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

Professor: Marcelo Pita

Projeto 4

Busca por Licitações Similares do Governo Federal

1. Caracterização do problema

O processo tradicional de aquisição de Governo é feito através do mecanismo de licitação. O órgão de Governo responsável ou beneficiário da licitação informa em plataforma específica suas necessidades de aquisição de bens e serviços, desde que elegíveis para licitação. Durante este processo, fornecedores oferecem opções, após o qual um dos fornecedores é selecionado.

No Governo Federal não existe um controle da história de licitações, seus resultados e lições aprendidas, que serviriam para as novas licitações cujos objetos guardam similaridade com anteriores. Desta forma, este projeto propõe um mecanismo baseado em medidas de similaridade textual para automaticamente selecionar as top-N licitações mais similares a uma fornecida. A equipe deverá aplicar técnicas de PLN (Processamento de Linguagem Natural) com o objetivo de desenvolver o mencionado mecanismo de similaridade.

2. Informações sobre os dados

Uma extração de Setembro/2019 do Portal da Transparência foi disponibilizada no Github da disciplina (arquivo “licitacoes.csv”). Esta extração contém 3421 registros. O arquivo de dados contém os seguintes campos: Número Licitação, Objeto. Foram mantidos apenas as licitações do Ministério da Ciência e Tecnologia, Ministério da Economia, Ministério da Educação, Ministério da Saúde e Ministério do Meio Ambiente. O textos a serem comparados estão no campo “Objeto”.

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 das métricas de similaridade exploradas 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_projeto4_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 4”.