# Classificação de patentes utilizando Random Forest

Henrique Cursino Vieira
Orientador Prof. Dr. Nikolai Valtchev Kolev

# Tópicos

- Introdução
- Materiais e métodos
- Resultados
- Conclusão

# Introdução

"A patente é o direito exclusivo concedido a uma invenção, que é um produto ou processo que proporciona, em geral, uma nova maneira de fazer algo ou oferece uma nova solução técnica para um problema." tradução direta - World Intellectual Property Organization, WIPO

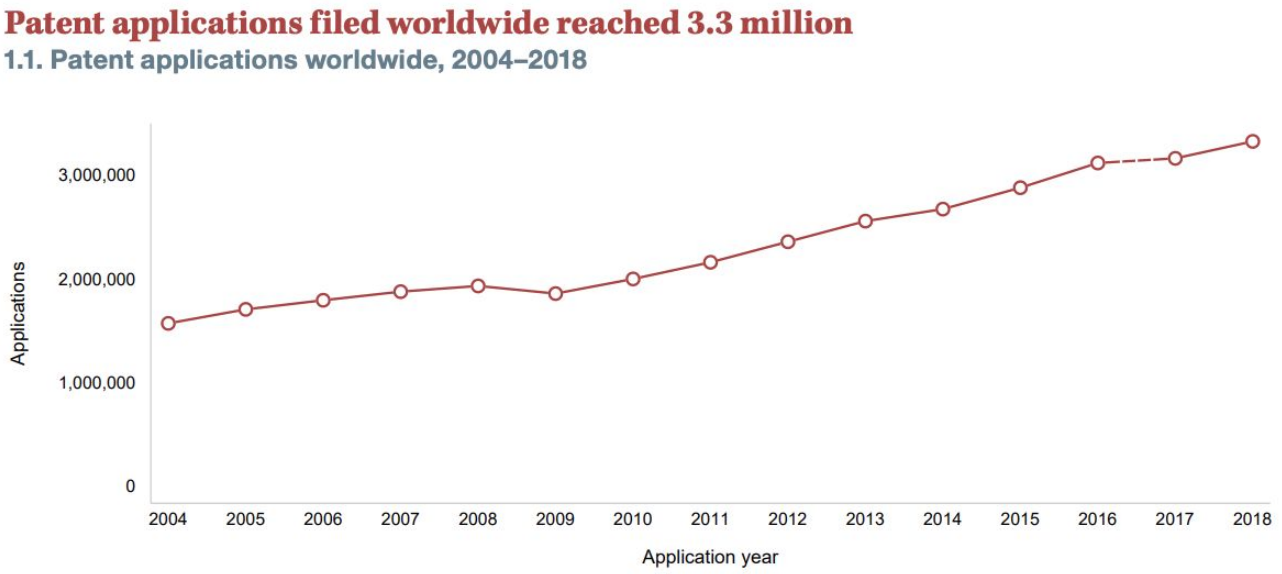



Gráfico retirado de World Intellectual Property Indicators 2019, WIPO 2019

# Introdução

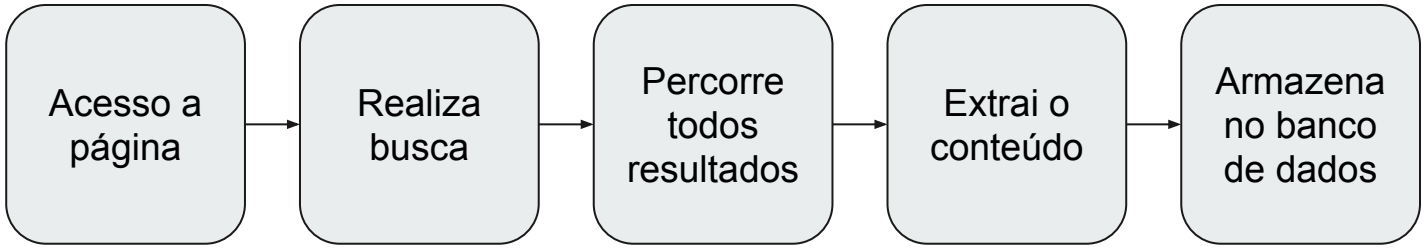
- Justificativa
  - Realizar busca por documentos mais relevantes por assunto e tema
- Objetivo geral
  - Classificação por múltiplos assuntos e temas
- Objetivo específico
  - Classificar e determinar a relevância de cada documento de patente por assunto e tema

