UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE INFORMÁTICA

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DISCIPLINA DE CLASSIFICAÇÃO E PESQUISA DE DADOS

PROF. LEANDRO KRUG WIVES

Alunos: Artur Rossi, Henrique Delazeri, João Pedro Silveira e Silva e Vinícius Fraga Coromberque.

Análise emocional de textos publicados no *Twitter*

Porto Alegre

2019

Artur Rossi

Henrique Delazeri

João Pedro Silveira e Silva

Vinícius Fraga Coromberque

Análise emocional de textos publicados no *Twitter*

Projeto para desenvolvimento de um banco de dados e um sistema de análise de sentimentos sobre texto para aplicação na disciplina Classificação e Pesquisa de Dados.
Professor responsável: Dr. Leandro Krug Wives

Porto Alegre

2019

DESCRIÇÃO DO PROBLEMA

A área de pesquisa e classificação de dados possui uma grande importância dentro da computação. Em conjunto com o avanço das tecnologias surge a expansão no volume de dados passïveis de serem manipulados e analisados e a necessidade da criação de ferramentas e tecnologias para lidar com estes dados.

Alguns exemplos de evolução na manipulação e armazenamento de dados estão nas redes sociais. Atualmente elas contam com milhares de usuários ativos, realizando postagens, compartilhando conteúdo de mídia ou texto e reagindo a conteúdos ao mesmo tempo, sendo elas responsáveis por uma significante fração do volume de visualizações de páginas e engajamento de usuários na *Internet*.

Muitas empresas e entidades governamentais utilizam informações das redes sociais na tomada de decisões e como forma de entrar em contato com seu público alvo. Atualmente muitas empresas usam suas próprias aplicações para analisar a satisfação de seus clientes e identificar pontos a melhorar pelas mídias sociais.

Uma das formas de se fazer esta análise é dada pela análise de emoções das publicações. Esta aŕea é muito explorada dentro de campos da Inteligência Artificial, onde é passada a tarefa de análise para um computador que consegue lidar com o grande volume de dados em um curto prazo.

Dado o objetivo de criar uma arquitetura para armazenar e manipular dados de forma eficiênte e a relação das redes sociais com a evolução da ciência de dados, sua importância no contexto atual da humanidade e sua utilidade na tomada de decisões de grandes empresas e órgãos governamentais foi decidido realizar uma aplicação que realiza a análise emocional de textos publicados em mídias sociais.

Por simplificação foi escolhida uma única rede social, o *Twitter*, que possui cerca de 330 milhões de usuários ativos por dia e fornece de forma simples e bem documentada dados para uso em aplicações externas.

A aplicação terá como objetivo principal analisar textos de *Tweets* de forma a classificá-los emocionalmente. A mesma poderá ser utilizada por qualquer entidade para analisar emoções relacionadas a seus interesses com a finalidade de ajudar em tomadas de decisões. Novamente por simplificação a aplicação irá considerar apenas os aspectos negativo e positivos para cada texto.

FUNCIONALIDADES PREVISTAS

Dado um conjunto de dados iniciais já pré analisados quanto a classificação emocional a aplicação poderá ser utilizada para analisar e classificar novos *Tweets* por localidade, *hashtag,* palavra chave ou popularidade em um determinado período de tempo. Os dados poderão ser analisados em tempo real.

Quando os textos forem analisados serão retornadas informações como:

- Número de *Tweets* negativos e positivos;
- Palavras mais citadas nos textos negativos;
- Palavras mais citadas nos textos positivos;
- Análise por localidade, quando forem utilizadas múltiplas localidades.

LIMITAÇÕES

A aplicação inicialmente estará disponível apenas em inglês devido a origem do conjunto de dados iniciais, porém poderá ser modificada para funcionar em outras línguas conforme que houver um conjunto de dados para treinamento nesta linguagem.

Haverá uma limitação quanto ao conjunto de caracteres aceitos para análise, se limitando inicialmente ao alfabeto romano.

Também haverá uma limitação aos Tweets que podem ser analisados conforme critérios de privacidade do *Twitter*.

A aplicação utilizará métodos probabilísticos de classificação de texto e estará sujeita aos erros do mesmo, incluindo a dependência da qualidade dos dados de treino para geração de bons resultados.

CRONOGRAMA DE ATIVIDADES

O trabalho será dividido nas seguintes etapas (prazos):

- Pesquisa de ferramentas, bibliotecas e formatos de representação que simplifiquem a construção da aplicação, permitindo um maior foco dos esforços na construção do Banco de Dados e das funcionalidades (25/10);
- Idealização e planejamento do banco de dados (01/11);
- Montagem da comunicação com o *Twitter* (01/11)*;*
- Pesquisa e preparação do conjunto de dados para treino (01/11);
- Montagem do banco de dados (08/11);
- Construção do processo de treinamento da ferramenta (08/11);
- Construção do processo de análise de *Tweets* (15/11*)*;
- Idealização e construção da forma de exibição dos resultados (22/11).

