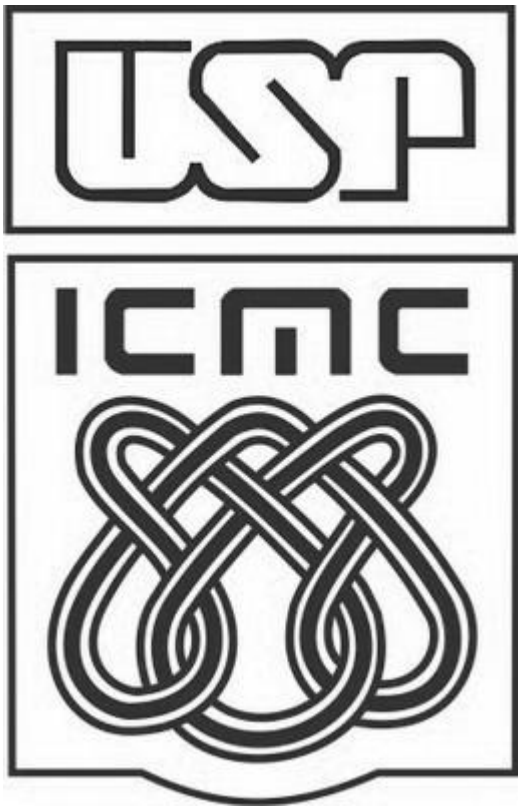
Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

PARTE 2

Vitor Albuquerque de Paula 8628220



1. Planejamento e execução

1-1. Execução do processo de mineração de dados

A execução do processo de mineração de dados se deu conforme planejado na parte anterior do trabalho. Primeiramente foi feita a filtragem do texto a ser classificado para uma otimização da classificação com uma certa "limpeza" do texto e padronização deste, então foi feito o cálculo da frequência e contagem de termos em todos os comentários através do TF-IDF, método que retira ainda mais termos desnecessários e classifica a importância dos restantes. Por fim utilizamos o algoritmo já escolhido na parte anterior, o Naive Bayes, para fazer a classificação dos dados, sendo esta classificação apenas a rotulação de um comentário como ofensivo a um usuário do chat ou não.

Os passos foram executados utilizado um algoritmo em python baseado na biblioteca sklearn (http://scikit-learn.org/stable/index.html), uma biblioteca de código aberto usada comercialmente. Esta biblioteca implementa todos os passos pensados no planejamento do projeto, com exceção da prévia filtragem do texto, sendo esta feita no próprio código utilizado.

1-2. Acuracidade no planejamento

Neste processo não foi necessária a aplicação de qualquer mudança, pois não foram encontrados problemas durante a execução do planejamento e por fim esse foi mantido.

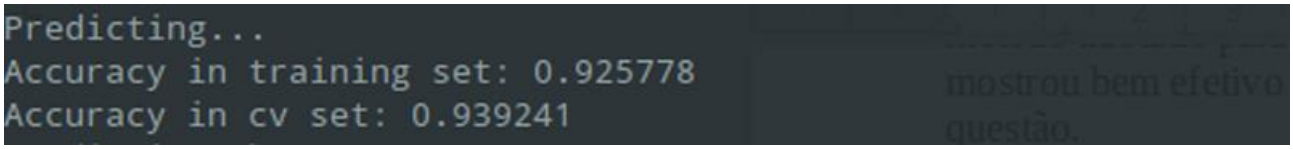
Após a execução do planejamento, a não necessidade de mudanças no planejamento inicial se tornou mais clara, pois o resultado obtido foi bastante satisfatório e mostrou a eficiência do método adotado para este problema de classificação textual. Sendo assim, no fim o planejamento se mostrou bem efetivo com uma boa precisão no que diz respeito a resolução do problema em questão.

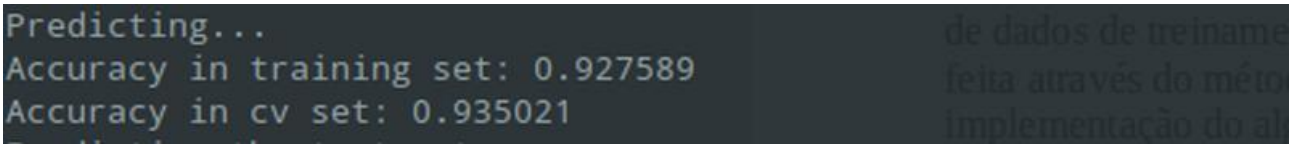
2. Resultados

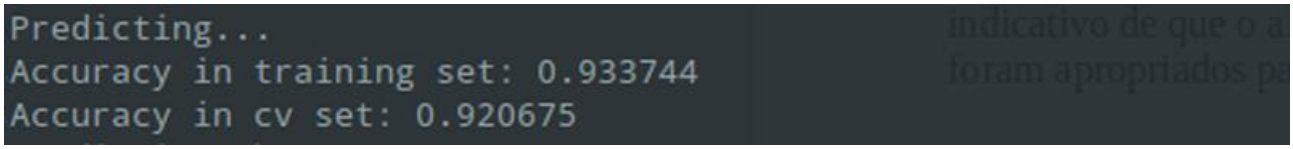
2-1. Apresentação dos resultados

Os resultados obtidos foram bem satisfatórios, mostrando um acerto a cerca de 92% no set de dados de treinamento e de 93% no de classificação. O cálculo dessa porcentagem de acerto é feita através do método score da classe GaussianNB que se encontra na biblioteca utilizada na implementação do algoritmo (sklearn).

Abaixo são mostradas imagens que mostram algumas execuções do programa:







2-2. Discussão final sobre os resultados obtidos

Os resultados obtidos foram muito satisfatórios mostrando um índice de acerto alto, indicativo de que o algoritmo escolhido no planejamento assim como o pré-processamento aplicado foram apropriados para o problema.

3. Experiências obtidas

Através deste trabalho, o grupo pôde perceber e ter uma ideia melhor de como funciona um processo de classificação de dados. Mais especificamente, foi notável como a classificação de texto é característica e requer um pré-processamento eficiente e bem planejado.

Os integrantes conheceram, no decorrer no trabalho, algoritmos e técnicas de pré-processamento que poderão ser usadas em outros problemas no decorrer da vida acadêmica e de trabalho.