Trabalho final de Classificação e Pesquisa de Dados

Arthur Zachow[1] **e Felipe de Almeida Graeff**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{azcoelho, fagraeff}@inf.ufrgs.br

Introdução

O software desenvolvido para o trabalho final da cadeira de Classificação e Pesquisa de Dados é um analizador de sentimentos de comentários sobre filmes. O fato de ter sido desenvolvido com o objetivo de ser utilizado para a avaliação de comentários de filmes não impede de forma alguma a utilização para a classificação de outras coisas, como por exemplo: Tweets e comentários do Facebook, desde que seja fornecida corretamente a entrada da forma especificada mais adiante no texto.

O funcionamento do programa e a classificação dos comentários depende grandemente do arquivo de entrada fornecido, pois ele é usado como base para os cálculos de pontuação.

O sistema de pontuação é o intervalo de 0–4 nos inteiros. Com 0 representando um sentimento ”Muito negativo”; 1 sendo um sentimento ”Negativo”; 2 representa um sentimento ”Neutro”; 3 codifica um sentimento ”Positivo”; 4 é classificado como um sentimento ”Muito positivo”.

Desenvolvimento

O software foi desenvolvido na linguagem C++, padrão C++11, com o auxílio de poucas estruturas de dados nativas, apenas as classes ”vector” e ”list” foram utilizadas além dos tipos de dados básicos (string, int, double, etc...), pois esta era uma das limitações impostas na descrição do trabalho.

As funções foram desenvolvidas e alocadas em arquivos separados da forma que os desenvolvedores julgaram conveniente no momento. Em 5 arquivos temos na pasta ”src” temos: ”file_functions.cpp”, o qual é a biblioteca a qual possui a implementação das funções diretamente relacionadas com a leitura do arquivo de entrada; ”hash_table.cpp” onde está a implementação da classe Hash Table (auto explicativo); ”main.cpp” onde está a função main do programa e a implementação dos menus e coisas feitas diretamente para a interatividade do usuãrio; ”review.cpp” ¿ a implementação da classe Review que representa um comentãrio; ”word.cpp” é a implementação da classe Word, a qual representa uma palavra a qual aparece ao menos em um comentário. A leitura do código pode ser feita ”https://github.com/Fxlipe115/CPD_Final” é encorajada pois ele não será apontado diretamente na descrição da maior parte do que foi implementado.

A classe Word possui os seguintes atributos: sum que é a soma dos valores da palavra em todos os reviews; key que é a string que é a palavra em si; occurrences que é a quantidade de ocorrências da palavra; pos é a quantidade de vezes na qual a pontuação da palavra foi ¿= 2; neg é a quantidade de vezes que a pontuação foi ¡ 2; reviews é uma lista de inteiros que são ponteiros para entradas na hash_table dos reviews onde essa

palavra está presente. Os métodos fora getters, setters e incrementadores de 1 nos atributos são apenas 2 métodos que são: wil_lower_bound, qual é o valor inferior do intervalo de confiança de Wilson; mean é o método que segue o algoritmo a seguir para o cálculo da pontuação da palavra.

O algoritmo de cálculo da pontuação da palavra é: primeiramente é lido o arquivo de entrada por completo e então utilizando os atributos da classe Word é cálculado a média simples, após isso esse valor calculado é feita a diferença para o valor neutro (2), então o resultado dessa operação é utilizado para multiplicar pelo valor do intervalo de confiança de Wilson e finalmente a esse último valor é somado 2 para retornar ao intervalo desejado [0–4].

Conclusion