Relatório Final do Projeto de BD

s a d b o y s ☹

Arthur Costa Lopes - **157699**
Gabriel Souza Franco - **155477**
Henrique Noronha Facioli - **157986**
Isadora Sophia - **158018**
Lauro Cruz e Souza - **156175**
Lucas Alves Racoci - **156331**
Luiz Fernando Rodrigues da Fonseca - **156475**
Matheus Diamantino - **156740**
Thiago Silva de Farias - **148077**
Willian Tadeu Beltrao - **157595**

1. Resumo

O problema escolhido para ser tratado condiz com o tópico de saúde psicológica, em especial, a depressão. A aproximação do problema seria feita a partir do estudo de dados que dizem respeito à depressão e suicídio, como a quantidade de casos de suicídios pelos estados dos Estados Unidos da América. Em seguida, seria extraído uma série de tweets dessas regiões de forma a correlacionar índices de depressão e suicídio por estado, com o sentimento atribuído ao tweet classificado usando técnicas de machine learning, e os casos de suicídio. O objetivo é investigar se há uma relação entre estes conjuntos de dados, o que permitiria analisar possíveis sintomas, vocabulários ou comportamento de pessoas que sofrem de depressão e se os dados conferem com a quantidade de suicídios. Também foi feito um grafo entre os usuários para saber se usuários com tweets mais tristes ou felizes seguem usuários tristes ou felizes.

2. Requisitos

Para resolver o problema descrito na seção anterior, os dois modelos conceituais adotados atenderam aos seguintes requisitos:

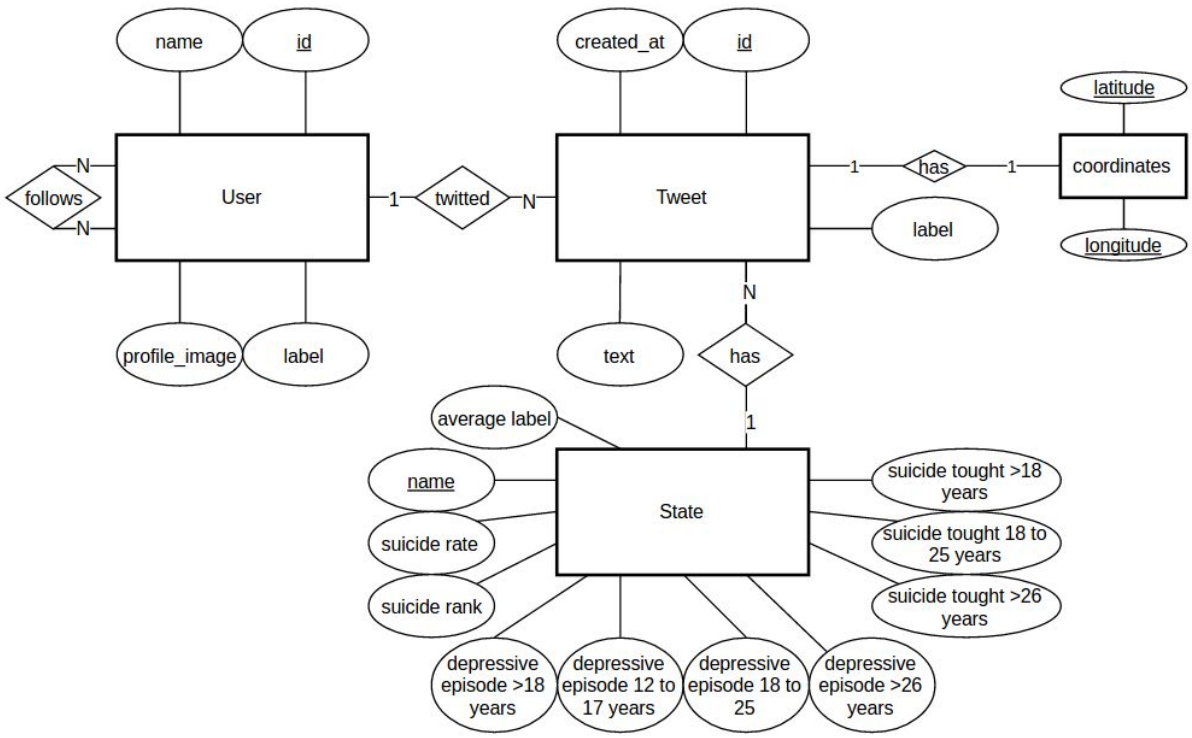
- O primeiro modelo guarda os tweets úteis que estão localizados em algum estado dos Estados Unidos, e os usuários que realizaram estes tweets.
- Entre os usuários e tweets deste primeiro modelo seriam necessárias arestas para saber que usuários publicaram quais tweets, para depois fazer uma média do sentimento (label) do usuário.
- Também haveriam arestas entre os usuários, indicando qual usuário segue qual usuário.
- E por fim existem os estados americanos e suas taxas de suicídio e depressão, e um relacionamento entre estado e tweet para calcular a média de sentimento em cada estado.
- Como, depois de filtrar os tweets em solo americano e os usuários que postaram estes tweets, houveram poucas relações entre os usuários, foi feito um segundo modelo para atender somente ao problema de visualização do grafo de usuários.
- No segundo modelo também tinham os tweets, e nesse caso eram só os em língua inglesa, e também há os usuários que publicaram estes tweets e a média do sentimento para cada usuário tirado da relação entre usuários e tweets.
- Seguindo estes requisitos, foi escolhido um banco de dados de grafos para resolver os problemas propostos para o projeto.

