Materiais e métodos

Revisão sistemática

O estudo de Revisão Sistemática da Literatura seguiu as recomendações Preferred Reporting Items for Systematic Reviews and Meta-Analisys – PRISMA (REFERENCIA).

Foram buscados os termos: “patent mining”, “patent”, “random forest”, “machine learning” - nas seguintes bases de dados: Periodicos CAPES, Microsoft research, Semantic Scholar e Google Scholar. O intervalo de publicação dos artigos selecionados estão entre 2012 a 2020 e restrito a somente artigos escritos em inglês.

1 – Descrição do objeto de estudo

Foi realizado a extração de dados de documentos de patentes no site Free Patents Online - FPO (<https://www.freepatentsonline.com/>). Este site contem os dados dos documentos de patentes de forma pública.

2 – Delineamento da pequisa

Foi buscado o termo “agronomy” e filtrado para somente documentos de patentes registrados nos Estados Unidos. Foi totalizado 12906 patentes, dos quais selecionamos uma amostragem das 200 primeiras patentes. Construímos uma aplicação de webscraping na linguagem Python para realizar a extração dos dados de documentos de patentes. Os dados extraídos foram armazenados em um banco de dados.

3 – Procedimentos específicos

Extração das dados

A aplicação de webscraping dos dados de documentos de patentes foi escrita na linguagem de programação Python, com uso das bibliotecas requests e BeautifulSoup. Essa aplicação é modular o suficiente para que seja definido quantos documentos de patentes terão suas informações extraídas, como também quais informações serão extraídas. Os dados são organizados na forma de tabela e armazenado em um pequeno banco de dados feito em SQLite.

FIGURA DE FLUXO E EXPLICACAO DO WEBSCRAPING E DO PROCESSAMENTO APLICADO

Construção do dicionário

A construção do dicionário que será utilizado no projeto é composto pelas seguintes etapas, geração de um corpora de documentos de patentes, pre processamento do corpora, obtenção da matriz de documento-termo (Document-Term Matrix – DTM) e aplicação do modelo Latent Dirichlet allocation (LDA). A partir dos tópicos apresentados pelo resultado do LDA, são adicionados ao tópicos, palavras relacionadas, tais como sinônimos, hiperônimos e hipônimos através do banco de dados wordnet.

FIGURA DO FLUXOGRAMA DE CONSTRUCAO DO DICIONARIO

4 – Validação do dicionário

A avaliação do dicionário obtido, consiste em observar se o valor k utilizado para geração de tópicos conseguiu separar adequadamente os assuntos contidos no corpora.

5 – Classificação a partir do dicionario

Esta etapa consiste em utilizar o dicionario para classificar os documentos de patentes a partir da iteração com cada termo do dicionario para o conjunto de termos de cada documento, classificando para cada tópico. Esta tarefa é demorada e o tempo necessário aumenta exponencialmente conforme aumenta o tamanho do corpora utilizado. A base de dados gerada será usada para ensinar ao modelo como classificar novos documentos de patentes.

6 – Modelagem

Utilizaremos três dos modelos mais citados na classificação de texto, o RandomForest, Naive Bayes e SVM para avaliar qual se adéqua melhor a essa classificação. Usaremos as técnicas de pré processamento para garantir que o mesmo dado será testado igualmente para cada modelo e escolheremos o modelo de melhor acurácia como modelo final.