos itens 3.1.2 e 3.1.3. Adicionalmente, foram removidos todos os termos que não compunham o dicionário do modelo treinado.

Observe-se que, conforme modelo abaixo, o documento extraído da base de dados do SIAJ continha originalmente uma série de ruídos:

> (e-STJ Fl.56) PODER JUDICIÁRIO : DESEMBARGADOR FEDERAL JOSÉ ANTONIO NEIVA AGRAVANTE INSTRUMENTO 2013.02.01.008385-8 : : 0008385-85.2013.4.02.0000 FUNDACAO HABITACIONAL DO EXERCITO-FHE : SEBASTIAO ZIMERMAN E OUTROS : MARCIO PEDRO DECISÃO instrumento interposto pela FUNDAÇÃO : JUREMA ALVES DO NASCIMENTO ALMA WI ORIGEM ADVOGADO AGRAVADO ADVOGADO RELATOR N° CNJ III - AGRAVO DE TRIBUNAL REGIONAL FEDERAL DA 2' REGIÃO Trata-se (200851010024757) - FHE visando à reforma da decisão proferidaHABITACIONAL DO EXÉRCITO de agravo de : VIGÉSIMA TERCEIRA VARA FEDERAL DO RIO DE JANEIRO pelo juízo da 23a Vara Federal do Rio de Janeiro, nos autos do processo n° 2008.51.01.002475-7, que indeferiu o pedido de consignação em folha de pagamento na razão de 30% dos vencimentos do executado, ora agravado. Sustenta a agravante, em apertada síntese, que o contrato em questão prevê que as parcelas mensais seriam pagas mediante consignação em folha, conforme documento acostado na execução. Aduz que realizou uma série de diligências para satisfazer seu crédito, porém sem sucesso. Preconiza o art. 649 do CPC que são absolutamente impenhoráveis: IV - os "(..) É o breve relato. Decido. vencimentos, subsídios, soldos, salários, remunerações, proventos de aposentadoria, pensões, pecúlios e montepios; as quantias recebidas por liberalidade de terceiro e destinadas ao sustento do honorários de profissional liberal (...); X - até o limite de 40 (quarenta) salários

mínimos, a quantia depositada em caderneta de poupança". mvb devedor e sua família, os ganhos de trabalhador autônomo e os 1 Documento recebido eletronicamente da origem (e-STJ Fl.57) III - AGRAVO DE TRIBUNAL REGIONAL FEDERAL DA 2' REGIÃO 2013.02.01.008385-8 >A I Olyq F salário do executado (na verdade, a conta bancária em que recebe a sua remuneração). INSTRUMENTO PODER JUDICIÁRIO de Justiça: Assim, não pode ser objeto de constrição nem de bloqueio remuneração ou FISCAL - CRÉDITO TRIBUTÁRIO - BLOQUEIO DE ATIVOS FINANCEIROS POR MEIO DO SISTEMA BACEN JUD APLICAÇÃO CONJUGADA DO ART. 185-A, DO CTN, ART. 11, DA ART. 655-A, A propósito do tema, vale conferir o seguinte julgado do E. Superior Tribunal EXECUÇÃO - DO CPC. PROPORCIONALIDADE NA EXECUÇÃO. LIMITES DOS ARTS. 649, IV e 620 DO CPC. E 655 ART. 6.830/80, N. LEI "PROCESSUAL CIVIL - RECURSO ESPECIAL - 1. Não incide em violação do art. 535 do CPC o acórdão que decide fazendo uso de argumentos suficientes para sustentar a sua tese. O julgador não é obrigado a se manifestar sobre todos os dispositivos legais levados à discussão pelas partes. 2. A interpretação das alterações efetuadas no CPC não pode resultar no credor público, principalmente no que diz respeito à cobrança do crédito tributário, que deriva do dever fundamental de pagar tributos (artigos 145 e seguintes da Constituição Federal de 1988). 3. Em interpretação sistemática do ordenamento jurídico, na busca de uma maior eficácia material do provimento jurisdicional, absurdo lógico de colocar o credor privado em situação melhor que o deve-se conjugar o art. 185-A, do CTN, com o art. 11 da Lei n. 6.830/80 e artigos 655 e 655-A, do CPC, para possibilitar a penhora de dinheiro em depósito ou aplicação financeira, independentemente do esgotamento de diligências para encontrar outros bens penhoráveis. Em suma, para as decisões proferidas a partir de 20.1.2007 (data da 2 .ic - _/L tributário ou não, aplica-se o disposto no art. 655-A do Código de Processo Civil, posto que compatível com o art. 185-A do CTN. 4. A aplicação da regra não deve descuidar do disposto na nova redação do art. 649, IV, do CPC, que estabelece a impenhorabilidade dos valores mvb entrada em vigor da Lei n. 11.038/2006), em execução fiscal por crédito Documento recebido eletronicamente da origem (e-STJ Fl.58) Il III - AGRAVO DE TRIBUNAL REGIONAL FEDERAL DA 28 REGIÃO 2013.02.01.008385-8 INSTRUMENTO PODER JUDICIÁRIO proventos de aposentadoria, pensões, pecúlios e montepios; às quantias devedor e sua família, aos ganhos de trabalhador autônomo e aos honorários de profissional liberal. ser sempre observado o recebidas por liberalidade de terceiro e destinadas ao sustento do princípio ferramenta, devendo 1074228/MG, Rel. Ministro MAURO CAMPBELL proporcionalidade na execução (art. 620 do CPC) sem descurar de sua finalidade (art. 612 do CPC), de modo a não inviabilizar o exercício da atividade empresarial." 6. Recurso especial parcialmente conhecido e, nessa parte, provido. MARQUES, SEGUNDA TURMA, julgado em 07/10/2008, referentes aos vencimentos, subsídios, soldos, salários, remunerações, da DJe 05/11/2008). REsp (STJ, 5. Também há que se ressaltar a necessária prudência no uso da nova Ainda que o contrato firmado entre a recorrente e o agravado autorize a consignação em folha de pagamento para o resgate das prestações acordadas, caberia a ela (agravante), no entanto, providenciar, junto à entidade a que a

recorrida estava vinculada quando da formalização do ajuste, a efetivação do ou então demonstrar que a cessação deu-se de forma unilateral, em descompasso com a jurisprudência do Superior Tribunal de Justiça acerca do tema, no sentido de que aquela medida é perfeitamente cabível e não configuraria penhora de vencimentos. desconto contratualmente previsto, fato este não comprovado nos presentes autos, Nesse sentido, vale conferir, a título exemplificativo, os seguintes julgados: CLÁUSULA CONTRATUAL. DESCONTO EM FOLHA DE PAGAMENTO. VALIDADE. AUSÊNCIA DE ABUSIVIDADE. DE PENHORA "EMBARGOS DE DIVERGÊNCIA. EMPRÉSTIMO BANCÁRIO. SUPRESSÃO UNILATERAL. IMPOSSIBILIDADE. que a cláusula que prevê, em contratos de empréstimo, o desconto em folha de pagamento, não configura a penhora vedada pelo art. 649, IV, mvb 1. A Segunda Seção desta Corte tem posição consolidada no sentido de 3 L--- L : VENCIMENTO. NÃO CONFIGURAÇÃO. Documento recebido eletronicamente da origem (e-STJ Fl.59) PODER JUDICIÁRIO - III AGRAVO DE INSTRUMENTO 2013.02.01.008385-8 do CPC, nem encerra qualquer abusividade, não podendo, em princípio, ser alterada unilateralmente, porque é circunstância especial para facilitar o crédito. TRIBUNAL REGIONAL FEDERAL DA 2' REGIÃO 2. Embargos de divergência acolhidos." (STJ, EREsp 537.145/RS, Rel. Ministro FERNANDO GONÇALVES, "AGRAVO REGIMENTAL. RECURSO ESPECIAL. CONTRATO VERBETES BANCÁRIO. INOCORRÊNCIA DE VIOLAÇÃO DO ARTIGO 535 DO CPC. ILEGITIMIDADE DA INSTITUIÇÃO FINANCEIRA. CAPITALIZAÇÃO 7/STJ. DESCONTO EM FOLHA. VALIDADE. 7/STJ. E 5 N.° MENSAL DOS JUROS. AGRAVO PARCIALMENTE PROVIDO. SUMULARES SÚMULA SEGUNDA SEÇÃO, julgado em 26/09/2007, DJ 11/10/2007 p. 285) acórdão recorrido aprecia a questão de maneira fundamentada. O julgador não é obrigado a manifestar-se acerca de todos os argumentos apontados pelas partes, se já tiver motivos suficientes para fundamentar sua decisão. 1. Não há violação do artigo 535 do Código de Processo Civil quando o 2. Necessidade de reexame fático-probatório para aferir responsabilidade do recorrente em proceder a suspensão do desconto em folha de pagamento, quando as instâncias ordinárias afirmam a existência de convênio, cabendo ao Banco operacionalizar todas as providências atinentes ao financiamento. 3. A cláusula que prevê, em contratos de empréstimo, o desconto em folha de pagamento, não encerra qualquer abusividade, não podendo, em princípio, ser alterada unilateralmente, porque é circunstância especial para facilitar o crédito. capitalização de juros, nem, tampouco, da data em que foi celebrado o c mvb Justiça. 4. As instâncias ordinárias não se manifestaram acerca da pactuação da _ - - verificação de tais requisitos, sob pena de afrontar o disposto nos enunciados sumulares nos 5 e 7 da Súmula do Superior Tribunal de ` contrato, o que impossibilita, nesta esfera recursal extraordinária a 4 Documento recebido eletronicamente da origem (e-STJ Fl.60) PODER JUDICIÁRIO TRIBUNAL REGIONAL FEDERAL DA 2' REGIÃO - III AGRAVO DE INSTRUMENTO 2013.02.01.008385-8 ao desconto em folha de pagamento." em 09/10/2007, penhora de salário, o que é vedado pelo art. 649, IV do CPC. 15/04/2013; TRF2, AG TRF2, AG TRF2, AG Nego seguimento ao agravo de instrumento nos termos do art. 557, caput, do --- 5 `- (STJ, AgRg no REsp 733.716/RS, Rel. Ministro HÉLIO QUAQLIA BARBOSA, QUARTA TURMA, julgado