# Materiais e métodos

1. Extração de dados
2. Construção do dicionário
3. Classificação a partir do dicionário
4. Modelagem

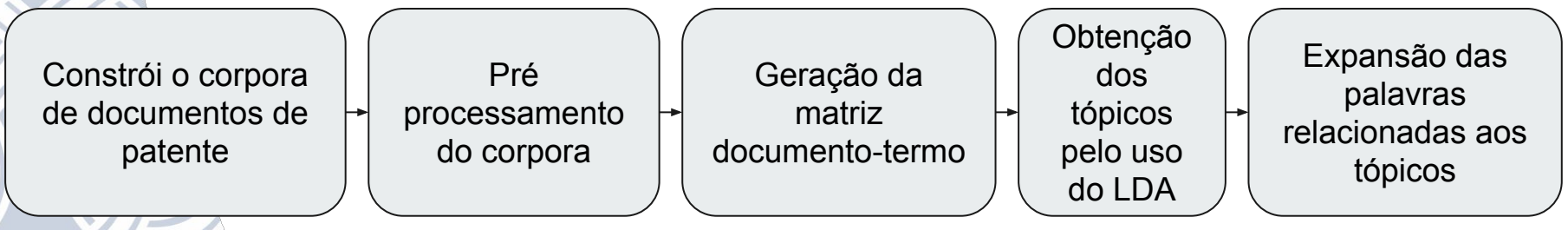
# Materiais e métodos - Extração dos dados

- Webscraping
  - www.freepatentsonline.com/
  - python
    - requests
    - BeautifulSoap



# Materiais e métodos - Construção do dicionario

- Processamento de linguagem Natural
- Corpora
  - É o plural de **corpus**. O corpus é a **junção de textos** para determinado **estudo**.
- Latent Dirichlet allocation (LDA)
  - Modelo estatístico, **agrupa** dados semelhantes a **conjuntos não observados**

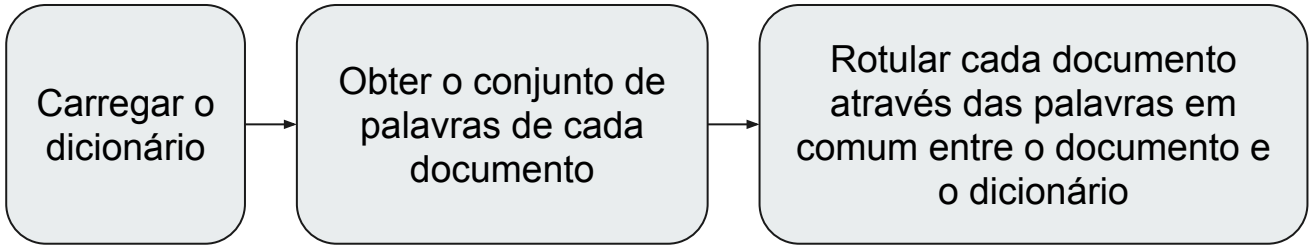


# Materiais e métodos - Validação do dicionario

- Avaliação dos tópicos
  - Correspondência entre o número de tópicos e assuntos
  - As palavras relacionadas ao tópico possuem sentido
  - Associar os tópicos gerados ao assunto

# Materiais e métodos - Classificação a partir do dicionário

- Processamento de linguagem Natural
  - tokens
  - normalização
    - pontuação
    - stopwords



