DOI: 10.5748/18CONTECSI/REX/DSC/6848

INFORMATION RETRIEVAL IN LEGAL DOCUMENTS ON HEALTH DEMAND

Diógenes Carlos Albuquerque Araújo - UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - Orcid: https://orcid.org/0000-0002-9859-9255

João Pedro Da Silva Lima - UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - Orcid: https://orcid.org/0000-0002-5706-6065

José Alfredo Ferreira Costa - UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE - Orcid: https://orcid.org/0000-0002-1290-6454

The Brazilian judicial system is currently one of the largest in the world with more than 77 million cases awaiting decision. One of the most slow steps is the identification of relevant information within the initial petition, with the aid of intelligence techniques, it would be possible to get around this problem. This article aims to build models capable of classifying health demands in the judiciary between requests for ICU, surgery, examination and drug supply. In addition to identifying information related to demand, such as diseases of the author and identification of the drug or surgery requested. For this, information retrieval models based on dictionaries, rules or hybrids were used. Within this article, the models obtained are evaluated on the recall rate. Among the models, the best result had a 90.22% recall rate, while the worst had a 37.03% recall rate.

Keywords: Information retrieval, Unstructured data, Legal documents, Health demands, Regular expression

RECUPERAÇÃO DE INFORMAÇÕES EM DOCUMENTOS JURÍDICOS COM DEMANDA DA SAÚDE

O sistema judiciário brasileiro é atualmente um dos maiores do mundo com mais de 77 milhões de processos aguardando decisão. Uma das etapas mais demorada é a identificação de informações relevantes dentro da petição inicial, com auxílio de técnicas de inteligência artificial seria possível contornar esse problema. Este artigo tem como objetivo construir modelos capazes de classificar demandas da saúde no judiciário entre pedidos de UTI, realização de cirurgia, realização de exame e fornecimento de medicamentos. Além de identificar informações relacionadas à demanda, como doenças da parte autora e identificação do medicamento ou cirurgia pedido. Para isso foram utilizados modelos de recuperação de informação baseado em dicionários, regras ou híbridos. Dentro desse artigo os modelos serão avaliados perante a sua taxa de recuperação (Recall). Dentre os modelos o melhor resultado teve 90,22% de taxa de recuperação, enquanto o pior teve uma taxa de recuperação de 37,03%.

Palavras-chave: Recuperação de informação, Dados não estruturados, Documentos jurídicos, Demandas da saúde, Expressão regular

18th International Conference on Information Systems and Technology Management
ISSN 2448-1041
Universidade de São Paulo/SP, 13 a 15 de Outubro de 2021 – Virtual
MODALIDADE RESUMO EXPANDIDO – GRADUAÇÃO E OUTROS NÍVEIS

RECUPERAÇÃO DE INFORMAÇÕES EM DOCUMENTOS JURÍDICOS COM DEMANDA DA SAÚDE

INFORMATION RETRIEVAL IN LEGAL DOCUMENTS ON HEALTH DEMAND

Diógenes Carlos Araújo
0000-0002-9859-9255
077.706.164-33
Universidade Federal do Rio Grande do Norte, Natal/RN
diogenes.carlos@hotmail.com

João Pedro Lima
0000-0002-5706-6065
701.335.194-65
Universidade Federal do Rio Grande do Norte, Natal/RN
joaopedrodasilvalima@gmail.com

José Alfredo Ferreira
0000-0002-1290-6454
538.201.264-49
Universidade Federal do Rio Grande do Norte, Natal/RN
jafcosta@gmaill.com

RESUMO: O sistema judiciário brasileiro é atualmente um dos maiores do mundo com mais de 77 milhões de processos aguardando decisão. Uma das etapas mais demorada é a identificação de informações relevantes dentro da petição inicial, com auxílio de técnicas de inteligência artificial seria possível contornar esse problema. Este artigo tem como objetivo construir modelos capazes de classificar demandas da saúde no judiciário entre pedidos de UTI, realização de cirurgia, realização de exame e fornecimento de medicamentos. Além de identificar informações relacionadas à demanda, como doenças da parte autora e identificação do medicamento ou cirurgia pedido. Para isso foram utilizados modelos de recuperação de informação baseado em dicionários, regras ou híbridos. Dentro desse artigo os modelos serão avaliados perante a sua taxa de recuperação (*Recall*). Dentre os modelos o melhor resultado teve 90,22% de taxa de recuperação, enquanto o pior teve uma taxa de recuperação de 37,03%.