> 29/10/2007.) 5. Agravo parcialmente provido para declarar válida a cláusula atinente DJ A ausência de comprovação dos fatos acima descritos acerca do desconto em folha de pagamento, comprovação essa que se mostra fundamental para análise das alegações recursais, inviabiliza o acolhimento do recurso interposto. Embora tenha o agravado autorizado a consignação em folha de pagamento na celebração do contrato de empréstimo, isto se deu na fase e para efeitos extrajudiciais, e respeitados os limites legais de consignação. Já agora, o desconto requerido pela agravante se dá para fins de execução judicial, e consiste, pois, em Nesta linha, merecem destaque os seguintes arestos desta e. Corte: TRF2, AG 201302010025111, Desembargador Federal GUILHERME COUTO, SEXTA TURMA ESPECIALIZADA, E-DJF2R 201202010155348, Desembargador Federal POUL ERIK DYRLUND, OITAVA TURMA ESPECIALIZADA, E-DJF2R 24/01/2013; 201202010022671, Desembargador Federal MARCELO PEREIRA DA SILVA, QUINTA TURMA ESPECIALIZADA, E-DJF2R 31/08/2012; 201202010021721, Desembargador Federal LUIZ PAULO DA SILVA ARAUJO Dessa forma, uma vez constatada a impenhorabilidade das verbas salariais e não comprovadas as alegações acerca do desconto em folha, em especial o início e eventual cessação indevida, impõe-se a negativa de seguimento do presente mvb CPC. Isto posto, recurso. FILHO, SÉTIMA TURMA ESPECIALIZADA, E-DJF2R 19/06/2012. Documento recebido eletronicamente da origem (e-STJ Fl.61) III - AGRAVO DE TRIBUNAL REGIONAL FEDERAL DA 2° REGIÃO 2013.02.01.008385-8 Feitas as anotações de estilo, baixem os autos à vara de origem para P.I. arquivamento. INSTRUMENTO PODER JUDICIÁRIO s Rio de Janeiro, 18 de junho de 2013. t r ' - Relator JOSÉ ANTONIO LISBOA NEIVA Desembargador Federal mvb 6 Documento recebido eletronicamente da origem 1

Os ruídos derivam da submissão ao Reconhecimento Ótico de Caracteres (OCR), que acabou gerando um documento permeado de termos truncados, alguns deles sem sentido. Parte destes ruídos deve desaparecer com a padronização do texto.

Após o pré-processamento, que foi executado de forma idêntica à descrita nos itens 3.1.2 e 3.1.3. O documento teve de seu corpo extraídos os termos que não compunham o dicionário do modelo. Isso resultou em um documento mais enxuto, conforme observamos abaixo.

> poder_judiciario desembargador federal instrumento fundacao habitacional exercito decisao instrumento_interposto fundacao alve origem cnj agravo tribunal_regional_federal_regiao trata visando reforma_decisao exercito agravo terceira vara_federal juizo_vara federal_auto processo indeferiu_pedido consignacao folha_pagamento razao vencimento executado sustenta sintese contrato questao preve parcela mensal paga consignacao folha documento_acostado execucao aduz realizou serie diligencia satisfazer credito sucesso preconiza cpc impenhoravel relato decido

vencimento subsidio soldo salario remuneracao provento_aposentadoria pensao peculio quantia recebida liberalidade terceiro destinada sustento honorario profissional liberal limite salario_minimo quantia_depositada poupanca devedor familia ganho trabalhador autonomo documento recebido eletronicamente origem_agravo tribunal_regional_federal_regiao salario executado instrumento poder_judiciario justica_pode objeto constricao bloqueio remuneracao fiscal credito_tributario bloqueio ativo_financeiro meio sistema bacen aplicacao conjugada ctn proposito tema vale conferir seguinte_julgado superior_tribunal execucao cpc proporcionalidade execucao limite cpc_lei processual_civil recurso_especial incide violacao_cpc acordao_decide fazendo uso argumento_suficiente sustentar_tese julgador_obrigado manifestar_dispositivo legal levado discussao parte interpretacao alteracao efetuada cpc_pode resultar credor publico principalmente diz_respeito cobranca_credito tributario deriva dever fundamental pagar tributo interpretacao_sistematica ordenamento_juridico busca maior eficacia material provimento_jurisdicional logico colocar credor privado situacao deve ctn_lei artigo_cpc possibilitar penhora_dinheiro deposito aplicacao_financeira independentemente esgotamento_diligencia encontrar penhoravel suma decisao_proferida partir execucao_fiscal credito documento recebido eletronicamente origem_agravo tribunal_regional_federal_regiao instrumento poder_judiciario provento_aposentadoria pensao peculio quantia devedor familia ganho trabalhador autonomo honorario profissional liberal observado recebida liberalidade terceiro destinada sustento principio ferramenta devendo proporcionalidade execucao descurar finalidade modo inviabilizar exercicio_atividade empresarial recurso_especial_conhecido_nessa parte_provido marque julgado referente vencimento subsidio soldo salario remuneracao stj ressaltar necessaria prudencia uso nova contrato_firmado autorize consignacao folha_pagamento resgate prestacao acordada caberia providenciar entidade recorrida vinculada formalizacao ajuste efetivacao demonstrar cessacao deu_forma unilateral descompasso jurisprudencia_superior_tribunal_justica tema sentido medida cabivel configuraria penhora vencimento desconto contratualmente previsto fato comprovado presente_auto sentido vale conferir titulo exemplificativo seguinte_julgado clausula_contratual desconto_folha_pagamento validade ausencia abusividade penhora embargo_divergencia emprestimo bancario supressao unilateral impossibilidade clausula preve contrato_emprestimo desconto_folha_pagamento configura penhora vedada desta_corte posicao consolidada_sentido vencimento configuracao documento recebido eletronicamente origem poder_judiciario agravo_instrumento cpc encerra abusividade podendo principio alterada unilateralmente circunstancia especial facilitar credito tribunal_regional_federal_regiao embargo_divergencia acolhido acordao aprecia questao maneira_fundamentada julgador_obrigado manifestar argumento apontado parte motivo_suficiente fundamentar_decisao violacao_artigo_codigo_processo civil necessidade_reexame_fatico_probatorio aferir responsabilidade proceder suspensao desconto_folha_pagamento instancia_ordinaria afirmam existencia convenio cabendo banco providencia atinente financiamento clausula preve contrato_emprestimo desconto_folha_pagamento encerra abusividade podendo principio