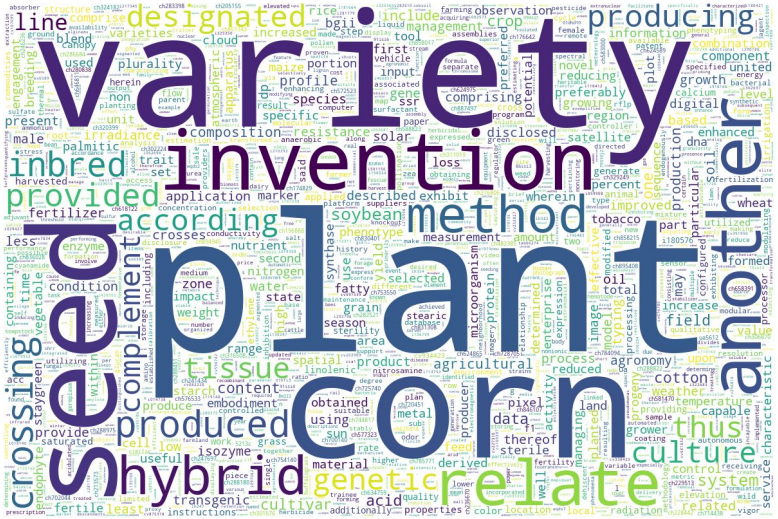
# Materiais e métodos - Modelagem

- Modelos aplicados
  - Random Forest
    - baseado em árvore
  - Naive Bayes
    - baseado em estatística bayesiana
  - SVM
    - baseado em aprendizado estatístico
- Estratégia de separação treino/teste
  - Cross Validation
- Métricas
  - média da acurácia

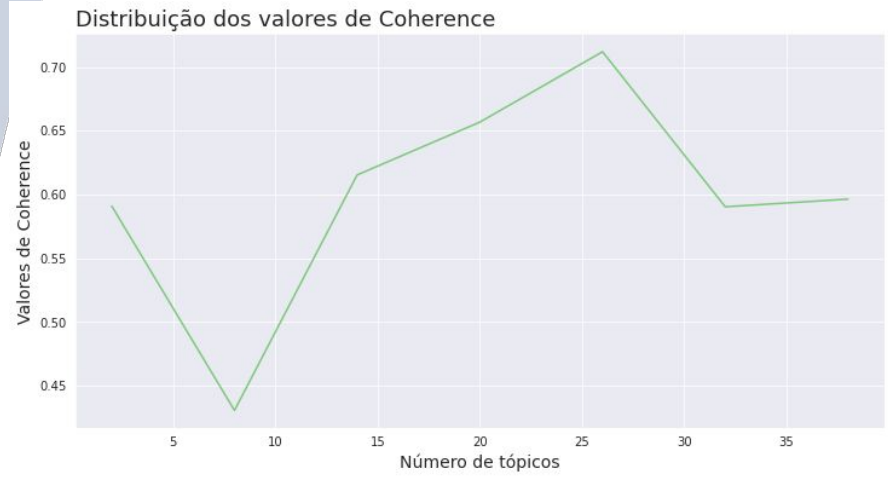
# Resultados

- Extração
  - Foi extraído uma **amostra** de **904** documentos de patentes sobre **agronomia**
- Dicionário
  - 25 **tópicos** e 901 **termos**