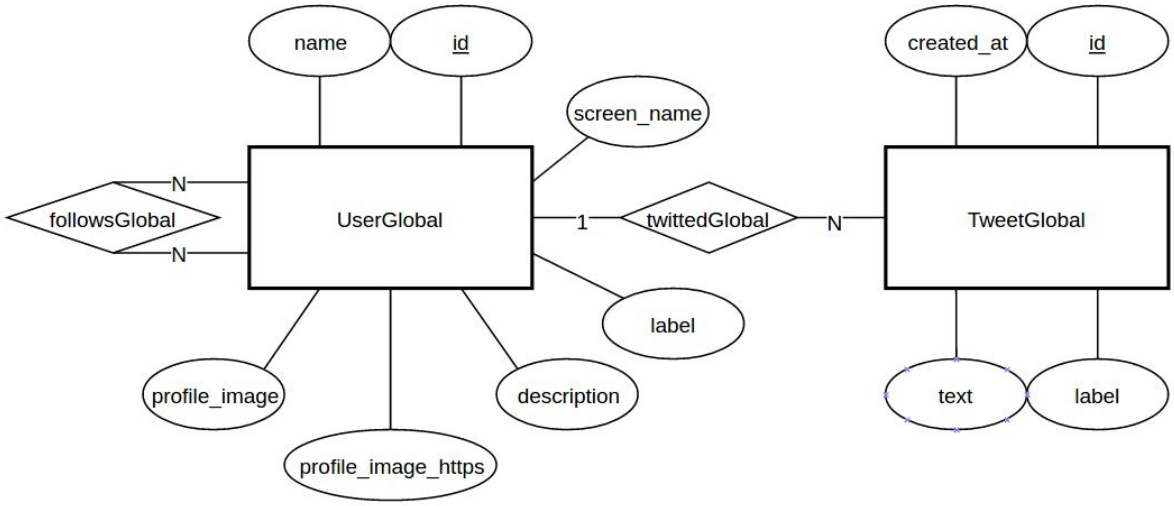
3. Sites

Os primeiros sites para extração de dados seriam fontes de pesquisa quanto à taxas de suicídio por região, idade e gênero, como [1], [2], [3] e [4]. Os dados do site [1] foram retirados de imagens e tabelas manualmente e convertidos para CSV. Os dados do site [2] foram baixados em arquivos xlsx de acordo com os artigos que pareceram interessantes. Os sites [3] e [4] possuíam sistemas de query (CoDQL e WISQARS) para filtragem de dados relativos a suicídio, sendo o [3] sobre países e o [4] sobre os EUA e seus estados. A segunda fonte partiria do próprio twitter [5], ao qual seria utilizada a sua API de forma a minerar a maior quantidade de dados possível pelas regiões do globo. Ou seja, o segundo banco de dados partiria de nossa própria extração de dados. Um terceiro banco de dados é considerado para, possivelmente, realizar a classificação de sentimentos, em que o site [6] e [7] disponibiliza. O site [8] disponibiliza uma API que auxiliou no download de tweets. O site 9 disponibiliza um dataset para a classificação binária de sentimento. O site 10 disponibiliza um otimizador para a função de perda [11] utilizada para o classificador de sentimentos.