alterada unilateralmente circunstancia especial facilitar credito capitalizacao_juro data celebrado justica instancia_ordinaria manifestaram pactuacao verificacao_requisito pena afrontar disposto_enunciado sumular sumula_superior tribunal contrato impossibilita nesta esfera recursal extraordinaria documento recebido eletronicamente origem poder_judiciario tribunal_regional_federal_regiao agravo_instrumento desconto_folha_pagamento penhora salario vedado cpc trf ag trf ag trf ag seguimento_agravo instrumento termo caput agravo_provido declarar valida clausula atinente ausencia_comprovacao fato_descrito desconto_folha_pagamento comprovacao mostra fundamental analise_alegacao recursal inviabiliza acolhimento_recurso interposto autorizado consignacao folha_pagamento celebracao_contrato emprestimo deu fase efeito extrajudicial respeitado limite_legal consignacao desconto fim execucao judicial consiste poil nesta linha merecem destaque seguinte aresto desta_corte trf ag desembargador federal guilherme especializada desembargador federal especializada desembargador federal especializada desembargador federal dessa_forma vez constatada impenhorabilidade verba_salarial comprovada alegacao desconto_folha especial inicio eventual cessacao indevida impoe negativa_seguimento presente cpc posto recurso especializada documento recebido eletronicamente origem_agravo tribunal_regional_federal_regiao feita anotacao auto vara origem arquivamento instrumento poder_judiciario junho desembargador federal documento recebido eletronicamente origem

Ao aplicar o modelo de Inteligência Artificial sobre os documentos, foram extraídos para cada um deles um vetor de 200 dimensões, conforme exposto abaixo:

-0.7217684388160706	-3.238701343536377	-9.480405807495117	1.0403672456741333
-3.0265471935272217	-6.043298721313477	-1.1627678871154785	3.936262845993042
-3.9727420806884766	0.9615909457206726	3.83803391456604	0.7638435959815979
2.3155009746551514	3.8920581340789795	-0.5999000668525696	2.753196954727173
8.847442626953125	7.651838779449463	-11.046567916870117	-4.317355155944824
-1.3314383029937744	-4.461985111236572	-2.670752763748169	-0.23997503519058228
6.125977516174316	1.810759425163269	-4.670225620269775	-0.10512460023164749
-2.538372039794922	-0.7031194567680359	1.3011951446533203	2.7550127506256104
-4.031469345092773	-0.6254139542579651	3.726902484893799	-2.2638444900512695
6.1773529052734375	-2.9896066188812256	-2.5951364040374756	-0.6882867813110352
3.5172958374023438	-12.66833209991455	-1.9690730571746826	-1.1374465227127075
-0.0830293595790863	4.855254173278809	-1.4020506143569946	-9.822463035583496
-0.5784963369369507	0.645620584487915	-4.9133710861206055	5.712505340576172
-0.6567866206169128	-2.6001484394073486	-4.015826225280762	1.6368789672851562
0.20575034618377686	-0.6669518947601318	0.2398321032524109	-0.4517885744571686
-0.20405727624893188	-1.4270719289779663	-6.653496742248535	2.857419967651367
0.8329592943191528	-3.086744546890259	-6.231448173522949	0.46599000692367554
3.7010223865509033	0.8851044178009033	-5.723743915557861	7.8737382888793945
8.957752227783203	0.7259459495544434	-1.323415756225586	-7.654653072357178
-2.4044618606567383	-0.1189049556851387	1.2612847089767456	-3.6721863746643066
2.5534467697143555	3.096687078475952	-5.292621612548828	-3.1905643939971924
-1.9133899211883545	1.8541370630264282	-5.068018436431885	-1.6663639545440674
-2.9355804920196533	1.2225030660629272	-1.7277363538742065	-4.377777099609375
-1.7290756702423096	-4.414225101470947	-6.328978538513184	-10.171880722045898
0.40061795711517334	-1.4186832904815674	7.494574069976807	3.0052433013916016
-4.83843994140625	10.022359848022461	6.580016613006592	2.656548500061035
-0.6985257863998413	-5.607921123504639	5.134721279144287	-4.176854133605957
3.621781826019287	-0.08601149916648865	4.1844635009765625	-4.602542400360107
5.181891441345215	-0.29307955503463745	-2.004587411880493	-0.0055449046194553375
4.569941997528076	-2.740199089050293	5.8810930252075195	-3.9744715690612793

-2.5121004581451416	2.356847047805786	-1.0859638452529907	8.06953239440918
9.248960494995117	-2.0956788063049316	8.862290382385254	-1.8475996255874634
1.5420610904693604	0.9036908149719238	-4.633780002593994	3.112339496612549
-5.972724914550781	2.8296937942504883	-0.6607450842857361	-3.0012454986572266
5.588530540466309	3.304886817932129	-0.9052466750144958	-4.842007160186768
-2.2270569801330566	-3.4605720043182373	-0.4605211615562439	0.6399190425872803
-1.6431432962417603	-3.7089924812316895	-2.609949827194214	2.423452377319336
1.6322895288467407	7.618996620178223	-7.271684646606445	-1.9013117551803589
-0.9252272248268127	6.003765106201172	0.3201490044593811	8.959856033325195
0.6245195865631104	4.841502666473389	-0.14068900048732758	4.708121299743652
-0.768897533416748	7.247696399688721	-0.7912754416465759	-5.3965559005737305
4.392187595367432	-0.4206462800502777	2.678196907043457	0.3834846615791321
-7.5970940589904785	-5.389682292938232	-3.8214097023010254	6.116034984588623
2.4382026195526123	3.34684419631958	3.0661070346832275	1.598451018333435
-1.4711161851882935	7.1295881271362305	0.9859408140182495	4.95748233795166
4.403463363647461	-1.956924557685852	1.2952207326889038	2.186674118041992
-1.6629129648208618	-2.4989571571350098	-0.21622861921787262	0.036035552620887756
-6.577131748199463	1.4660937786102295	-0.804520845413208	-4.893164157867432
-0.5389730930328369	1.3079426288604736	5.188241958618164	3.320317029953003
-4.512632846832275	1.0352330207824707	2.9813313484191895	0.5604121088981628

Tabela 4. Vetor de um documento do corpus de teste.

As mesmas atividades foram repetidas para todos os documentos do corpus.

Essa atividade consumiu um tempo total de 8’30’’.

4.2 - A redução de dimensionalidade

Segundo DIAS (2012), em muitos problemas, o conjunto de dados apresenta muitos atributos, isto é, possuem uma dimensionalidade alta. No entanto, há motivos para se acreditar que os dados pertençam a uma variedade de baixa dimensão, isto é, características ou parâmetros podem ser dependentes entre si, ou em outras palavras, a dimensionalidade dos dados pode ser maior do que a necessária, como no caso de variáveis altamente correlacionadas. Assim, o grande conjunto de parâmetros poderia ser reduzido a um conjunto menor, seja pela eliminação de variáveis consideradas irrelevantes ou pela transformação das variáveis, aliada a preservação de características importantes do conjunto de dados inicial. Em todos os casos, reduzir a dimensionalidade dos dados resulta em uma representação mais eficiente dos mesmos.

No experimento, o vetor gerado pelo modelo possui 200 dimensões. Esse número de dimensões acaba por impedir o agrupamento utilizando o algoritmo DBSCAN, que se encarregará da clusterização de documentos. Por conta disso, foi utilizada uma técnica de redução de dimensionalidade chamada *Principal Component Analisis* (PCA). Segundo Dias (2012), a ideia central do PCA é reduzir a dimensionalidade de um conjunto de dados no qual existe um grande número de variáveis que estão relacionadas entre si, ao mesmo tempo que procura reter ao máximo a variação presente no conjunto de dados. Esta redução é alcançada pela transformação das variáveis em um novo conjunto, as componentes principais, as quais são descorrelacionadas, e ordenadas a fim de que as primeiras componentes principais retenham o máximo da variação presente em todas as variáveis originais.