**Termos de maior destaque no corpora**

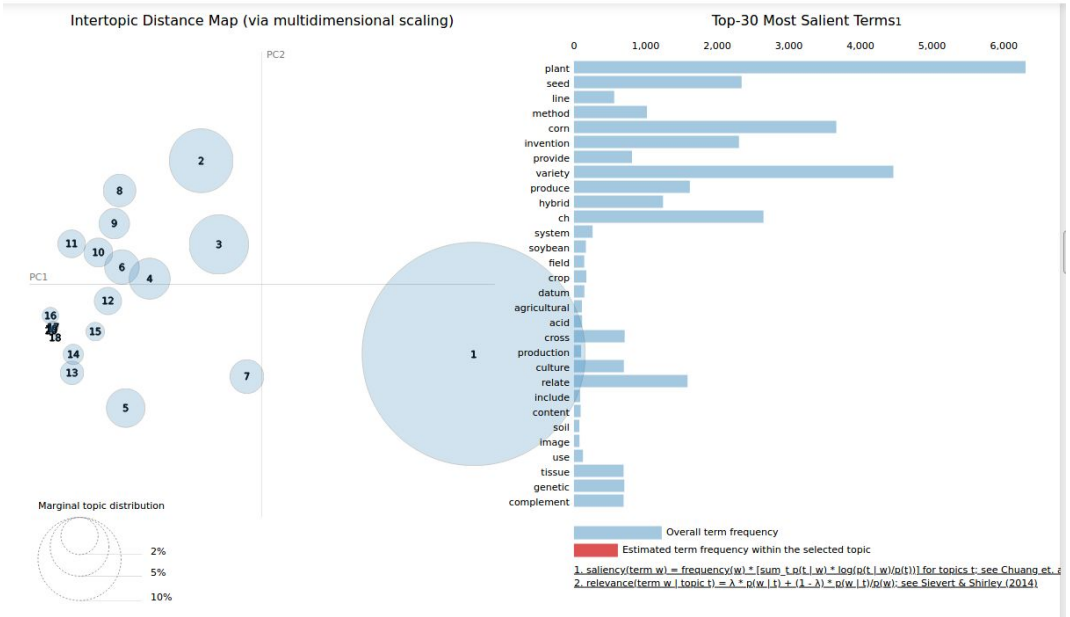


# Resultados




**Métrica de auxílio na escolha da do número de tópicos**

**painel pyLDAvis**



# Resultados

Tópicos gerados **automaticamente**

| topic | term | feature |
|---|---|---|
| digital_display_different | digital | digital |
| digital_value_field | digital | digital |
| field_datum_value | digital | digital |
| digital_display_different | display | display |
| digital_display_different | display | show |
| digital_display_different | display | exhibit |

Dicionário construído

Sinônimos, hiperônimos e hipônimos **adicionados**

# Resultados

- Random Forest
  - 10 folds
  - Acurácia média : **0.83**
- Naive Bayes
  - 10 folds
  - Acurácia média : **0.80**
- SVM
  - 10 folds
  - Acurácia média : **0.78**

# Conclusão

- Conclusão do trabalho
  - Possível classificar patentes em múltiplos tópicos
  - Construção automática de dicionário
  - Random Forest é o modelo mais adequado para este tipo de problema
- Trabalhos futuros
  - Melhorar a construção automática de dicionários
  - Testar modelos de redes neurais

# Bibliografia


- ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. World Patent Information, Elsevier Ltd, v. 37, p. 3–13, 2014. ISSN 01722190. Disponível em: <http://dx.doi.org/10.1016/j.wpi.2013.12.006>.
- ANNE, C. et al. Multiclass patent document classification. Artificial Intelligence Research, v. 7, n. 1, p. 1, 2017. ISSN 1927-6974.
- BREITZMAN, A. F.; MOGEE, M. E. The many applications of patent analysis. Journal of Information Science, v. 28, n. 3, p. 187–205, 2002. ISSN 01655515.
- LI, G. A Literature Review on Patent Texts Analysis Techniques. International Journal of Knowledge and Language Processing, v. 9, n. 3, p. 1–15, 2018.
- SHAHID, M. et al. Automatic patents classification using supervised machine learning. In: SPRINGER. International Conference on Soft Computing and Data Mining. [S.l.], 2020. p. 297–307.
- WANG, G. et al. Extraction of Principle Knowledge from Process Patents for Manufacturing Process Innovation. Procedia CIRP, The Author(s), v. 56, p. 193–198, 2016. ISSN 22128271. Disponível em: <http://dx.doi.org/10.1016/j.procir.2016.10.053>.
- WILLIAMS, H. L. How Do Patents Affect Research Investments? Annual Review of Economics, v. 9, n. 1, p. 441–469, 2017. ISSN 1941-1383.
- ZHU, H. et al. Patent automatic classification based on symmetric hierarchical convolution neural network. Symmetry, v. 12, n. 2, p. 1–12, 2020. ISSN 20738994.