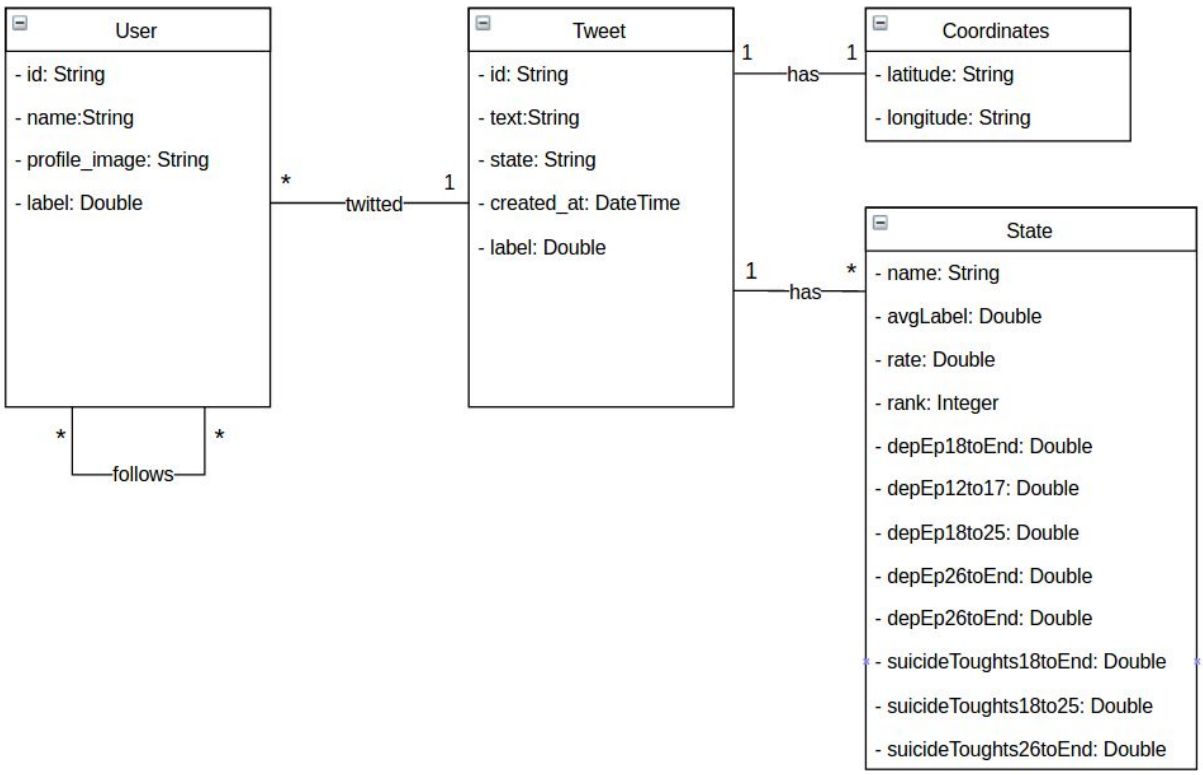
4. e 5. Modelagem

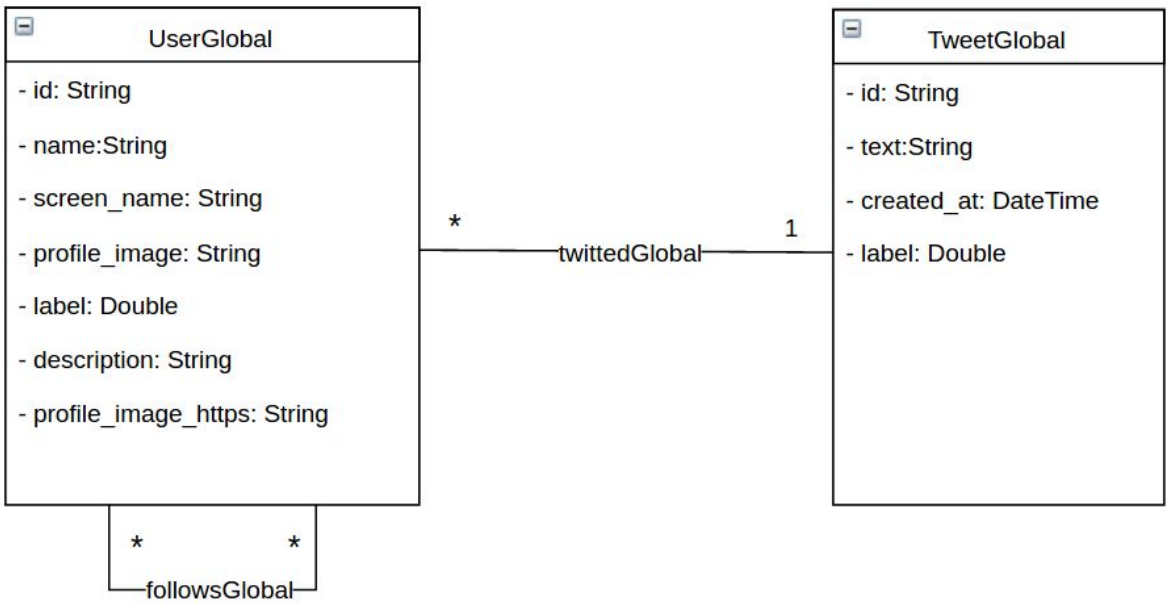
ER



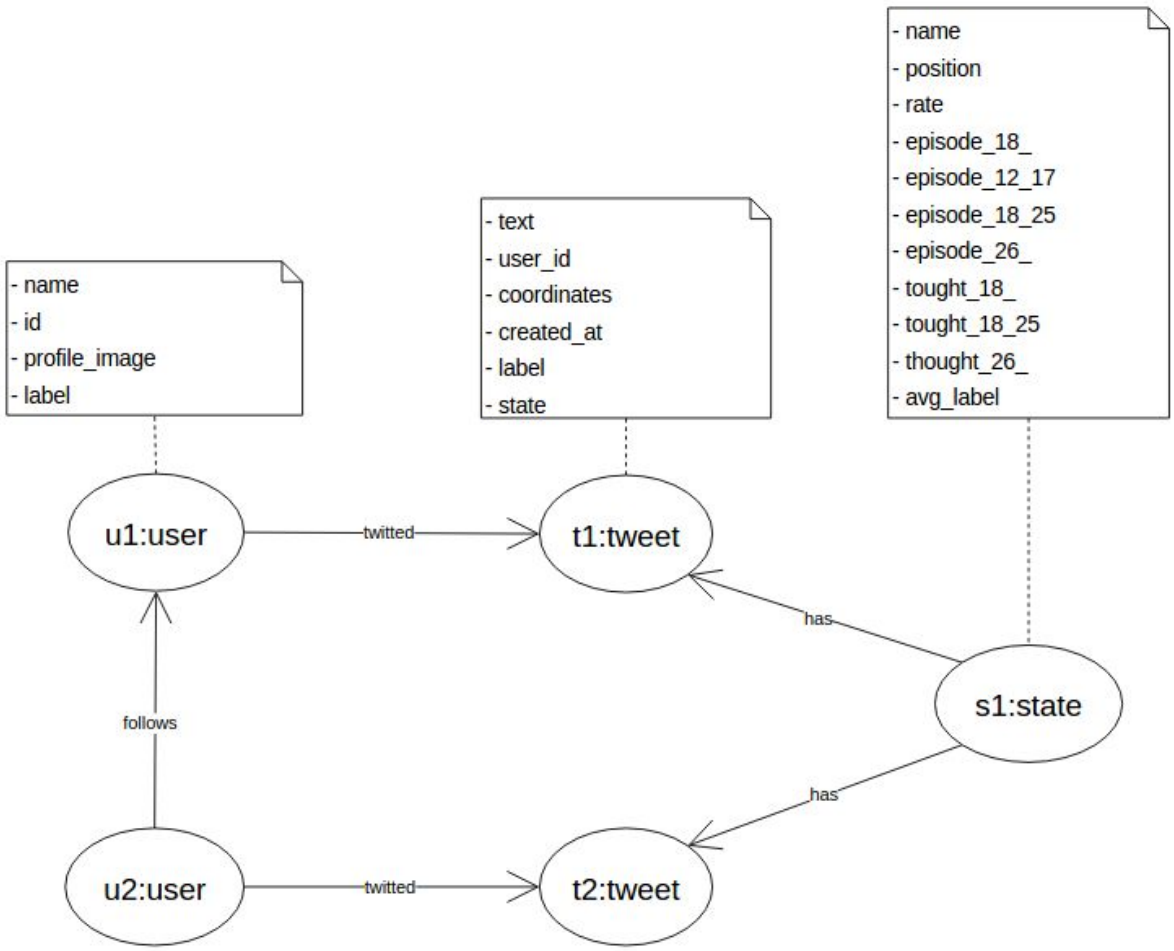

name
id
created_at
id
screen_name
N
followsGlobal
N
UserGlobal
1
twittedGlobal
N
TweetGlobal
label
profile_image
description
text
label
profile_image_https