Após a redução da dimensionalidade, é possível plotar os documentos em um plano de duas dimensões. Isso nos permite observar a distribuição dos documentos no espaço.

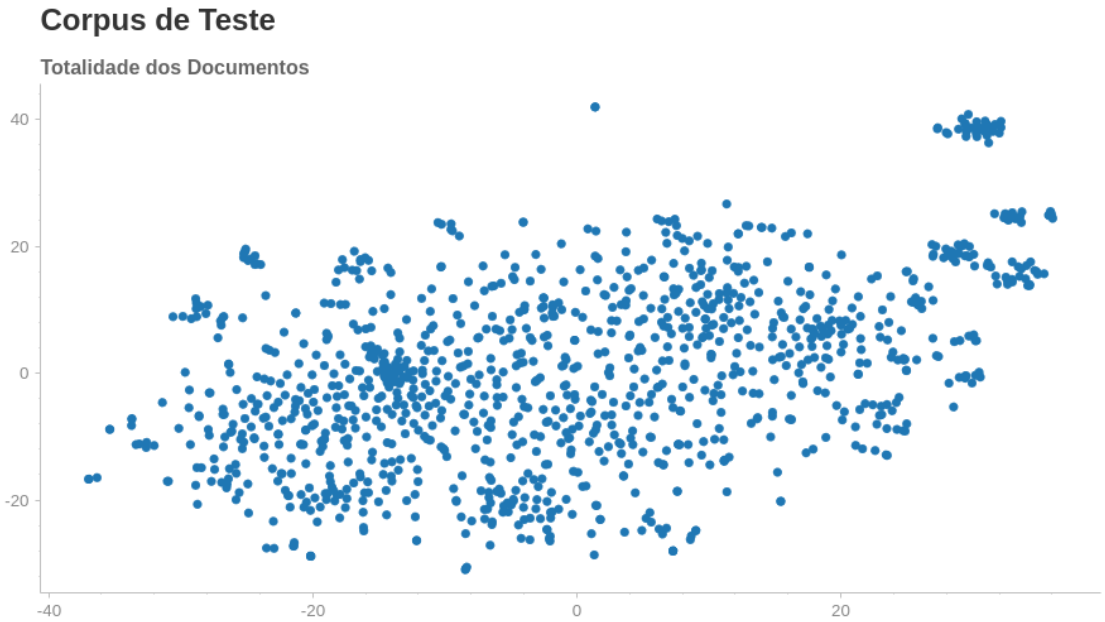

Gráfico 5. Distribuição dos documentos do corpus de teste em um espaço bidimensional.

Essa atividade consumiu um tempo total de 1’.

4.3 - O agrupamento

Após a redução do número de dimensões de 200 para 2, o corpus de teste foi submetido a agrupamento utilizando o algoritmo DBSCAN, abreviação do termo ‘Density Based Spatial Clustering of Application with Noise’ (Agrupamento Espacial Baseado em Densidade de Aplicação com Ruído). Inicialmente proposto por Ester et al (1996), o DBSCAN é destinado à clusterização baseada em densidade, sendo capaz de identificar de forma automática, sem intervenção, grupos (clusters) de objetos em espaços de duas ou três dimensões.

A ideia chave do método DBSCAN é que, para cada ponto de um cluster, a vizinhança para um dado raio contém, no mínimo, certo número de pontos, ou seja, a densidade na vizinhança tem que exceder um limiar. (CASSIANO, 2014).

No experimento, o algoritmo foi ajustado para operar com os liminares (distância máxima entre dois integrantes do grupo, calculada pelo cosseno) iguais a 0.30, 0.35, 0.40, 0.45 e 0.50 e pelo menos 2 integrantes em cada grupo.

Essa atividade consumiu um tempo total de 30’’ para cada agrupamento.

O resultado do agrupamento com o uso do DBSCAN sobre os documentos com os limiares informados gerou os seguintes grupos:

Com o limiar de 0.30 foram gerados 125 grupos que abarcaram 262 documentos, com uma cobertura de 23,12% da amostra. O maior grupo continha 3 elementos.

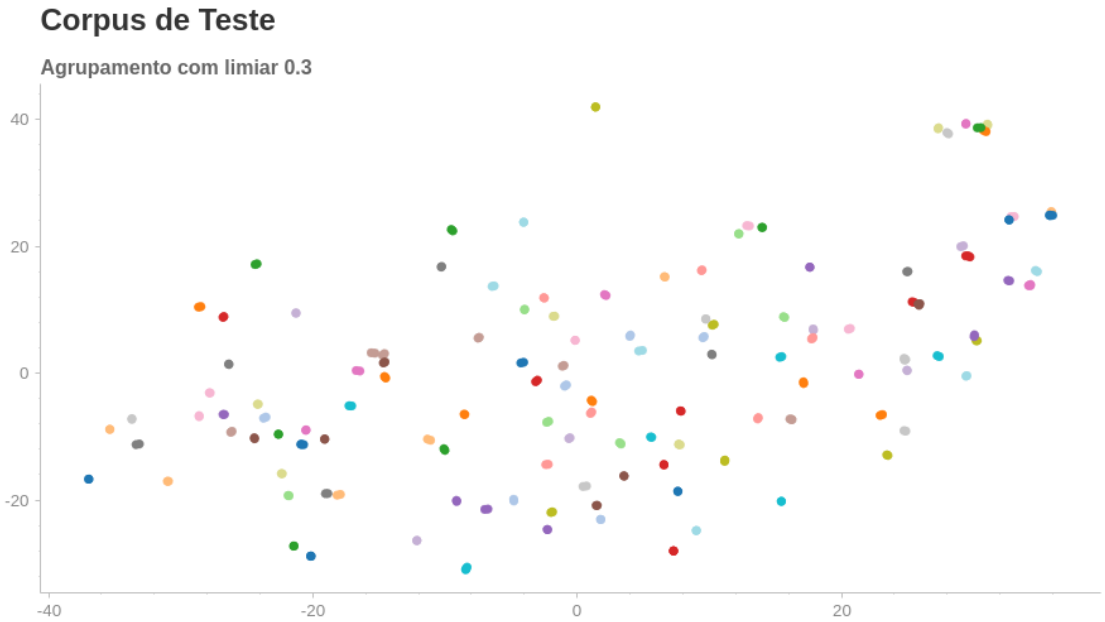

Gráfico 6. Agrupamentos com limiar de 0.30.

Com o limiar de 0.35 foram gerados 147 grupos que abarcaram 313 documentos, com uma cobertura de 27,62% da amostra. O maior grupo continha 4 elementos.

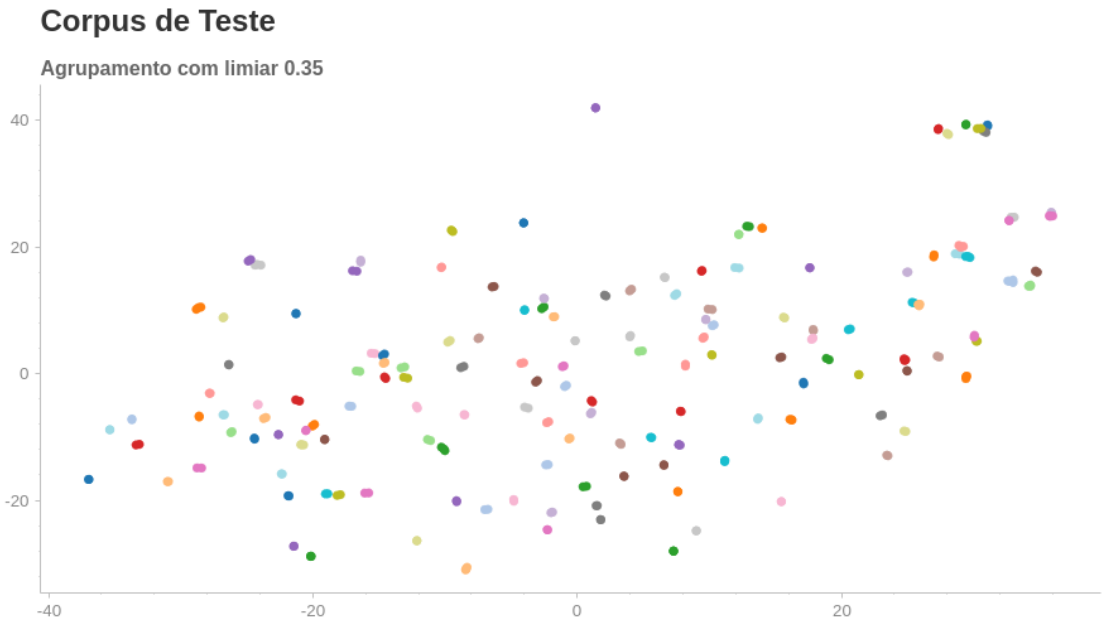

Gráfico 7. Agrupamentos com limiar de 0.35.

