# Extração de dados

Extraímos uma amostra no total de 904 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 do texto em minúsculo. Estes dados foram concatenados e usados para a montagem do corpora de documentos de patentes, que poderá ser utilizado para outros projetos.

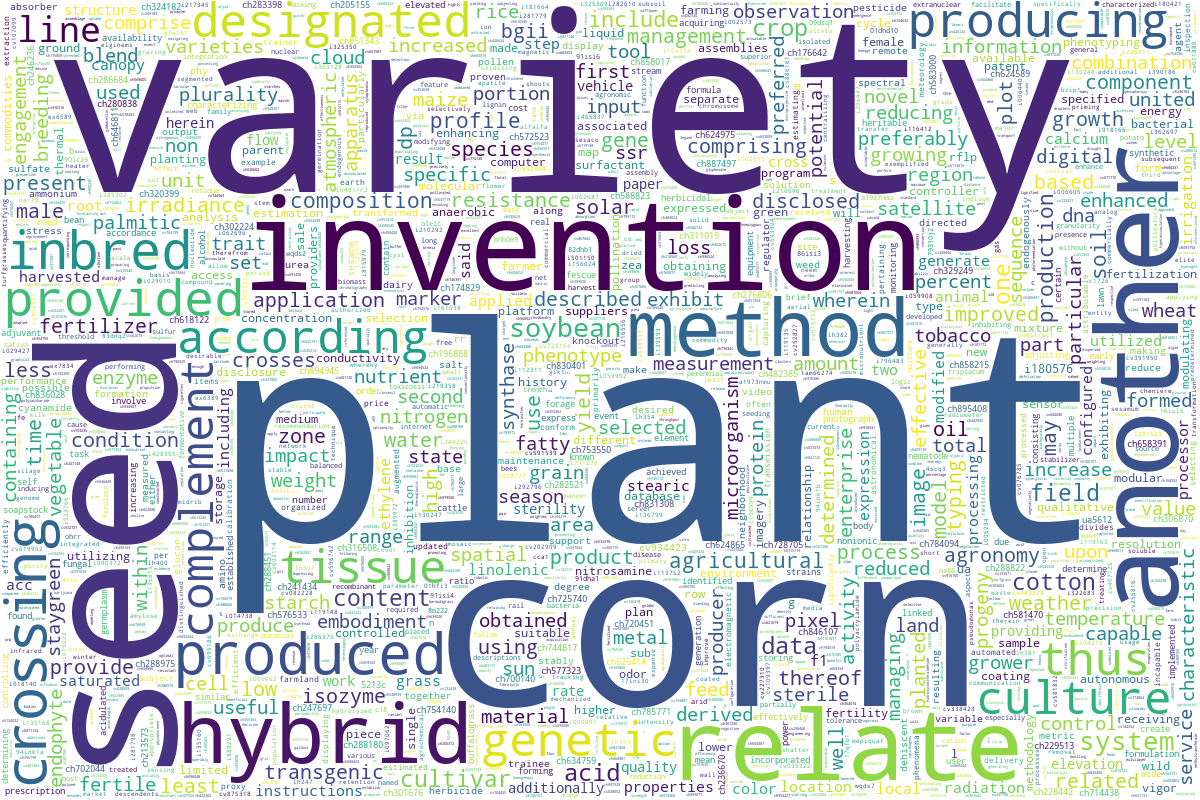


Fig. Construção nuvem de palavras dos termos mais representativos para este corpora.

# Construção do dicionário

A construção do dicionário engloba o levantamento de tópicos, validação dos tópicos, expansão do dicionário.

## Levantamento de tópicos

Foi utilizado o corpora de documentos de patentes feito no passo anterior, onde foi removido as stopwords (palavras que não possuem importância a frase, por exemplo em Inglês: The, from, a, an, with, etc.), foram removidos também caracteres numéricos e especiais. O conteúdo de cada corpus foram separados em uma lista palavras, este processo denomina-se como geração de tokens e cada token foi desflexionado para a sua palavra raiz (lemmas). Obtivemos 904 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.

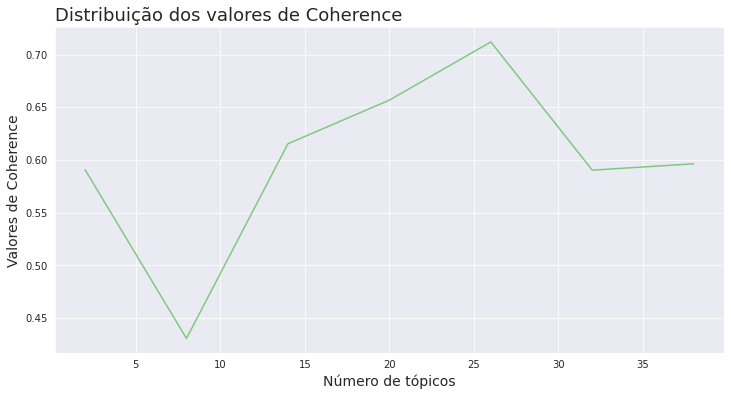


Fig. A distribuição dos valores de Coherence ao longo da variação do parâmetro k, permite que observemos qual a quantidade de tópicos mais relevantes a se utilizar.

