PROJETO DE PESQUISA

Classificação de documentos de patentes por assunto e categorias específicas com uso de processamento de linguagem natural

Aluno: Henrique Cursino Vieira

Orientador: Nikolai Kolev

1. Justificativa e importância (Justificativas e importância do projeto de pesquisa em termos de relevância para a área)

No desenvolvimento geral de produtos, a pesquisa por documentos de patentes visa a garantia de não infringimento de propriedades intelectuais que ainda não estão em domínio público (BREITZMAN E MOGEE, 2002). Em uma pesquisa de documentos de patentes, podemos também buscar em entender em qual ponto esta tecnologia está. Mas como realizar esta tarefa havendo algumas centenas de documentos de patentes sobre um assunto específico? Hoje já há portais web que oferecem ferramentas das quais algumas auxiliam ao pesquisador a reduzir essa pesquisa (ABBAS, ZHANG, KHAN, 2014), mas eles classificam os documentos em uma relevância geral. Esse resultado somente demonstra que dentro daquela amostra de documentos uma visão macro sobre o assunto, muitas vezes o pesquisador está em busca de um subassunto, como quais mercados essa tecnologia está presente, quais os processos de produção desta tecnologia ou qual a formulação desse composto. Este projeto busca atender essa necessidade, classificando os documentos de patentes em categorias e atribuindo um valor de relevância para esta categoria, permitindo ao pesquisador buscar as patentes que o interessa e em um volume menor (SHAHID et al., 2019).

2. Objetivos (Objetivos gerais e específicos do Projeto de Pesquisa)

Objetivos gerais: Este projeto se propõe a classificar documentos de patentes por assunto específico e categorias de interesse do pesquisador, reduzindo o escopo de documentos de patentes a serem estudados à somente os mais relevantes para o que se procura.

Objetivos específicos: A classificação de documentos de patentes envolverá o uso técnicas de processamento de linguagem natural para o tratamento e preparação dos dados que serão usados no modelo de classificação por relevância que será desenvolvido. Este modelo usará inicialmente a medida estatística TF-IDF (term frequency - inverse document frequency) e avaliaremos outras medidas. Haverá a necessidade de criação de dicionários que auxiliem na classificação dos documentos de patentes.

3. Metodologia (Metodologia a ser utilizada no Projeto de Pesquisa)

Para a realização desse projeto, faremos um levantamento bibliográfico sobre as principais formas de classificação de documentos por relevância. Iremos usar a base dados Google Patents para obter os documentos de patentes sobre o assunto “telefonia móvel”. Através de técnicas de processamento de linguagem natural, iremos processar e normalizar as palavras, contar a frequência de palavras, isto será usado para a construção de um dicionário inicial e para calcular o TF-IDF do modelo de relevância, classificar os documentos de patentes nas categorias a partir do dicionário (SHAHID et al., 2019). O resultado será apresentado inicialmente como uma planilha.

4. Cronograma (Relação itemizada das atividades previstas, em ordem sequencial e temporal, de acordo com os objetivos traçados no projeto e dentro do período de um ano)

5. Levantamento bibliográfico e obtenção dos documentos de patentes - 1 mês;

6. Construção do dicionário - 1 mês;

7. Tratamento da base de dados - 1 mês;

8. Desenvolvimento do modelo de relevância e avaliação de medidas - 3 meses;

9. Aplicação e ajustes a classificação dos documentos de patentes - 3 meses;
10. Discussão dos resultados - 2 meses;

5. Resultados e Impactos Esperados (Relação dos resultados ou produtos que se espera obter após o término da pesquisa)

Obter um modelo de relevância que seja rápido e confiável; Automatizar a classificação de documentos de patentes em categorias de interesse ao pesquisador; Desenvolvimento de uma aplicação que apresente os resultados visualmente;

6. Referências Bibliográficas (Relação itemizada das referências que subsidiam a proposta de pesquisa em ordem alfabética, com no máximo 10 referências)

- ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. World Patent Information, v. 37, p. 3-13, 2014.
- BREITZMAN, A. F.; MOGEE, M. E. The many applications of patent analysis. Journal of information science, v. 28, n. 3, p. 187-205, 2002.
- SHAHID, M. et al. Automatic Patents Classification Using Supervised Machine Learning. In: International Conference on Soft Computing and Data Mining. Springer, Cham, 2020. p. 297-307.