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

Professor: Marcelo Pita

Projeto 2

Tópicos em Compras do Governo Federal

1. Caracterização do problema

As compras de produtos e serviços pelos vários órgãos do Governo Federal constituem a execução do orçamento destino aos ministérios e suas unidades constituintes, tais como secretarias, autarquias e empresas públicas. Refletem, portanto, as despesas e investimentos do Governo Federal necessários para a prestação do serviço público.

O Governo Federal poderia conhecer melhor suas compras se conseguisse identificar, a partir das suas descrições, seus objetos. Contudo, a rotulação manual de tópicos é impossível devido ao grande volume de compras que um servidor teria que se inteirar, lendo suas descrições, para indicar quais tópicos estão presentes em uma despesa.

Este projeto consiste na implementação de um modelo de descoberta de tópicos a ser aplicado sobre as descrições textuais das compras do Governo Federal. Para isso a equipe deverá aplicar técnicas de PLN (Processamento de Linguagem Natural) com o objetivo de descobrir tópicos interessantes.

2. Informações sobre os dados

Um extração de Novembro/2019 do Portal da Transparência foi disponibilizada no Github da disciplina (arquivo “compras_governo.csv”). Esta extração contém uma amostra aleatória uniforme de 500 registros do conjunto de dados original. O arquivo de dados contém um único campo “descricao_compra”.

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

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_projeto2_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 2”.