Com o limiar de 0.40 foram gerados 167 grupos que abarcaram 363 documentos, com uma cobertura de 32,03% da amostra. O maior grupo continha 4 elementos.

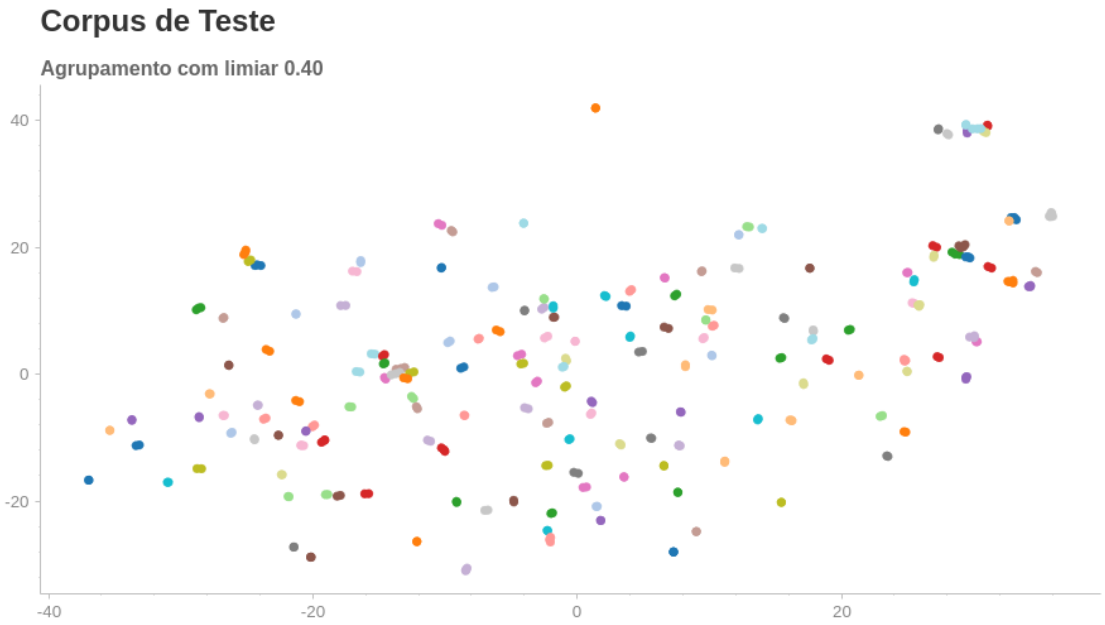

Gráfico 8. Agrupamentos com limiar de 0.40.

Com o limiar de 0.45 foram gerados 182 grupos que abarcaram 404 documentos, com uma cobertura de 35,65% da amostra. O maior grupo continha 6 elementos.

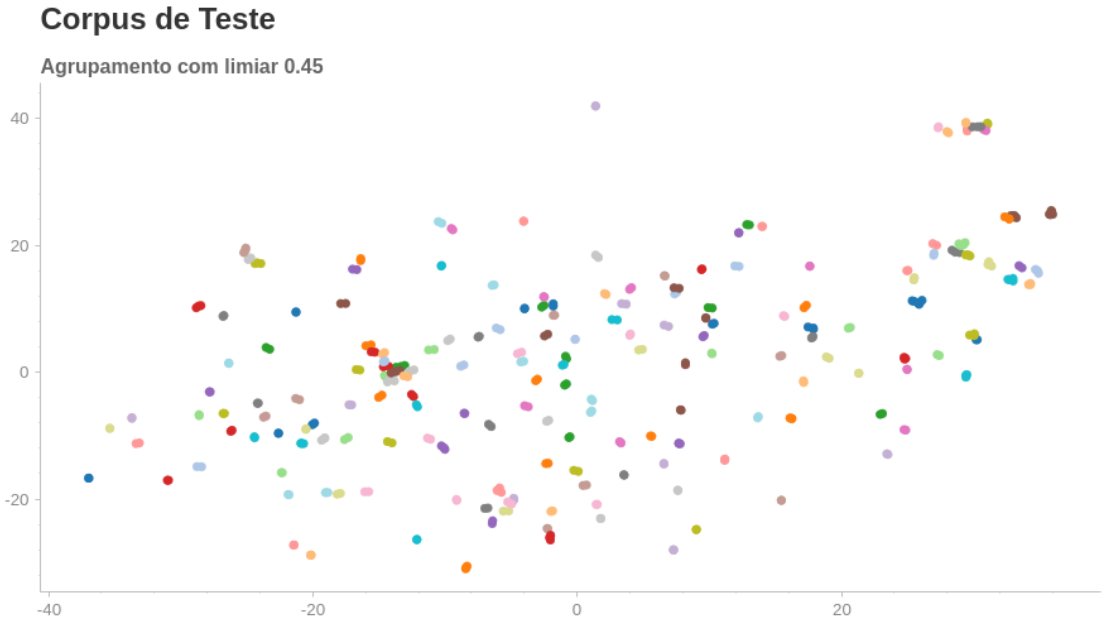

Gráfico 9. Agrupamentos com limiar de 0.45.

Com o limiar de 0.50 foram gerados 194 grupos que abarcaram 451 documentos, com uma cobertura de 27,62% da amostra. O maior grupo continha 7 elementos.

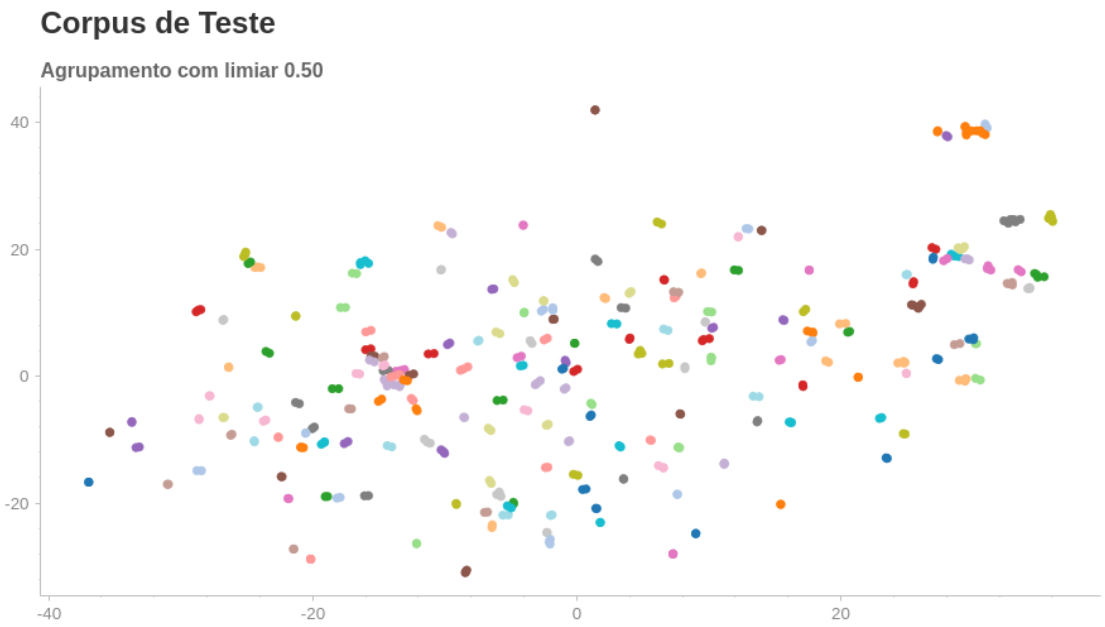

Gráfico 10. Agrupamentos com limiar de 0.50.