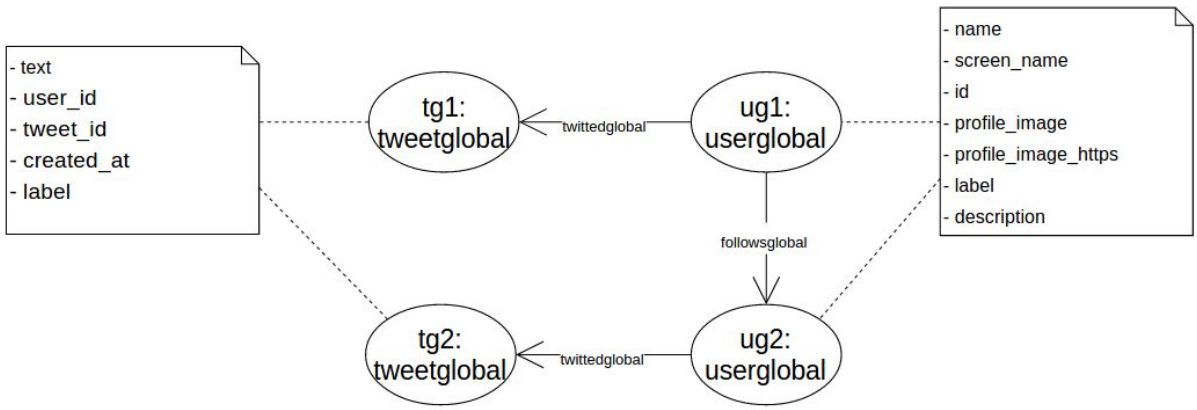
UML


User
- id: String
- name:String
- profile_image: String
- label: Double
*
twitted
1
Tweet
- id: String
- text:String
- state: String
- created_at: DateTime
- label: Double
1
has
1
Coordinates
- latitude: String
- longitude: String
1
has
*
State
- name: String
- avgLabel: Double
- rate: Double
- rank: Integer
- depEp18toEnd: Double
- depEp12to17: Double
- depEp18to25: Double
- depEp26toEnd: Double
- depEp26toEnd: Double
- suicideToughts18toEnd: Double
- suicideToughts18to25: Double
- suicideToughts26toEnd: Double
*
*
follows



Mapeamento para o Grafo





6. Esquema

O banco de dados de grafo por ser um modelo novo ainda não apresenta uma descrição formal de esquema que seja consenso, mas, assim como mostrado nas figuras acima, podemos dizer que nosso esquema trata-se de dois grafos diferentes, um apenas com estados dos EUA e outro global, contendo grande parte dos tweets coletados, essa divisão foi útil devido a natureza dos dados coletados sobre suicídio e depressão e facilitou em muitos aspectos a organização e prevenção de falhas durante o andamento do projeto.

O primeiro, contendo os tweets dos EUA, com seus respectivos usuários pode ser descrito como uma família de três conjuntos de nós: User, Tweet e State e três famílias de conjuntos de arestas: Twitted, com arestas ligando nós de User com nós de Tweet; Follows, com arestas ligando nós de User com nós de User e Has, com arestas ligando nós de State com nós de Tweet.

O segundo, semelhantemente pode ser descrito como uma família de dois conjuntos de nós UserGlobal e TweetGlobal e uma família com dois conjuntos de aresta: TwittedGlobal, com arestas ligando nós de UserGlobal com nós de TweetGlobal; FollowsGlobal, com arestas ligando nós de UserGlobal com nós de UserGlobal.

7. Análise

A análise do projeto foi basicamente feita de forma visual e foi dividida em duas partes, isso depois de cada tweet ter o seu sentimento classificado pelo algoritmo de machine learning. A primeira consistiu em fazer um mapa de calor (heatmap) dos Estados Unidos, onde em cada estado haveria uma bola indicando a taxa de suicídio, porcentagem de intenção de suicídio e a porcentagem de pessoas com depressão para aquele estado, sendo que neste caso foram usados aproximadamente 57 mil tweets.

A conclusão foi que aparentemente existe uma relação direta entre o sentimento demonstrado pelos tweets de um estado e suas taxas de suicídio e depressão. Porém, vale ressaltar que, pela forma não uniforme com que os pontos estão distribuídos no heatmap (algumas regiões quase não tem pontos), a intensidade da cor no heatmap pode estar mais relacionada com um maior número de tweets em uma certa região do que propriamente o

valor dos sentimentos desses tweets (fenômeno que acontece devido ao fato de o raio dos pontos aumentar quando se diminui o zoom do mapa).

A segunda análise consistiu de fazer um grafo entre usuários, partindo de um usuário inicial e buscando as pessoas que ele segue e quem essas pessoas seguem, existindo dois níveis. Cada usuário teve sua média de sentimento calculada dadas pelos seus tweets, sendo que neste caso foram usados 1 milhão de tweets e aproximadamente 1200 usuários. Isso teve que ser feito pois no primeiro caso havia poucos usuários e relações entre eles considerando apenas os tweets nos Estados Unidos. Esta análise mostrou se usuários mais tristes ou felizes tendem a seguir usuários mais tristes ou felizes. Também, se clicar em um usuário, serão mostrados em um grafo o usuário e no máximo 500 tweets dele, para ver o nível dos sentimentos dos tweets de um determinado usuário.

A conclusão foi que o número de nós não foi suficiente para termos uma boa noção do grafo. Além disso, o fato de a média de sentimento da grande maioria dos usuários ser na faixa de 0.4 a 0.6 faz com que a cor dos nós seja a mesma, o que dificulta a análise.

8. Papel das Duplas

# Dupla	Integrantes	Papel da Dupla
1	Isadora Sophia Matheus Diamantino	Machine learning para classificação de sentimentos dos tweets
2	Arthur Costa Lopes Gabriel Souza Franco	Coleta e filtragem de dados do Twitter
3	Lucas Alves Racoci Luiz Fernando Rodrigues da Fonseca	Modelo de Banco de Dados e prover comunicação entre os componentes do cluster
4	Henrique Noronha Facioli Thiago Silva de Farias	Apresentação visual dos Dados
5	Lauro Cruz e Souza Willian Beltrao	Coleta de dados sobre depressão e suicídio

Dupla 1: Machine learning para classificação dos tweets

A dupla 1 foi responsável pela parte relacionada ao processamento de dados (com tópicos relacionados a IA e machine learning), de forma a lidar com a classificação dos dados.

A partir do banco de dados "Large Movie Review Dataset" [9], foi treinado e desenvolvido um classificador de sentimentos, o qual classifica um determinado texto em um sentimento ruim (0) ou bom (1). Assim, o modelo foi aplicado para cada tweet de forma a classificá-los em um sentimento/emoção. Permitindo, dessa forma, visualizar a relação de taxas de depressão/suicídio com os sentimentos de tweet a partir da região e idade, na

parte final do trabalho. A relação entre as informações obtidas é feita por meio da visualização.

Este trabalho foi dividido em duas partes principais:

1) Definição e treinamento;

2) Aplicação do modelo para classificação dos tweets e a comunicação com o banco de dados (query para buscar os tweets e atualizar os valores de sentimento com interface implementada pela dupla responsável);

Na fase de definição do modelo, foi adotado dois modelos dentre deep learning: LSTM e uma arquitetura que utiliza redes convolucionais 1D (CNN), com métodos como max pooling e fully-connected layers para extrair características relevantes da amostra sendo analisada. Apesar de ambos os modelos terem apresentado precisão em torno de 82%, o segundo modelo mostrou-se mais eficiente na prática para classificação dos tweets.

Na fase de aplicação do modelo, foi adotado o mesmo pré-processamento utilizado no dataset "Large Movie Review Dataset" [9] como parâmetro do modelo: foi implementado um dicionário para cada palavra encontrada no dataset, com sua respectiva frequência (0 para a palavra mais frequente, 1 para a segunda palavra mais frequente etc.), e cada palavra de um dado tweet era traduzida para a sua respectiva frequência. Assim, utilizando a interface para acesso ao banco de dados de grafo dos tweets, foi possível extrair os dados para a classificação e atualizar o sentimento a partir do output produzido pelo modelo.

Dupla 2: Coleta de dados Twitter

A dupla 2 foi responsável pela coleta e filtragem de dados do Twitter. Após o estudo inicial das necessidades do grupo a dupla extraiu mais de 11 milhões de tweets com grande concentração nos Estados Unidos. Para fazer a extração foi utilizado um método recursivo para a escolha dos usuários. Primeiramente foram selecionados celebridades e grandes canais de TV como usuários alvo, e em seguida, foram coletados seus seguidores e recursivamente seus seguidores. A cada iteração os seguidores eram colocados numa fila e os tweets do usuário atual eram salvos em um banco de dados.

Para a coleta foi criado um script utilizando a API tweepy [8] que faz download dos tweets representados por um objeto JSON. Para guardar os tweets utilizamos o MongoDB que armazena diretamente o objeto JSON, facilitando o acesso, a filtragem e a transferência para outras duplas.

Uma das grandes dificuldades de retirar dados do Twitter são as limitações da API em relação a quantidade de downloads (limite de 3000 tweets por usuário e download de blocos com apenas 200 tweets). No total foram necessários cerca de 10 dias para realizar o download dos dados.

Ao final do processo foram obtidos uma grande quantidade de tweets em inglês (cerca de 60% dos mais de 11 milhões) e mais de 100 mil tweets com localização. Como a opção por fornecer a localização do tweet é opcional, poucos usuários ativam essa opção por uma questão de privacidade, dificultando o agrupamento dos dados.

Após a filtragem, que envolve selecionar tweets apenas realizados nos EUA e em inglês, os tweets foram enviados para a dupla 3 para preparar os dados para a análise de sentimentos.

A região foi escolhida com base na disponibilidade dos dados coletados pela dupla 5.

Dupla 3: Modelo de Banco de Dados e Prover Comunicação entre os Componentes do Cluster

A dupla 3 se responsabilizou por encontrar o melhor modelo de Banco de Dados para representar as informações coletadas dados os recursos disponíveis. A princípio pensou-se que o mais apropriado seria usar um Banco de Dados em grafo devido a alta complexidade semântica dos dados coletados. A inspiração inicial para usar banco de dados em grafo veio do fato de facilitar a recuperação de dados que sejam muito interconectados. Os bancos de dados em grafo permitem, “by design”, a recuperação rápida de dados organizados em complexas estruturas semânticas, como seria o caso de dados de tweets e usuários que se interconectam. Para estes fins resolveu-se adotar o SGBD Neo4j.

Como a API utilizada para extrair dados do Twitter retornava documentos JSONs, a dupla 2 decidiu guardar os tweets e usuários no banco de documentos MongoDB. Então, a adaptação e passagem destes dados para o grafo do neo4j e a maneira como eles seriam organizados ficou de responsabilidade da dupla 3.

Para obter os tweets dos estados americanos, foram extraídos tweets dentro de um retângulo de coordenadas geográficas que indicaria os Estados Unidos, e para conseguir exatamente os tweets dos estados americanos, foi utilizada a API geopy Nominatim do Open Street Map. Foram conseguidos aproximadamente 57 mil tweets nos estados unidos para o primeiro grafo e aproximadamente 1 milhão de tweets sem localização para o segundo grafo.

Outra responsabilidade da dupla 3 foi integrar ao grafo os dados coletados pela dupla 5, de suicídio e depressão por faixa etária por estado. Isso torna possível estudar a relação que esses dados apresentam com com a análise de sentimento realizada.

Outro ponto importante foi o acesso provido ao banco de dados para as duplas de machine learning e apresentação visual dos dados. Para isso, foram feitas APIs para que eles acessassem os dados de que precisassem, como o valor dos dados de suicídio por estado, e também atualizassem valores no grafo, como os valores dos sentimentos analisados das frases dos tweets.

Para que o acesso aos dados fosse feito de maneira mais rápida, foram criados índices no neo4j para aumentar a eficiência.

Dupla 4: Apresentação visual dos dados

A dupla 4 ficou responsável pela apresentação visual dos dados com qualidade, facilitando a compreensão dos resultados obtidos e tornando o projeto esteticamente mais atrativo.

Foi usada a API do Google Maps para a geração do mapa de calor (heatmap) dos Estados Unidos, onde em cada estado temos um círculo indicando a taxa de suicídio, porcentagem de intenção de suicídio e a porcentagem de pessoas com depressão para aquele estado (quanto maior o raio do círculo, maior a porcentagem de pessoas que cometeram suicídio ou têm depressão naquele estado). Além disso, para cada um dos 57 mil tweets coletados, foram criados pontos no mapa referentes à sua localização (local de

onde o tweet foi enviado), e atribuídos pesos referentes ao valor de sentimento calculado para aquele tweet.

Para os grafos, foi usada a API do vis.js. No primeiro grafo temos cada nó representando um usuário, a cor do nó representando o sentimento médio do usuário (média do valor de sentimento dos seus tweets), e as arestas ligando dois nós representando que um usuário segue o outro. No segundo grafo (que pode ser obtido ao se clicar no usuário desejado), temos o nó central representando o usuário e arestas ligando esse nó a cada um dos tweets referentes à esse usuário (a cor dos nós dos tweets diz respeito ao valor analisado para o seu sentimento).

Por fim, foi feito um campo para que o usuário possa analisar o sentimento de alguma frase. Basta ele inserir a frase desejada e pedir para que a análise seja feita (esta análise é referente ao trabalho realizado pela dupla 1).

Dupla 5: Coleta de dados sobre depressão e suicídio

A dupla 5 ficou responsável por coletar dados de saúde relacionados à depressão e suicídio. Estes dados envolvem taxas de ocorrência, sintomas e atitudes relacionadas a esse tipo de patologia, levando-se em conta a faixa etária, gênero e região.

Os dados tiveram como fonte a Web e foram obtidos por extração manual, download direto ou download filtrado por meio de sistemas de query. Para cada site fonte foi utilizado o procedimento adequado para a extração.

Após a extração, foi feita a interligação, limpeza e organização dos dados obtidos. Para isso removemos as redundâncias, verificamos sua consistência, se para um mesmo grupo alvo não havia divergência de dados nas diferentes fontes, e eliminamos atributos desnecessários para o trabalho.

Os dados foram obtidos em sua maioria no formato csv, assim foi feita uma padronização neste formato, convertendo-se o restante dados para csv também, facilitando a inclusão de dados no banco.

Essas informações foram consolidadas, melhor organizadas no banco de dados e utilizadas pela dupla 1 para obter informações relevantes dos tweets coletados pela dupla 2.

9. Destaques do Trabalho:

Desafios de extração e Data Cleaning

Pré-processamento para o classificador de sentimentos

Para realizar a classificação de uma determinada sentença, é necessário, inicialmente, processar o texto de forma a se apresentar como um input válido para o classificador. Assim, foi criado um dicionário de palavras baseado no dataset de treino, "Large Movie Review Dataset" [9], em que cada palavra possui um valor de frequência associado a ela.

Dessa forma, cada palavra do texto de input é associado a um valor de frequência. Como cada sentença foi definido um número máximo de palavras (chamando esse valor de N, por exemplo), ela é convertida em um vetor de inteiros

de tamanho N. Esse vetor é, então, levado como entrada para o modelo de classificação de sentimentos.

Localização geográfica dos tweets

Para fazer a análise utilizando o mapa de calor, foi preciso pegar apenas os tweets que tinham localização geográfica dentro dos Estados Unidos. Para isso foi utilizada a API geopy Nominatim do OpenStreetMap. O desafio de usar esta API é que ela tem uma quota máxima possível que é permitida de requisições, que é de uma requisição por segundo. Assim, vários erros precisaram ser tratados para se conseguir apenas tweets em algum estado americano, e o tempo total de processamento apenas do arquivo mongo2neo.py foi de aproximadamente 48 horas. Abaixo está um recorte de uma parte do tratamento de erros e uso da API.

```
# Para cada tweet encontrado, checa se ele esta dentro de um quadrado
# que engloba os Estados Unidos e chama o geolocator para conseguir
# as informacoes dos locais do tweet, sendo que e necessario dormir
# durante um segundo para nao exceder a quota da API
i = 1
for tweet in cursor:
    if i > offset:
        place = tweet['place']
        coordinates = tweet['coordinates']
        location = None

        if coordinates != None:
            lon = coordinates['coordinates'][0]
            lat = coordinates['coordinates'][1]
            if -170 < lon < -60 and 12 < lat < 72:
                tries = 3
                while tries > 0:
                    try:
                        time.sleep(1)
                        location = geolocator.reverse("%f, %f" % (lat, lon))
                        break
                    except GeocoderTimedOut as e:
                        print "GeocoderTimedOut"
                        time.sleep(1)
                        tries -= 1
                    except GeocoderQuotaExceeded as e:
                        print "QuotaExceeded"
                        time.sleep(10)
                    except GeocoderUnavailable as e:
                        print "GeocoderUnavailable"
                        time.sleep(30)
```

Breves descrições ou recortes de algoritmos implementados

Classificador de sentimentos

Como mencionado anteriormente, foi utilizada uma arquitetura específica para o classificador de sentimentos, a qual mostrou resultados melhores que LSTM, a qual utiliza redes neurais recorrentes e seria a escolha natural para problemas de NLP (Natural Language Processing).

O modelo consiste em:

Convolução 1D => *Max Pooling* => ***FC*** => *ReLu* => ***FC*** => *Sigmoid* => ***Y***