O gráfico aponta que um k igual a 25 resulta no mais alto valor de Coherence. Utilizaremos este valor para k para se definir os títulos de tópico.

## 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.

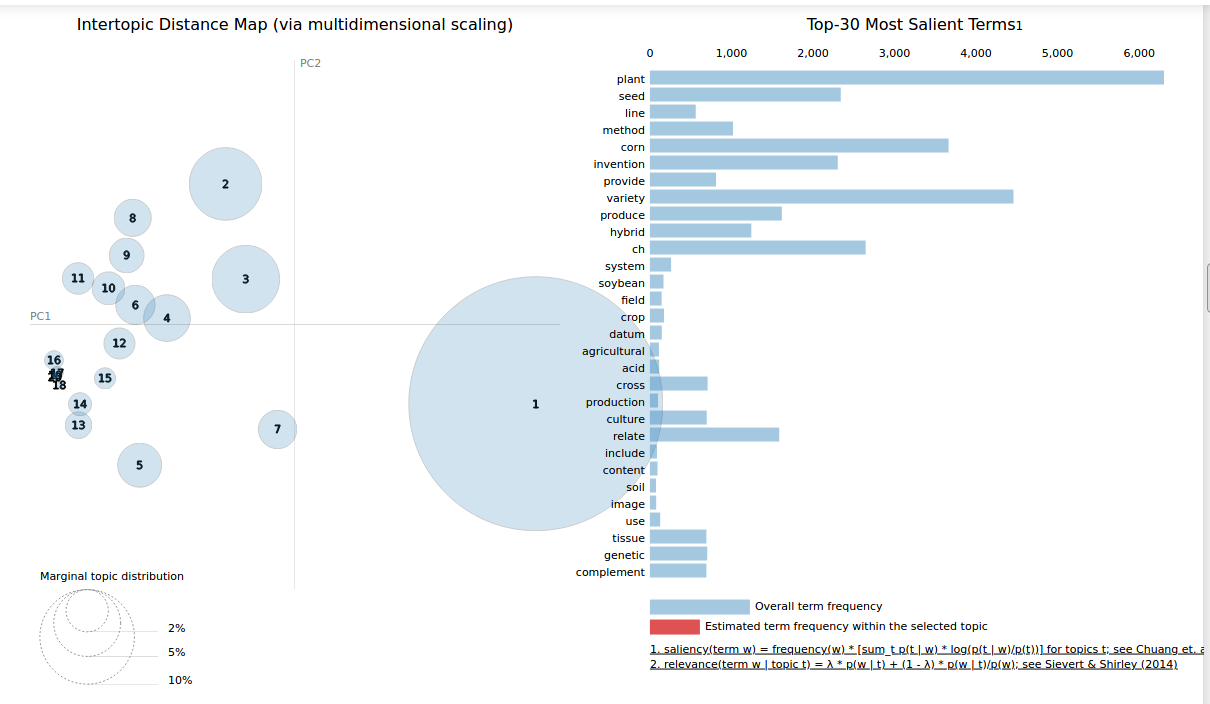


Fig. O gráfico de bolhas, cada bolha representa um tópico, o tamanho da bolha representa a prevalência do tópico e a sobreposição de bolhas aponta a similaridade entre os tópicos. O gráfico da direita, as barras representam a relevância do termo para o tópico observado.

## Expansão do dicionario

Antes de expandir o dicionário, realizamos a remoção dois tópicos que estavam muito similares. Os tópicos geraram no total de 144 termos únicos que foram submetidas ao wordnet e adicionado os sinônimos, hiperônimos e hipônimos destes termos, totalizando 616 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 901 linhas e três colunas, que foi utilizado para fazer uma classificação inicial dos documentos de patentes.

# Construção da base de dados

A base de dados composta a partir das informações de identificação documento de patente, o título do documento de patente e o seu resumo. A partir do dicionario fizemos uma classificação, onde atribuímos a cada documento de patente os tópicos a que se referem. A estrutura da base de dados é de 817 linhas e 7 colunas, sendo que 87 linhas foram removidas por não conterem título ou resumo.

# Construção do **m**odelo

Para a construção do modelo principal, os seguintes modelos foram testados, Random Forest, Naive Bayes e SVM, os principais modelos aplicados em classificação de texto. O seguinte fluxo de analise de dados foi aplicado:

pre processamento

Matriz documento-termo

Conversão da tabela de entrada 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.

Com 817 linhas e 3492 colunas.

Remoção de stopwords

Ao converter a tabela de entrada em uma matriz de documento-termo, as colunas são todas palavras de todos os documentos de patente, fazendo-se necessário a remoção de stop-words, resultando em uma tabela com 817 linhas e 3402 colunas.

Seleção de características

A seleção de características tem como objetivo selecionar as colunas que possuam maior relevância a coluna alvo. Utilizamos o método RFE para reduzir o numero de colunas para somente 20. Nossa matriz final possui 817 linhas por 20 colunas.

Modelo

RandomForest: critério de separação Gini, valor 15 de profundidade máxima, utilização de validação cruzada de 10 folds, obtendo uma acurácia igual a 0,84.

NaiveBayes:

SVM: