Extração de dados

Extraímos uma amostra no total de 200 documentos de patentes através do uso da técnica de webscraping. Destes documentos, os dados de Titulo e Resumo foram pré processados, removendo as quebras de linhas, espaços no inicio e fim da frase, uso de somente um espaço como separador e transformação em minúsculo. Estes dados foram concatenados e usados para a montagem do corpora de documentos de patentes, que poderá ser utilizado para outros projetos.

IMAGEM DA NUVEM DE PALAVRAS

Construção nuvem de palavras dos termos mais representativos para este corpora.

Construção do dicionário

A construção do dicionário engloba o levantamento de tópicos, expansão dos termos, avaliação dos tópicos e expansão dos termos.

Levantamento de tópicos

O corpora de documentos de patentes foi carregado, removido as stopwords, removido também caracteres numéricos e especiais, depois separados em palavras (tokens) e desflexionados para a sua raiz (lemmas). Resultando em 200 conjuntos de palavras normalizadas representando cada documento de patente e que esta pronto para ser utilizado em modelos de Processamento de Linguagem Natural e em modelos de Aprendizado de Maquina.

Aplicamos o modelo LDA, com os seguintes parâmetros - random\_state igual a 100, update\_every igual a 1, chuncksize igual a 100, passes igual a 10 e alpha automático. Para definir a quantidade de tópicos k, usamos um laço de 40 interações e anotamos o valor da métrica Coherence.

FIGURA DO GRAFICO DE Coherence SCORE

A distribuição dos valores de Coherence ao longo da variação do parâmetro k, permite que observemos qual a quantidade de tópicos devemos ter.

O gráfico aponta que um k igual a 20 resulta no segundo mais alto valor de Coherence. Utilizaremos este valor para k, pois é um número menor de conjuntos de palavras que precisará serem interpretadas para se definir qual o título do tópico se referem.

Validação dos tópicos

Examinamos o tópicos obtidos através da ferramenta pyLDAvis. Os termos que compõe os tópicos gerados representam bem o corpora usado. Temos pouca sobreposição, com exceção do tópico 18, e os termos de cada tópico possuem uma alta relevância com o tema agronomia.

FIGURA DO GRAFICO pyLDAvis

O gráfico de bolhas, cada bolha representa um tópico, o tamanho da bolha representa a prevalência do tópico e a sobreposição de bolhas aponta a similaridade entre os tópicos. O gráfico da direita, as barras representam a relevância do termo para o tópico observado.

Expansão do dicionario

Antes de expandir o dicionário, realizamos a remoção dois tópicos que estavam muito similares. Os tópicos geraram no total de 98 palavras únicas que foram submetidas ao wordnet e adicionado os sinônimos, hiperônimos e hipônimos destes termos, totalizando 433 termos que representam cada tópico. A estrutura do dicionário criado é composta por três colunas, a primeira é o tópico, a segunda são os termos que estão atrelada ao tópico e a terceira coluna são as palavras derivadas dos termos.

Obtivemos no final um dicionário com com 954 linhas e três colunas, que foi utilizado para fazer uma classificação inicial dos documentos de patentes.

Modelo

Para a construção do modelo, realizamos um pre processamento convertendo o conteúdo do documento de patente em uma matriz documento-termo, esta matriz tem a estrutura da seguinte forma:

colunas: palavras de relevância

linhas: documentos

valores: correspondem ao valor de TF-IDF obtido, quando a palavra não consta na entrada, o valor será igual a zero.

O modelo testado foi o Random Forest, com critério de separação Gini, obtendo um score igual a 0,56.