Todo o grupo estará inicialmente responsável por todas as etapas do processo e as atividades serão divididas conforme disponibilidade e dificuldades encontradas durante o processo de desenvolvimento.

FERRAMENTAS E BIBLIOTECAS A SEREM UTILIZADAS

A aplicação será desenvolvida na linguagem Python, utilizando bibliotecas para serialização de dados, escrita e leitura em disco, manipulação e análise de Strings e cálculos estatísticos.

Para desenvolvimento do código em grupo será utilizado um repositório no *GitHub* que permite o armazenamento, edição e versionamento do projeto de forma *online*.

PyCharm será utilizado como IDE padrão do projeto, podendo ser utilizado outras IDEs conforme necessidade e/ou facilidade.

O *Trello* será utilizado para divisão e planejamento de tarefas, permitindo que sejam criadas e dívidas tarefas entre o grupo de forma simples e organizada.

Para obtenção de dados atuais da rede social será utilizada a API (*Application programming interface*, interface de programação de aplicação) do próprio *Twitter* por via de uma aplicação já registrada e autorizada.

O conjunto de dados de entrada será extraído do Kaggle que possui diversas opções, das quais selecionamos a que melhor se adaptar a nossa ideia de projeto.

PROJETO DE ARQUIVOS

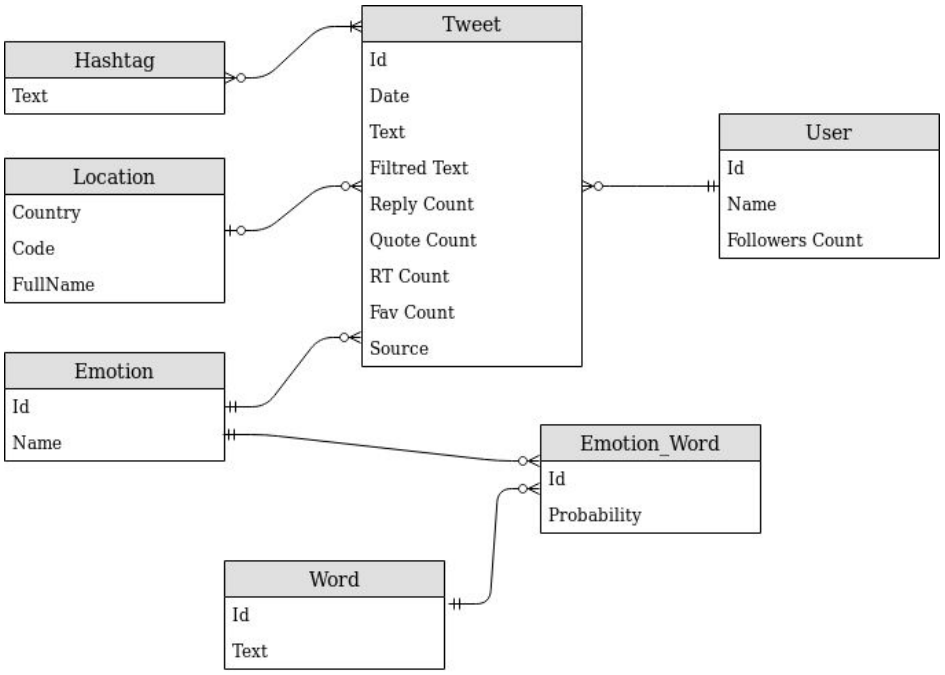
Inicialmente serão desenvolvidas estruturas baseadas nativamente nas entidades do *Twitter*, como a entidade *Tweet* com código de identificação (*Id*), data (*Date*), texto (*Text*), número de respostas (*Reply*), citações (*Quote*), *re-tweets* (*RT*) e favoritos (*FAV*), dispositivo fonte (*Source*), local (*location*) e *hashtags* associadas (Twitter Developers, 2019).

Cada *Tweet* estará relacionado a um *User*, que representa os usuários da rede social com nome (*Name*), código de identificação (*Id*) e número de seguidores (*Followers Count*), podendo estes ter vários *Tweets*.

A entidade *Hashtag* pode estar relacionada a um ou vários *Tweets* e será utilizada para filtragem dos resultados, ela possui apenas seu nome (*Text*). *Location* segue os mesmo padrões e funcionalidade da entidade *Hashtag* contendo um país (*Country*), um código (*Code*) e um nome completo (*FullName*).

A entidade *Emotion* será responsável por armazenar as emoções que serão analisadas. Cada *Tweet* analisado pode estar relacionado a uma *Emotion*. *Word* é responsável por armazenar as palavras analisadas nos dados de treino e em conjunto com *Emotion_Word* irá relacionar cada palavra com a sua probabilidade de estar relacionada a uma emoção da entidade *Emotion* (Figura 1).

Figura 1 - Diagrama de entidades da aplicação.



Fonte: autoria própria.