# Algoritmo de classificação de textos jurídicos a partir da segmentação semântica

Danilo Panzeri Carlotti
João Eduardo Ferreira

## 1) Introdução

Como projeto de pós-doutorado do primeiro autor, foi utilizado um algoritmo previamente desenvolvido de segmentação lógica dos textos jurídicos a partir de suas frases. Este algoritmo foi inicialmente utilizado para fazer a sumarização dos textos e posteriormente ele foi utilizado para selecionar partes do texto que serviram para a classificação das decisões de segunda instância em um projeto de análise de precedentes com o Tribunal de Justiça de São Paulo em convênio com a Universidade de São Paulo (IME-USP). A literatura sobre aprendizado de máquina e direito se vale da segmentação de textos para realizar a sumarização de textos jurídicos, como abordado no artigo escrito pelos autores e já publicado[1].

## 2) Descrição da essência do algoritmo

Com base em experiências anteriores de processamento de textos jurídicos, os grandes diferenciais da abordagem do algoritmo resumem-se a dois aspectos metodológicos que nortearam e norteiam a construção do algoritmo desde o início do projeto: 1) utilização de classificação multirrótulo (implementada com classificadores binários) com uma abordagem computacional simples, no caso a regressão logística; 2) sumarização semântica de textos de acordo com a segmentação lógica de conteúdos de relevância jurídica: "pedidos", "fatos", "leis" e "decisões".

O principal argumento do aspecto 1 considera a complexidade dos textos jurídicos, que normalmente exigem uma diversidade de argumentações e com combinações múltiplas , em consonância com a complexidade do mundo real. Assim, desde o início do projeto de classificadores, a equipe do IME-USP sustenta e insiste que a família de algoritmos não poderia ignorar essa especificidade do TJ e que, mandatoriamente, teria que assumir a abordagem multirrótulo como um caminho plausível de solução. Essa visão não foi facilmente aceita por parte dos membros do projeto até a execução do experimento em questão solicitado pelo TJ. Depois da apresentação dos resultados do experimento, a abordagem multirrótulo apresentada pelo IME-USP foi assumida como um padrão de implementação para os demais participantes do projeto. Adicionalmente ao aspecto 1, apresentamos, em várias oportunidades do projeto, a existência de uma certa semelhança e equivalência dos algoritmos de classificação no que se refere ao desempenho. Por isso, implementamos o classificador multirrótulo por meio de n classificadores binários com regressão logística, que não demanda uma grande complexidade computacional de processamento.

O principal argumento do aspecto 2 considera a necessidade de sumarização de textos jurídicos, mas sem gerar a perda da generalidade e representação do arcabouço jurídico. Por isso denominamos sumarização semântica com base em segmentos lógicos jurídicos. Mais concretamente, consideramos os segmentos lógicos "pedidos", "fatos", "leis" e "decisões". Insistimos nesse aspecto de segmentos lógicos, pois embora a sumarização de textos seja um importante campo de pesquisa em processamento de linguagem natural, ela não pode ser diretamente aplicada em documentos jurídicos. Existem diferenças importantes entre sumarização de textos gerais e jurídicos. No caso de textos jurídicos, a linguagem é técnica e com um vocabulário específico e os documentos tendem a ter uma estrutura comum. . Existem restrições lógicas inerentes a uma decisão e informações (como precedentes legais) importantes em si mesmas e que podem

[1] https://reedrevista.org/reed/article/view/600/418

ser extraídas. Portanto, é imperativo que os métodos para sumarizar os documentos legais abordem essas particularidades de segmentação lógica. As palavras, frases e combinação de frases para cada segmento lógico ("pedido", "fato", "lei", "decisão") foram colecionadas tendo como referência centenas de milhares de textos jurídicos. A partir dessa coleção, os padrões dos segmentos são utilizados para extrair significados jurídicos dos acórdãos do experimento.

Além das apresentações realizadas durante as reuniões do projeto e do código computacional disponibilizado, detalhes mais acadêmicos que sustentam os dois argumentos podem ser encontrados no artigo citado.

## 3) Resultados do Experimento do IME-USP

Como primeiro resultado de nossa abordagem, podemos destacar a redução dos acórdãos para 20% do tamanho dos textos originais, sem a perda da representatividade da estruturação jurídica. Cabe observar aqui que existem várias formas de resumir texto, tais como frequências de palavras, por tipo de arquivo e por tipo de tema, mas estudos e experimentos mostram que poderá haver perda de generalidade do significado jurídico. Nossa abordagem para sumarização, além de excluir ruídos e trechos desnecessários, gerando uma redução do tamanho do texto, **preserva a essência jurídica dos textos dos acórdãos** e evita os conhecidos fenômenos de "oversampling" muito comuns em sumarizações específicas, parciais e tendenciosas.

Como segundo resultado, temos que em média o algoritmo teve um desempenho de acurácia superior a 95% em quase todos os cenários aos quais ele foi exposto.