

Universidade de Brasília (UnB)
Departamento de Ciência da Computação

Estratégias para Anotação de Textos do DODF

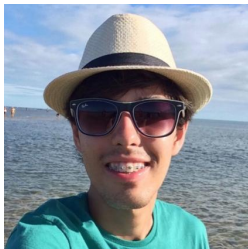
Equipe KnEDLe/Data Annotation



Equipe



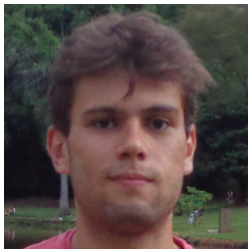
Lívia



Matheus



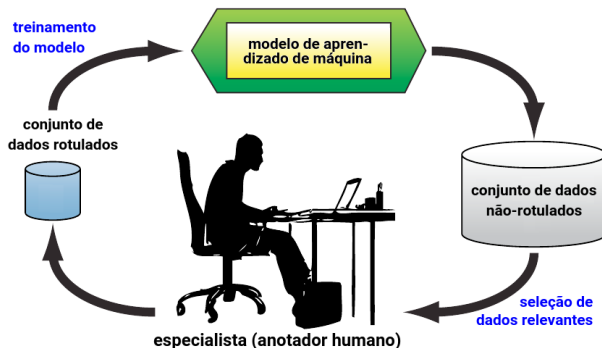
Tatiana



Vinícius

- Algumas tarefas definidas no projeto KnEDLe se baseiam em abordagens supervisionadas de aprendizado de máquina (AM)
 - classificação de textos
 - reconhecimento de entidades nomeadas
- O aprendizado dos modelos de AM supervisionados depende de dados rotulados
 - Os textos do Diário Oficial do Distrito Federal (DODF) são coletados sem informações de rótulos.

- **Princípio:** deve-se manter a acurácia de um modelo de AM treinado com menos dados rotulados quando comparado com o treinamento com todos os dados ¹




¹http://github.com/oraclesknedle/activeLearning_DODF

Representações de textos

- Os textos do DODF são naturalmente não-estruturados;
- Deve-se explorar representações estruturadas apropriadas;
- Estudos ² de representações para anotação de textos em modelos de classificação:
 - Modelo espaço vetorial (*Term Frequency-Inverse Document Frequency*);
 - *Word embeddings* (*word2vec*, *fastText*, BERT).

²http://github.com/oraclesknedle/text_experiments

- Visualização de coleções de textos para apoiar as tarefas de aprendizado de máquina e processamento de linguagem natural;
- Empregar visualização interativa para gerar representações gráficas intuitivas dos textos, visando facilitar a interpretação dos padrões e relações de similaridade;
- Integração de visualização com modelagem de tópicos para auxiliar os especialistas na anotação dos textos ³ e outros processos de descoberta de conhecimento.

³<http://github.com/oraclesknedle/LDATopicModel> 

- **Demanda:** geração de coleções de textos rotulados do DODF **padrão ouro** (*gold standard*);
- A literatura relata que a anotação manual dos textos é a abordagem mais indicada ⁴;
 - A alta qualidade de *Corpora* Gold Standard exige a análise das anotações por vários especialistas, de maneira independente;
 - Exige um grau de concordância para assegurar qualidade, tornando o processo extremamente custoso.
- Anotação manual realizada pela ferramenta TeamTat ⁵

⁴Wissler, L., Almashraee, M., Díaz, D. M., Paschke, A.. The Gold Standard in Corpus Annotation. In IEEE Germany Student Conference, 2014

⁵Islamaj, R., Kwon, D., Kim, S., Lu, Z.. TeamTat: a collaborative text annotation tool. arXiv preprint arXiv:2004.11894, 2020

TeamTat