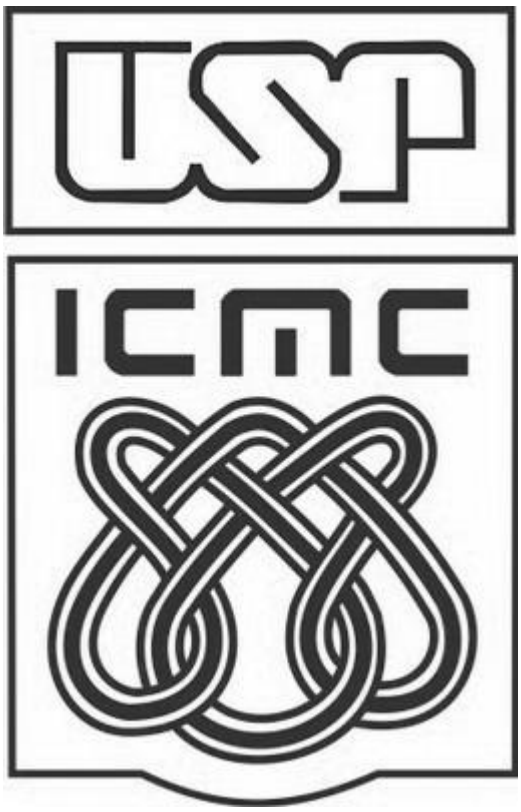
Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

PARTE 1

Vitor Albuquerque de Paula 8628220



1. Levantamento de dados

Como base de dados foi escolhida a base da competição “Detecting Insults in Social Commentary”, promovida no site kaggle.com. Esta simula comentários feitos numa seção qualquer, como um blog por exemplo, e traz informações adicionais sobre cada um deles.

2. Descrição da base de dados

2.1 Função dos arquivos

A base de dados consiste em dois arquivos .csv, sendo um o set de treinamento, utilizado na construção das regras de análise do algoritmo, e outro o set de teste, no qual as regras aprendidas são aplcadas.

2.2 Formatação dos arquivos

Cada linha do arquivo corresponde a um dado, o qual possui três colunas, sendo cada uma um campo do dado.

No arquivo de treinamento, o primeiro campo possui um dado booleano que identifica se o comentário é ofensivo (1) ou não (0). O segundo campo possui a data em que o comentário foi feito formatado como “YYYYMMDDhhmmss”, terminado com um caracter Z ou está em branco por não ter sido possível obter esta data. O terceiro campo possui o comentário feito entre aspas (“”).

No arquivo de teste, o primeiro campo possui a ID do comentário, dado único que identifica cada comentário e o segundo e terceiro campos são iguais aos correspondentes no arquivo de treinamento.

3. Objetivo da análise

Com uma análise nesse banco de dados, objetiva-se identificar comentários ofensivos a outros usuários e apenas a eles, ou seja, um comentário que ofenda um grupo de pessoas que não seja direcionado a um usuário do serviço de comentários não deverá ser considerado como ofensivo.