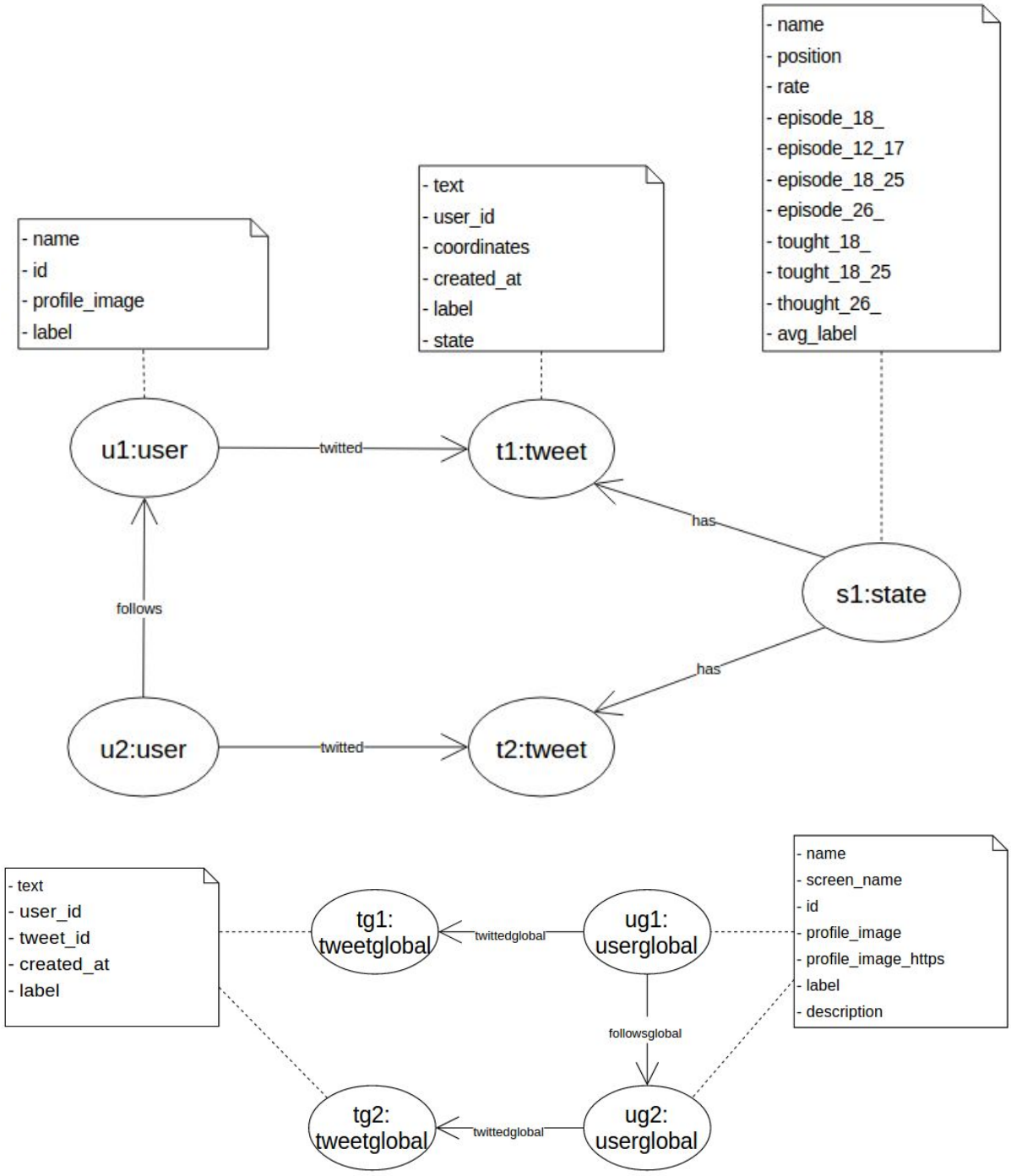
Y seria o output final do modelo, com 0 (ruim) e 1 (bom). A função de perda utilizada foi a Binary Cross Entropy [11], com o otimizador Adam [10].

Entre os hiperparâmetros escolhidos, foi utilizado:

- Apenas as 7000 palavras mais relevantes do banco de dados "Large Movie Review Dataset" [9] (caso encontre uma palavra que não se enquadre nessas 7000 palavras, ela é ignorada, por exemplo);
- O máximo de palavras em cada sentença é 70;
- A dimensão do filtro de convolução é 3;
- O *output* da convolução é 32;
- O *Max Pooling* tem tamanho 2;
- A primeira *FC* mapeia 250 features, enquanto a segunda mapeia para 1 feature;
- O *batch size* utilizado (quantidade de inputs múltiplos para o modelo) foi de 128.

Mapeamentos Interessantes de Modelos

Como foi usado o modelo de grafos, os modelos Entidade Relacionamento precisaram ser mapeados de formas diferentes para o grafo em si. Para isso, foram estudadas as melhores maneiras de integrar os tweets, usuários e estados americanos nos dois grafos, sendo que seus mapeamentos estão abaixo. Nas figuras, label se refere ao valor do sentimento que varia de 0 a 1.



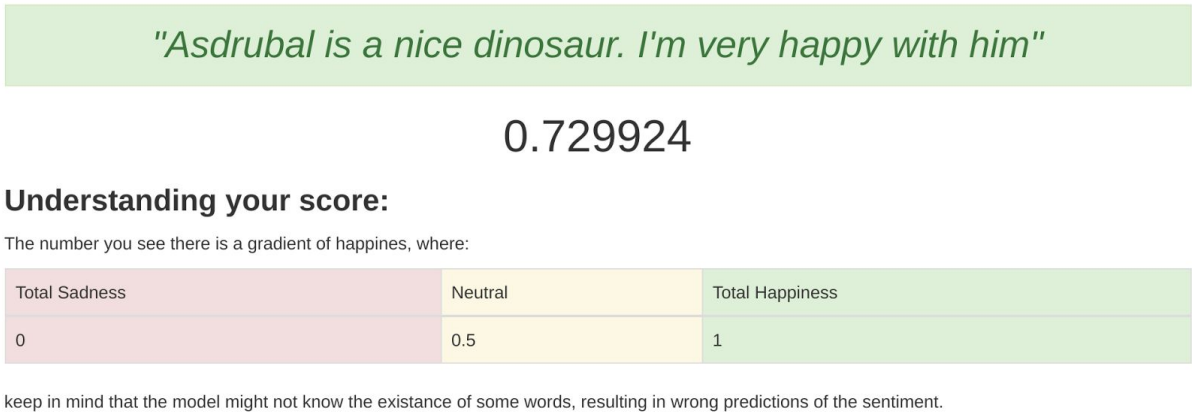
Foram usados dois grafos pois o primeiro grafo não teve usuários e arestas entre estes usuários em número suficientes, haviam 57 mil tweets, 430 usuários e 150 arestas entre os usuários apenas. Por isso, foi feito um segundo modelo com mais tweets, cerca de 1 milhão, e os estados americanos não foram usados, isso para apenas visualizar as relações entre os usuários e seus tweets, além da média de sentimentos calculados por seus tweets. Então o segundo foi mais simples e foi melhor aproveitado nesse caso, com 1200 usuários e 1750 arestas entre eles aproximadamente.