O quadro abaixo que sintetiza os resultados obtidos com o algoritmo DBSCAN.

Limiar	Documentos	Cobertura	Grupos	Maior Grupo
0.30	262	23,12%	125	3
0.35	313	27,62%	147	4
0.40	363	32,03%	167	4
0.45	404	35,65%	182	6
0.50	451	39,80%	194	7

Tabela 5. Consolidação dos grupos por limiar.

A primeira observação possível é que, embora o limiar tenha chegado a 0.50, a cobertura dos agrupamentos não chegou a 40% dos documentos integrantes do corpus de teste. Essa observação parece contrastar com a análise de Muniz (2018), que indica um altíssimo percentual de repetitividade nas matérias trazidas ao STJ. O gráfico abaixo ilustra a avaliação.

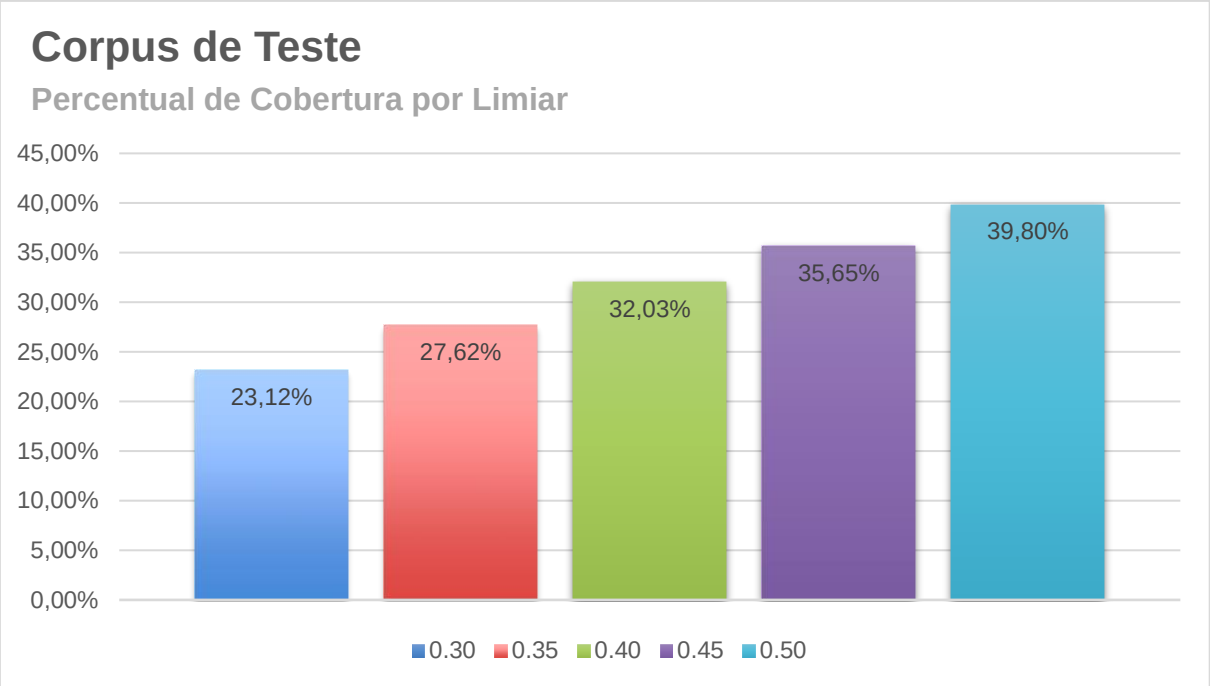

Gráfico 11. Cobertura do corpus de teste por limiar.

O que observamos no gráfico é que, embora o corpus utilizado no treinamento seja volumoso, não era suficientemente abrangente para garantir a aplicação a qualquer controvérsia trazida na peça sob análise.

4.4 – O avaliador humano

O avaliador humano encarregado da validação dos grupos é mestre em direito constitucional com mais 10 anos de experiência em gabinete que opera com direito público.

A escolha do avaliador deu-se por conta de sua larga experiência na área, inclusive em atividades de triagem processual.

4.5 - A avaliação dos grupos

Na avaliação dos grupos, optou-se por analisar amostras de todos os limiares, estabelecendo-se os seguintes critérios para seleção dos grupos:

a) Serão avaliados 5 grupos de cada limiar;
b) Será selecionado um grupo com o menor número de integrantes e um com o maior número aferido dentro do limiar;
c) Os outros três grupos serão selecionados considerando seu tamanho, em ordem decrescente.

Os grupos foram nomeados segundo o limiar aplicado, seguido do identificador numérico atribuído pelo algoritmo. Aplicando as regras de seleção foram avaliados os seguintes grupos:

Grupo	Documentos
30.01	201702917974, 201703162609
30.21	201703363750, 201800397702, 201801070591
30.30	201800022636, 201800413757, 201801127676
30.46	201800106084, 201800173435, 201800199060
30.09	201703257508, 201800428340, 201800889574
35.01	201702917974, 201703162609
35.10	201703257508, 201800428340, 201800889574
35.02	201703129895, 201800655969, 201800940477
35.03	201703154380, 201800167390, 201800478649
35.71	201800183913, 201800388501, 201800389179, 201800563320
40.01	201702917974, 201703162609
40.03	201703129895, 201800655969, 201800940477
40.04	201703154380, 201800167390, 201800478649
40.82	201800183913, 201800388501, 201800389179, 201800563320
40.86	201800192823, 201800443076, 201800838391, 201801183879
45.01	201702917974, 201703162609
45.46	201800034058, 201800177426, 201801082963, 201801184222, 201801213030
45.61	201800061597, 201800102929, 201800106084, 201800173435, 201800199060, 201800281711
45.90	201800183913, 201800388501, 201800389179, 201800563320
45.93	201800192823, 201800443076, 201800838391, 201801183879
50.01	201702917974, 201703162609
50.35	201800000797, 201800010770, 201800049991, 201800282810,

	201800401300, 201800470578, 201800901440
50.49	201800034058, 201800177426, 201801082963, 201801184222, 201801213030, 201801243358
50.66	201800061597, 201800102929, 201800106084, 201800173435, 201800199060, 201800281711
50.82	201800128756, 201800162061, 201800178790, 201800180778, 201800435230, 201800511151, 201800563457

Tabela 6. Grupos submetidos ao avaliador.

A relação dos grupos foi repassada ao avaliador humano que examinou no SIAJ os acórdãos recorridos de cada um dos feitos. Os quesitos avaliados foram os seguintes:

1) Quantos integrantes do grupo tratam da mesma matéria ou de matérias intimamente correlatas?

2) O agrupamento realizado no experimento é suficientemente preciso para possibilitar o julgamento conjunto dos feitos, considerada devolvida toda a matéria tratada nos documentos?

3) Adicionalmente, foi solicitado ao avaliador que descrevesse o grupo com um pequeno número de palavras.

4.5.1 - Resultados da Avaliação – Considerações Gerais:

A transcrição das avaliações resultou no quadro abaixo:

Grupo	**Quesitos**		
	1	**2**	**3**
30.01	2	S	Execução Fiscal - Nulidade da CDA
30.09	3	S	Execução Individual - Ação Coletiva - Interrupção da Prescrição - Súmula 150/STF
30.21	3	S	Empréstimo Compulsório - Energia Elétrica
30.30	3	S	Fiocruz - Reconhecimento Administrativo de Débito - Servidor Público
30.46	3	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
35.01	2	S	Execução Fiscal - Nulidade da CDA
35.02	3	S	CFA - Anuidades - Holdings
35.03	3	S	Empréstimo Consignado - Execução com desconto em folha.
35.10	3	S	Execução Individual - Ação Coletiva - Interrupção da Prescrição - Súmula 150/STF
35.71	4	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
40.01	2	S	Execução Fiscal - Nulidade da CDA
40.03	3	S	CFA - Anuidades - Holdings

40.04	3	S	Empréstimo Consignado - Execução com desconto em folha.
40.82	4	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
40.86	4	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
45.01	2	S	Execução Fiscal - Nulidade da CDA
45.46	5	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
45.61	6	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
45.90	4	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
45.93	4	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
50.01	2	S	Execução Fiscal - Nulidade da CDA
50.35	**0**	**N**	**Descarte - Diversas matérias diferentes**
50.49	6	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
50.66	6	S	Direito Previdenciário - Ecs 20/1998 e 41/2003
50.82	7	S	Direito Previdenciário - Ecs 20/1998 e 41/2003

Tabela 7. Consolidação das avaliações do especialista.

