Universidade Federal do Ceará
Campus de Quixadá

Mineração de Dados

Trabalho Prático – Parte 1

Aluna: Bárbara Stéphanie Neves
Professora: Lívia Almada

Setembro
2019

Universidade Federal do Ceará
Campus de Quixadá

Mineração de Dados

Trabalho Prático – Parte 1

Este documento consiste em descrever os detalhes do problema a ser resolvido e qual(is) *dataset(s)* será(ão) usado(s) para o Trabalho Prático da disciplina de Mineração de Dados

Aluna: Bárbara Stéphanie Neves

Matrículas: 388713

Professora: Lívia Almada

Curso: Ciência da Computação

Setembro
2019

Conteúdo

1 Proposta para o Trabalho Prático

1.1 Definição

O problema escolhido foi o ***Toxic Comment Classification Challenge*: Identifique e classifique comentários "tóxicos"**. Este problema foi retirado das competições da **Plataforma *Kaggle*** e se trata de um problema de **Regressão** e **Processamento de Linguagem Natural (LPN)**.

1.2 Descrição

A equipe *Conversation AI*, uma iniciativa de pesquisa fundada por *Jigsaw* e *Google* (ambos parte do *Alphabet*), está trabalhando em ferramentas para ajudar a melhorar a conversação *online*.

Uma área de estudo aborda os comportamentos *online* negativos, como comentários considerados "tóxicos", ou seja, comentários rudes, desrespeitosos ou com probabilidade de fazer alguém sair de uma discussão.

Até o momento, eles criaram uma variedade de modelos disponíveis ao público, servidos por meio da *API* do *Perspective*. Mas, estes modelos ainda cometem erros e não permitem que os usuários selecionem quais tipos de comentários tóxicos estão interessados em encontrar. Por exemplo: algumas plataformas autorizam palavrões, mas não outros tipos de conteúdo tóxico.

1.3 Objetivo

Criação de um *multi-headed model* capaz de detectar diferentes tipos de comentários tóxicos, como os que possuem ameaças, obscenidade, insultos e ódio baseado em identidade. Este modelo deve prever a probabilidade de comportamento tóxico para cada comentário.

1.4 Conjunto de Dados

Será usado um conjunto de dados de comentários das edições da página de discussão da *Wikipedia*. Este *dataset* contém textos que podem ser considerados profano, vulgar ou ofensivo.

Os comentários da *Wikipedia* foram rotulados por avaliadores humanos de acordo com o seu comportamento em:

- *toxic* (tóxico)
- *severe_toxic* (tóxico_grave)

- *obscene* (obsceno)
- *threat* (ameaça)
- *insult* (insulto)
- *identity_hate* (odio_a_identidade)

1.4.1 Arquivos

- **train.csv**: conjunto de treinamento. Contém comentários com seus rótulos binários.
- **test.csv**: conjunto de teste onde será previsto as probabilidades de conteúdo tóxico para esses comentários.
- **sample_submission.csv**: arquivo de envio de amostra no formato correto.
- **test_labels.csv**: rótulos para os dados de teste. O valor -1 indica que não foi usado para pontuação.