Faculdade Ibratec - Pós em Inteligência Artificial
Processamento de Linguagem Natural (Dezembro/2019)

Professor: Marcelo Pita

Projeto 3

Classificação de Convênios do Governo Federal

1. Caracterização do problema

Convênios são os principais instrumentos de repasse de recursos do Governo Federal para a implementação de Programas de Governo nos Estados e Municípios. Os repasses dos convênios estão sempre associados a Programas de Governo e vinculados ao orçamento dos ministérios e secretarias especializadas.

Este trabalho propõe o seguinte desafio de classificação de texto: *será que é possível descobrir qual o ministério ao qual um convênio está vinculado com base na descrição do seu objeto?* Para isso a equipe deverá aplicar técnicas de PLN (Processamento de Linguagem Natural) com o objetivo de classificar convênios quanto ao seu ministério.

2. Informações sobre os dados

Uma extração de 06/12/2019 do Portal da Transparência foi disponibilizada no Github da disciplina (arquivo “convenios.csv”). Esta extração contém 10779 registros. O arquivo de dados contém os seguintes campos: NÚMERO CONVÊNIO, NOME MUNICÍPIO, OBJETO DO CONVÊNIO e NOME ÓRGÃO SUPERIOR. Foram mantidos apenas os convênios do Ministério da Ciência e Tecnologia, Ministério da Economia, Ministério da Educação, Ministério da Saúde e Ministério do Meio Ambiente. O texto a ser classificado está no campo “OBJETO DO CONVÊNIO” e o ministério a ser inferido está no campo “NOME ÓRGÃO SUPERIOR”.

3. Protótipo de software

A equipe deverá entregar um protótipo de software desenvolvido em Python, R ou KNIME. O desenvolvimento deve ser orientado pelo processo de ciência de dados,

contemplando pelo menos as seguintes fases: (1) acesso e tratamento de dados; (2) representação de dados e modelagem; (2) apresentação dos resultados.

O protótipo deve obrigatoriamente implementar as seguintes tarefas de PLN:

- Normalização de texto
- Rotulação de partes da fala
- Reconhecimento de entidades nomeadas
- Exploração de mais de um modelo de representação de texto, a saber, um dos *modelos vetoriais* e *média de vetores de palavras*

4. Relatório de projeto

A equipe deverá entregar um relatório de projeto com pelo menos os seguintes tópicos:

- Entendimento do negócio (introdução, motivação, objetivo e resultados esperados);
- Descrição dos dados e do tratamento de dados realizado, incluindo normalização de texto e outras filtragens consideradas (e.g. filtragem por partes da fala);
- Descrição dos modelos de representação dos dados;
- Descrição dos modelos analíticos explorados e sua avaliação;
- Apresentação e explicação dos resultados obtidos;
- Conclusão.

5. Detalhes da entrega

O código documentado do protótipo de software e relatório deverão estar compactados em um único arquivo que deverá ter o seguinte nome: `ibratec-pos-ia-pln_projeto3_2019.zip`.

A entrega deverá ser feita até o fim do dia dia **10/01/2020**, impreterivelmente. O arquivo do projeto deverá ser enviado por e-mail para marcelo.souza.pita@gmail.com, contendo como título “[Ibratec-PLN] Entrega Projeto 3”.