Observa-se ainda que são idênticos os grupos:

a) 35.71 e 40.82;

b) 35.03 e 40.04;

c) 30.09 e 35.10;

d) 30.01, 35.01, 40.01, 45.01 e 50.01;

e) 35.02 e 40.03.

O fato de haver grupos idênticos mesmo com a aplicação de limiares diferentes demonstra a estabilidade do algoritmo DBSCAN.

Excetuado o grupo 50.35, que agrupou documentos que tratam de assuntos tão variados quanto matrícula em instituição de ensino superior e execuções fiscais, para todos os demais agrupamentos a avaliação considerou viável a aplicação de decisões idênticas.

4.5.2 – Avaliações por matéria tratada.

Ao plotar no gráfico abaixo apenas os documentos que compõem os grupos analisados, unificando-os pelo assunto descrito pelo analista, observa-se que o modelo conseguiu reunir os documentos de acordo com a matéria tratada, agrupando em uma mesma região do plano aqueles que tratam de assuntos semelhantes ou correlatos.

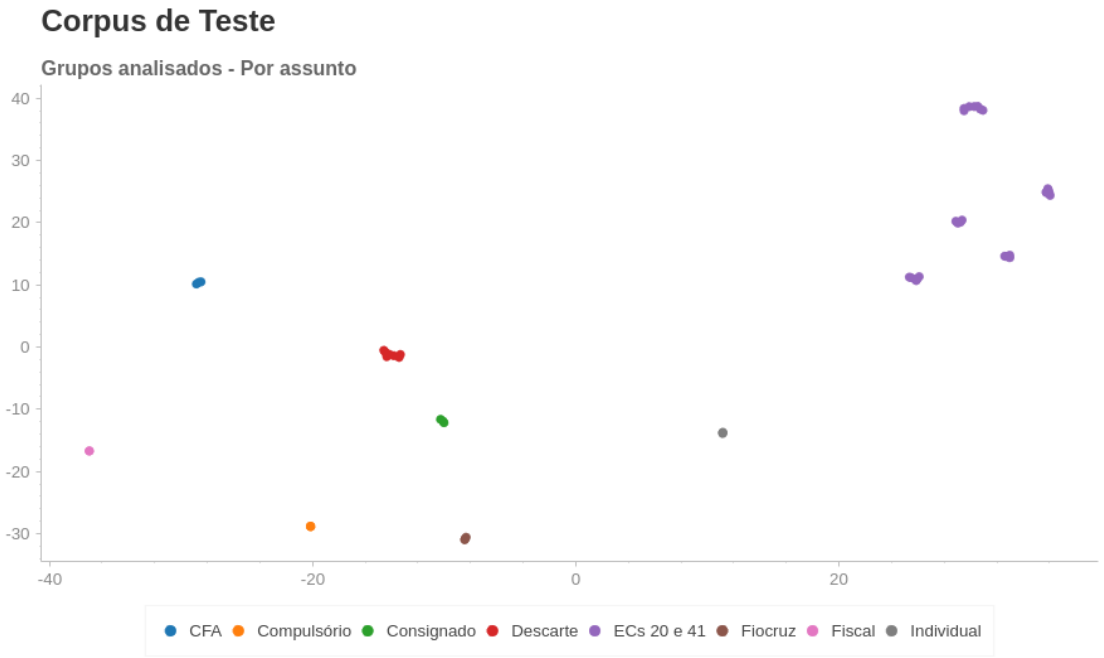

Gráfico 12. Documentos avaliados agrupados por assunto.

Em um segundo passo, solicitou-se ao analista que reavaliasse em conjunto os documentos integrantes dos grupos que foram por ele classificados como “Direito Previdenciário - Ecs 20/1998 e 41/2003” para que verificasse se poderiam ser reunidos em um único grupo, mantidos os quesitos iniciais da avaliação. Com a resposta positiva do analista, pode-se afirmar que, embora o experimento seja efetivo em juntar documentos semanticamente semelhantes, as distâncias calculadas entre os documentos ainda não podem ser consideradas ótimas, demandando uma abordagem mais refinada no treinamento do modelo.

Reforça esse raciocínio a avaliação do analista sobre o grupo 50.35, que trouxe matérias que resultaram em resposta negativa para os dois quesitos iniciais da análise.

Observando os documentos dos grupos avaliados em relação aos demais documentos do corpus de teste, percebe-se que o grupo “Descarte” está compondo uma das áreas mais densamente povoadas do gráfico. Isso indica que o grupo estava em uma área limítrofe entre várias matérias.

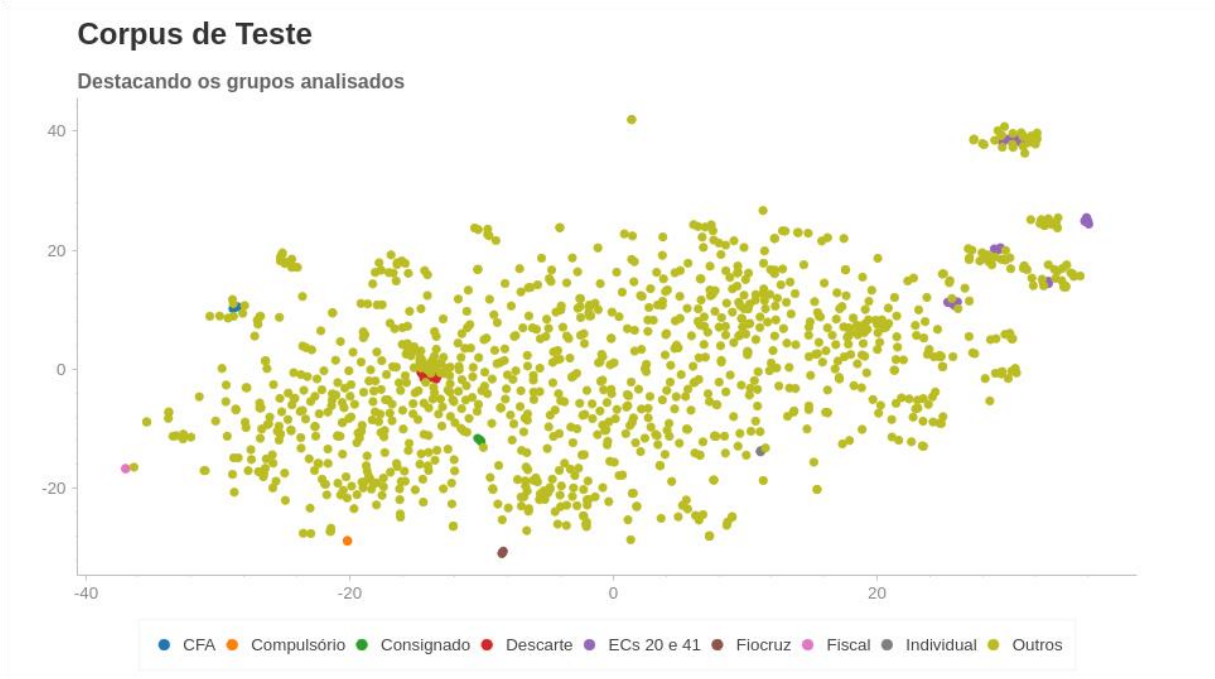

Gráfico 13. Destaque dos grupos analisados, por assunto, dentro do corpus de teste.

Capítulo 5 – Conclusões e Trabalhos Futuros

Este capítulo expõe as conclusões do experimento, limitações e potencialidades da solução bem como possibilidades futuras.

5.1 – Conclusões

O objetivo deste trabalho foi avaliar a utilização do algoritmo *Paragraph Vector* no agrupamento de documentos jurídicos no âmbito do Superior Tribunal de Justiça. A triagem de feitos baseada em seus documentos jurídicos constitui um problema concreto no âmbito da Corte, consumindo tempo e recursos e retardando a prestação jurisdicional.