ABSTRACT: The Brazilian judicial system is currently one of the largest in the world with more than 77 million cases awaiting decision. One of the most slow steps is the identification of relevant information within the initial petition, with the aid of intelligence techniques, it would be possible to get around this problem. This article aims to build models capable of classifying health demands in the judiciary between requests for ICU, surgery, examination and drug supply. In addition to identifying information related to demand, such as diseases of the author and identification of the drug or surgery requested. For this, information retrieval models based on dictionaries, rules or hybrids were used. Within this article, the models obtained are evaluated on the recall rate. Among the models, the best result had a 90.22% recall rate, while the worst had a 37.03% recall rate.

PALAVRAS-CHAVE: Recuperação de informação. Dados não estruturados. Documentos jurídicos. Demandas da saúde. Expressão regular.

KEYWORD: Information retrieval. Unstructured data. Legal documents. Health demands. Regular expression.

1 INTRODUÇÃO

O mundo está enfrentando uma transformação digital em passo avançado. Desde a revolução dos computadores, a qual possibilitou um maior controle de informação, o tamanho e a complexidade da base de dados está sempre aumentando. A digitalização tem transformado economias e vidas. O avanço na computação proporcionou um barateamento nos custos de armazenamento de informações, possibilitando um grande volume de dados sobre os mais diversos negócios ou aplicações. Como nos sistemas judiciários. Alguns avanços, como aprendizado profundo proporcionaram às máquinas um avanço e acurácia nunca antes vista. No entanto tais modelos possuem melhor desempenho com dados estruturados, o que nem sempre é o caso.

Como dito por Bhatt [2], a inovação precisa ter valor público e ser moldada para trazer todos para a era digital. O sistema judiciário brasileiro é grande e complexo, com quase 80 milhões de casos não concluídos [4]. Devido à forte política de digitalização judicial, a maioria desses processos é processada integralmente de forma digital, por exemplo, na plataforma PJe, que é utilizada por diversos Tribunais brasileiros. A aplicação de técnicas de aprendizado de máquina ao sistema judiciário é altamente desejável, devido ao tamanho e à complexidade de textos jurídicos não estruturados e não rotulados. A ideia geral é automatizar algumas etapas do processo, como tarefas repetitivas, visando a eficácia judicial e celeridade. O Brasil já conta com diversos projetos voltados para o uso de Inteligência Artificial (IA) na justiça, com aplicações com foco em automatização e melhor recuperação de informações. Projetos como a plataforma Sinapses podem levar a modelos de inteligência artificial robustos e fáceis de usar [3]. Sinapses foram inicialmente desenvolvidos pelo TJRO em colaboração com o CNJ com os objetivos de criar um ambiente colaborativo, que pode condensar e compartilhar modelos de aprendizado de máquina desenvolvidos em vários tribunais brasileiros.

Uma das etapas mais repetitivas para os servidores do tribunal seria a extração de informações importantes através da leitura das petições iniciais, sendo assim esse trabalho tem como objetivo extrair informações da petição inicial de causas relacionadas à saúde, para isso foi utilizado técnicas de busca por palavras chaves. O algoritmo busca extrair a doença, realizar uma nova classificação de assunto entre 4 áreas: fornecimento

de medicamentos, internação em UTI, realização de exame, realização de cirurgia. Dependendo do assunto, o algoritmo pode buscar pelo medicamento ou cirurgia. A seguir será descrito a metodologia utilizada, então desenvolvimento, após os resultados obtidos e por último as considerações finais.

