Faculdade de Tecnologia Bandeirantes - BandTec 2^o semestre de 2015

Exercício Programa de Text Mining: classificador de intenções de perguntas

Pós Graduação em Big Data Prof. Fabrício Jailson Barth

O objetivo deste exercício programa é desenvolver um classificador de intenções de perguntas levando-se em consideração alguns artefatos já desenvolvidos ao longo da disciplina:

- Um dataset sobre um domínio específico com perguntas e intenções mapeadas.
- Um classificador gerado a partir do uso do serviço *Natural Language Classifier* (NLC) existente na plataforma Bluemix. Este classificador deverá ser considerado como *gold standard* neste estudo.
- Um processo, implementado em R, que realiza o pré-processamento e constrói o classificador. Este processo está disponível em [1]. O modelo resultante deste processo será considerado como o *baseline* neste estudo.

Para a entrega deste projeto deverão ser realizadas as seguintes atividades:

1. Aumentar o dataset para uma amostra com no mínimo 15 classes e 10 exemplos para cada classe;
2. Re-treino de uma instância do NLC e cálculo da estimativa de acurácia.
3. Execução do script em R [2] com o dataset criado, criação do modelo e estimativa de acurácia.
4. Alteração do script em R visando o aperfeiçoamento do modelo gerado pelo mesmo. Criação de um modelo e estimativa de acurácia. As alterações podem ser feitas no pré-processamento do texto, na filtragem dos atributos ou na construção do modelo.

Cada equipe deverá:

- Disponibilizar o projeto em uma conta no GitHub.
- A conta no GitHub deverá conter o dataset criado (formato CSV), todos os scripts utilizados para o desenvolvimento do projeto e um relatório descrevendo o que foi realizado durante o projeto.
- Enviar o link do projeto no GitHub para o email *fabricio.barth@gmail.com* até o dia 19 de dezembro de 2015.

[1] https://github.com/fbarth/mlr/blob/master/scripts/textClassify/classifyQuestions.R
[2] https://github.com/fbarth/mlr/blob/master/scripts/textClassify/classifyQuestions.R