O objetivo específico de criar um modelo de inteligência artificial capaz de inferir uma representação vetorial para um documento jurídico foi alcançado.

Ao modelo produzido foi submetido um corpus de 1.133 documentos jurídicos diversos dos utilizados no corpus de treinamento, de forma de evitar o *overfitting*. Em um segundo momento, foi reduzida a dimensionalidade dos vetores inferidos pelo modelo e sobre eles foi aplicado um algoritmo de agrupamento baseado em densidade. Uma amostra dos grupos gerados foi avaliada por um especialista humano.

5.2 – Resultados Obtidos

As avaliações constantes do capítulo 4, em especial a efetuada pelo especialista humano, indicam que o modelo foi capaz de lidar bem com a variabilidade dos textos e com os ruídos decorrentes da aplicação do reconhecimento ótico de caracteres. Entretanto, o agrupamento 50.35 indicou algum tipo de desvio que deve ser, em uma eventual repetição do experimento, observada com maior atenção. Essa ressalva, porém, não inviabiliza a utilização imediata da técnica para a execução das atividades de triagem no âmbito do STJ.

O objetivo geral do experimento, consistente na avaliação da aplicação de técnicas de IA no agrupamento de documentos jurídicos no âmbito do STJ, foi integralmente atingido, podendo-se concluir pela aplicabilidade da abordagem na triagem de documentos jurídicos.

O algoritmo *Paragraph Vector* mostrou-se adequado para a solução do problema, lidando bem com a variabilidade vocabular dos documentos integrantes do

corpus de treinamento e identificando adequadamente as relações semânticas entre os termos.

A avaliação demonstra a potencialidade da técnica empregada para a solução de problemas concretos, podendo constituir uma ferramenta de aceleração do trâmite processual no âmbito do STJ. Para fins de registro, o experimento consumiu um tempo total de 10’ para processar 1.133 documentos, um tempo médio < 0.5’’ por documento, algo impossível para um operador humano.

Em todas as atividades foi utilizado um notebook com processador Intel Core I7-7500U de (2.7 a 3.5GHz), equipado com disco SSD SanDisk Plus SDSSDA-480G-G26 e 16GB de memória RAM.

A aplicabilidade da solução transcende o agrupamento de documentos jurídicos semanticamente semelhantes, sendo útil também para a classificação de documentos e indicação de precedentes no âmbito da Corte.

5.3 – Trabalhos Futuros

Em relação a trabalhos futuros, algumas atividades podem ser sugeridas:

a) Extensão do corpus de treinamento, tanto temporalmente, incluindo acórdãos mais antigos, quanto em conteúdo, para que inclua, além dos acórdãos, também os relatórios e votos, que poderão agregar novas relações semânticas ao modelo. É também desejável que sejam incluídos julgados do Supremo Tribunal Federal. Essa inclusão permitiria o mapeamento de relações semânticas específicas do direito constitucional, recurso não compreendido no experimento;
b) Avaliação da aplicação de outros algoritmos de *clustering* que não necessitem da aplicação de redutores de dimensionalidade. Isso reduziria a perda de informação inerente ao PCA.
c) A avaliação do desempenho do modelo quando submetidos a ele documentos menores, tendo em vista que os documentos do corpus de teste, acórdãos proferidos pelo TRF3, são costumeiramente extensos;

d) Uma mudança na abordagem de treinamento do modelo, deixando de operar com documentos inteiros e passando a trabalhar com parágrafos destes documentos. Essa abordagem permitiria, em tese, que o modelo identificasse com maior precisão temas variados dentro de um mesmo documento;
e) Utilização da técnica para testar a classificação de documentos jurídicos.

REFERÊNCIAS

AHIRE, Jayesh B. **Introduction to Word Vectors.** 2018. Disponível em <https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf>. Acesso em 20/11/2018.

CASSIANO, Keila M. **Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade**. Tese (Doutorado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro - Rio de Janeiro - RJ . 2014.

CASSIANO, Keila M., CORDEIRO, Douglas D. **Representação semântica vetorial para análise de similaridade de documentos textuais.** in Anais da VI Escola Regional de Informática de Goiás. Goiânia: SBPC, 2018

CUESTA, Hector. "**Practical Data Analysis**". Birmingham, UK: Packt Publishing, 2013. Print.

DIAS, Marina S. **Regressão Construtiva em Variedades Implícitas.** Tese (Doutorado em Matemática) - Pontifícia Universidade Católica do Rio de Janeiro - Rio de Janeiro - RJ. 2012.

ESTER, M., KRIEGEL, H. P., SANDER, J. & XU, X. "**Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**". in KDD96 Proceedings. 1996

FRUTUOSO, Danielle G. **Recuperação de Informação e Classificação de Entidades Organizacionais em Textos não Estruturados**. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Pernambuco - Recife - PE. 2014.

GLAVAS, F. G., KARAN, M., SNAJDER, J. & BASIC, B. D. "**TakeLab: Systems for Measuring Semantic Text Similarity**". in SEM 2012: The First Joint Conference on Lexical and Computational Semantics, 2012

JORDAN, Michael I. "**Artificial Intelligence - The Revolution Hasn't Happened Yet**." 2018, disponível em <https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7>. Acesso em 16/08/2018.

LAVELLI, A. , SEBASTIANI, F. e ZANOLI, R. "**Distributional Term Representations: an experimental comparison**". in: CIKM 04: "Proceedings of the Thirteenth ACM 130 International Conference on Information and Knowledge Management", 2004.

LE, Quoc V. MIKOLOV, Tomas: "**Distributed Representations of Sentences and Documents**", Mountain View, CA: Google Inc. 2014

MACCORMICK, Neil. **Retórica e o Estado de Direito**. Rio de Janeiro: Elsevier. 2008.

MANNING, C. & SCHÜTZE, H. "**Foundations of Statistical Natural Language Processing**". MIT Press, 2000

MANNING, C. D., RAGHAVAN, P., and SCHUTZE, H. "**Introduction to Information Retrieval**". Cambridge University Press, Online edition. 2009

MIKOLOV, Tomas. SUTSKEVER, Ilya. CHEN, Kai. CORRADO, Greg. DEAN, Jeffrey. "**Distributed Representations of Words and Phrases and their Compositionality**", Mountain View, CA: Google Inc. 2013

MINSKY, M. (editor). "**Semantic Information Processing**". Cambridge: The MIT Press, 1968.

MITCHELL, T. "**Machine learning**". McGraw-Hill, 1997.

MUNIZ, Montgomery. W. **Gerenciamento de Processos Judiciais - Superior Tribunal de Justiça – Corte de "casos fáceis"**, 2018, no prelo

NUNES, Filipe V. "**Verbal Lemmatization and Featurization of Portuguese with Ambiguity Resolution in Context**". Dissertação (Mestrado em Engenharia Informática) - Universidade de Lisboa - Lisboa - PT. 2007.

OLIVEIRA, Ricardo Mariz de. **Incertezas que entravam o desenvolvimento**. In: SANTI, Eurico Marcos Diniz de (Org.). Tributação e Desenvolvimento. São Paulo: Quartier Latin, 2011.

SILVA, T. S. **Reconhecimento de Entidades Nomeadas em Notícias de Governo**. Dissertação (Mestrado em Engenharia de Sistemas e Computação) - Universidade Federal do Rio de Janeiro - Rio de Janeiro. 2012.

SINCLAIR, J. 2005. "**Corpus and Text - Basic Principles in Developing Linguistic Corpora: a Guide to Good Practice**". ed. M. Wynne. Oxford: Oxbow Books: 1-16. Disponível em https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm. Último acesso em 21/08/2018

TAN, P., STEINBACH, M. & KUMAR, V. "**Introduction to Data Mining**". Pearson Education, Inc. 2006.

VIEIRA, Renata & LOPES, Lucelene. **Processamento de Linguagem Natural e o Tratamento Computacional de Linguagens Científicas**. in PERNA et al. Linguagens Especializadas em Corpora - Modos de Dizer e Interfaces de Pesquisa. Porto Alegre: EDIPUCRS, 2010.