2 METODOLOGIA

O projeto será dividido em três etapas, o primeiro é o estudo das tecnologias mais utilizadas na área de recuperação de informação, sendo essa etapa necessária para a seleção da técnica utilizada. Nessa etapa foi possível observar diversos estudos relacionados à obtenção de dados, específicos a área médica [5][6][7][9], como também casos mais genéricos [1][8]. Analisando os artigos foi observado a existência de quatro abordagens clássicas para extração de informação, são elas: baseada em dicionário, a partir de um lista de palavras pré determinada realizar a busca, baseada em regra, observando uma estrutura textual ou lógica do texto para encontrar a informação, baseada em máquina de aprendizado, treinar um modelo capaz de detectar a informação, e baseada em modelos híbridos, que combinam as técnicas previamente citadas.

A segunda etapa do projeto é o desenvolvimento, nela será definido a técnica utilizada e criado códigos necessários para o seu funcionamento, nessa etapa também será realizado a extração dos documentos e todos os pré-processamentos necessários. A terceira e última etapa será a avaliação do modelo, perante a sua taxa de recuperação (*Recall*).

3 DESENVOLVIMENTO

O primeiro passo foi a definição do modelo utilizado para a extração de cada informação, a seleção desses modelos priorizou o baixo custo computacional. A nova definição de temas dos processos foi realizada através de um modelo baseado em um dicionário com palavras relacionadas ao tema como: uti, uci, cirurgia, procedimento cirúrgico, remédio, medicamento, medicação, exame entre outros. Para encontrar a doença foi utilizado um modelo híbrido entre o baseado em regras e dicionário. O modelo de regra busca pela palavra “cid” seguida por uma estrutura de uma palavra e dois ou três números, sendo essa estrutura padrão das doenças na CID 10, que seria a Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde. Já o dicionário foi construído exclusivamente para esse problema visando detectar doenças em casos que a estrutura da cid não é encontrada, portanto engloba palavras relacionadas a doenças

mais comuns como: diabetes, hipertensão, câncer, insuficiência renal, trombose venosa e trombofilia.

Para a detecção de medicamento foi utilizado exclusivamente o modelo baseado em dicionário, sendo usado como dicionário a lista de medicamentos da Anvisa, usando tanto o nome do fármaco como seu princípio ativo, já para a detecção da cirurgia foi utilizado um modelo baseado em regras, a regra utilizada foi a detecção do procedimento cirúrgico dentro de um formulário anexado junto a petição inicial.

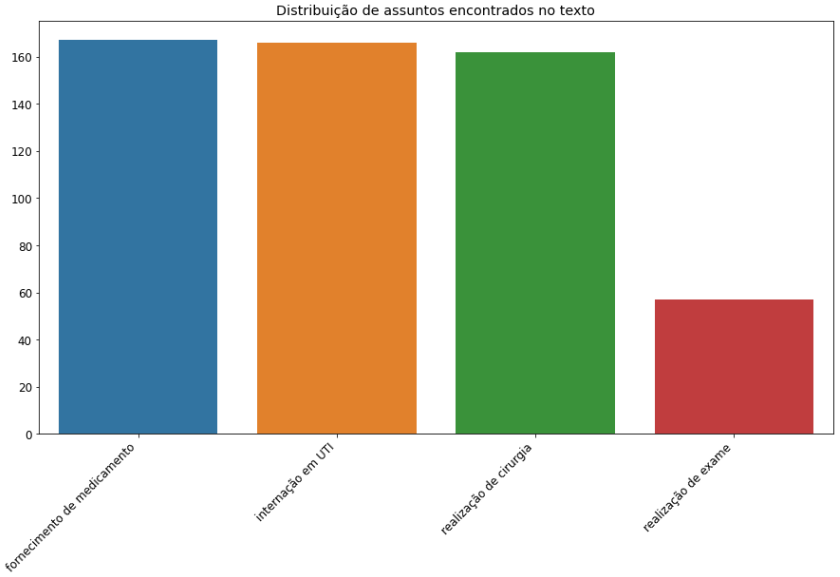
Sendo assim a próxima etapa é a de extração dos documentos, para isso foi utilizado um *web crawler* dentro do sistema de homologação do Tribunal de Justiça do Rio Grande do Norte, toda essa etapa foi realizada com o acompanhamento de servidores do tribunal. Após a extração dos documentos em pdf foi utilizado um sistema para transformá-los em texto, tal sistema desenvolvido pelos próprios autores utilizando python.

4 RESULTADOS OBTIDOS

Com a extração foi obtido 348 documentos todos são relacionados a judicialização da saúde, portanto todos devem ter a citação de pelo menos uma doença, usando o algoritmo de recuperação de informação foi possível detectar a doença em pelo menos 285 documentos, sendo assim deve uma taxa de recuperação de 81,89% uma taxa bastante significativa levando em conta que nem todas as petições possuem a informação da cid e trabalhamos com dicionário pequeno.

A avaliação do modelo de classificação de assuntos também vai ser em cima do total de arquivos, o modelo foi capaz de identificar o assunto em 314 documentos, dando uma taxa de recuperação de 90,22% um valor dentro do esperado, já que os quatro assuntos cobrem quase todas as demandas de saúde, sobrando apenas alguns casos especiais. Para avaliar os demais modelos será utilizado a quantidade de documentos com o tema referente a aquela informação, uma visão geral da quantidade de documento por assunto pode ser vista na figura 1, lembrando que um mesmo documento pode ter mais de um assunto.

Figura 1 - Distribuição de documentos por assunto



Fonte: Autor

O total de documentos sobre fornecimento de medicamento é 167 tendo o modelo detectado o medicamento em apenas 68 casos, tendo assim uma taxa de recuperação de 40,71% indicando que muitos dos pedidos de fornecimento de medicamento não são de medicamentos da tabela de referência da Anvisa. Já a quantidade de pedidos para a realização de cirurgia é 162, sendo o modelo capaz de detectar em 60 petições dando uma taxa de conversão de 37,03%, demonstrando que um modelo simples baseado em regras sozinho não tem grande eficiência para a detecção de cirurgia.

Com isso podemos obter a tabela 1 comparando a taxa de recuperação para cada modelo e com suas lógicas de implementação.

Tabela 1 - Taxa de recuperação por modelo

Uso do Modelo	Taxa de Recuperação	Tipo de Modelo
Identificação da Doença	81,89%	Híbrido (Regras e Dicionário)
Identificação do Assunto	90,22%	Dicionário
Identificação do Medicamento	40,71%	Dicionário
Identificação da Cirurgia	37,03%	Regra

Fonte: Autor

5 CONSIDERAÇÕES FINAIS

A partir dos resultados podemos concluir que os modelos propostos para a identificação da doença e do assunto tiveram resultados dentro das expectativas, resolvendo assim o problema inicialmente proposto, com algumas possíveis melhorias na quantidade de

palavras do modelo de identificação de doenças. No entanto o modelo de identificação de medicamento e cirurgia apresentaram resultado inferior ao esperado necessitando uma mudança na abordagem utilizada, uma possível modificação seria adicionar um modelo de aprendizado de máquina a esses identificadores, tornando-os assim em modelos híbridos.

REFERÊNCIAS

1- Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. Vol. 463. New York: ACM press, 1999.
2 - Bhatt, G.: The Haves and Have-nots (2021)
3 - CNJ - National Council of Justice: SINAPSES (2019)
4 - CNJ - National Council of Justice: Justiça em Números: ano-base 2019 (2020)
5 - Hersh, William, Hersh, and Weston. Information retrieval: A biomedical and health perspective. Springer, 2020.
6 - Gurulingappa, H.; Rajput, A.; Roberts, A.; Fluck, J.; Hofmann-Apitius, M.; Toldo, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J. Biomed. Inform. 2012, 45, 885–892.
7 - Mulligen, E.; Fourrier-Reglat, A.; Gurwitz, D.; Molokhia, M.; Nieto, A.; Trifiro, G.; Kors, J.A.; Furlong, L.I. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. J. Biomed. Inform. 2012, 45, 879–884.
8 - Ferrucci, D.; Lally, A. UIMA: An architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. 2004, 10, 327–348.
9 - Liu, Shengyu, et al. "Drug name recognition: approaches and resources." Information 6.4 (2015): 790-810.