Diversidade da Apresentação

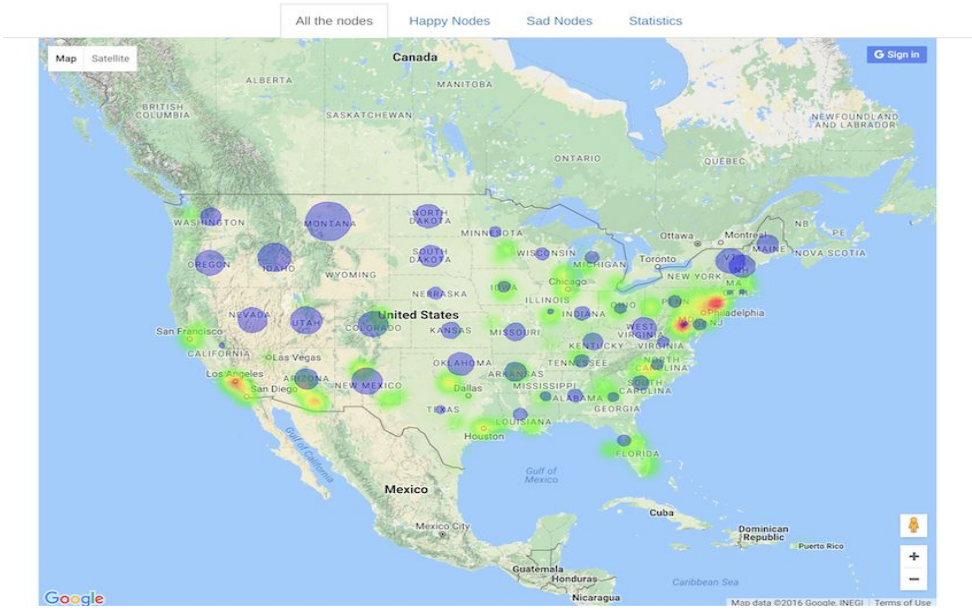
Por fim, destacamos as várias formas de apresentação dos dados como um dos diferenciais do trabalho do nosso cluster, o que, além de tornar o projeto esteticamente mais atrativo, facilita a compreensão do usuário e incentiva a sua interação com os dados analisados. A seguir, apresentaremos imagens de cada um dos tipos de apresentação fornecidos, juntamente com uma breve descrição do que elas representam.

Na imagem abaixo, temos a opção dada ao usuário de analisar o valor de sentimento de uma frase fornecida por ele:

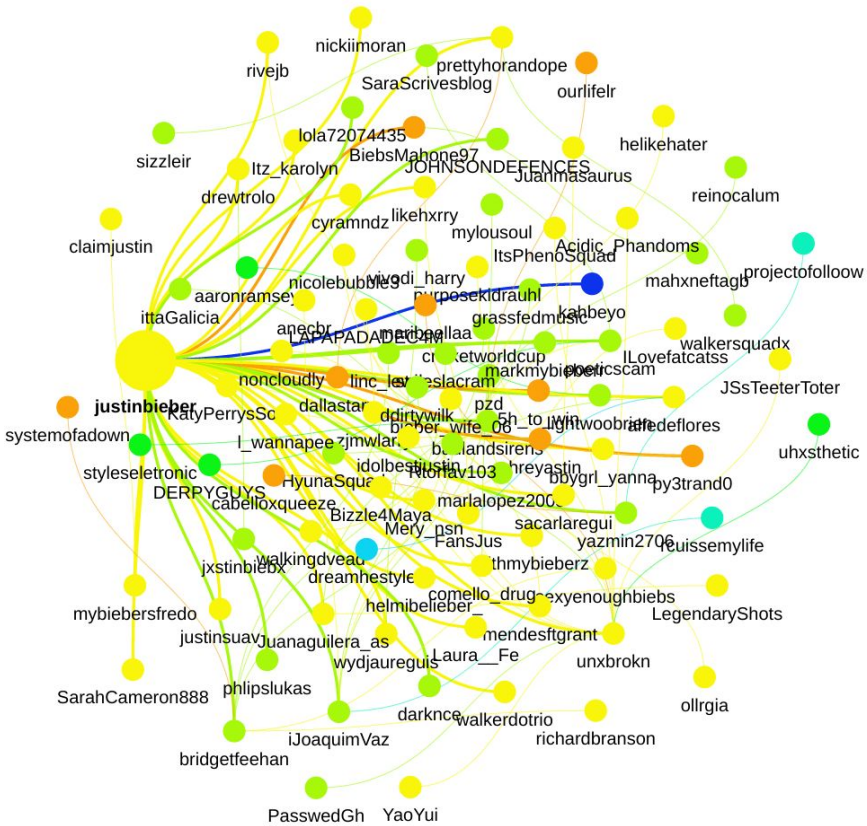
Na próxima imagem, temos um exemplo do resultado dessa análise:



A seguir, temos a apresentação dos dados em forma de um mapa de calor, que também oferece a opção de verificar as estatísticas de cada estado individualmente.



Por fim, temos um exemplo da apresentação em grafos, que também oferece a opção de verificar um outro grafo que representa os sentimentos de cada um dos tweets daquele usuário:



Referências

[1] Statista, “http://www.statista.com/statistics/187478/death-rate-from-suicide-in-the-us-by-gender-since-1950/,”

[2] A. F. for Suicide Prevention, “https://afsp.org/about-suicide/suicide-statistics/,”

[3] World Health Organization, “http://www.who.int/en/,”

[4] Centers for Disease Control and Prevention, “https://www.cdc.gov/,”

[5] Twitter, “https://twitter.com/,”

[6] Kaggle, “https://inclass.kaggle.com/c/si650winter11/data,”

[7] S. Analytics, “http://www.sananalytics.com/lab/twitter-sentiment/,”

[8] Tweepy, “http://tweepy.readthedocs.io/en/v3.5.0/#”

[9] Large Movie Review Dataset, "http://ai.stanford.edu/~amaas/data/sentiment/"

[10] Adam: A Method for Stochastic Optimization, "https://arxiv.org/abs/1412.6980/"

[11] Cross Entropy, "https://en.wikipedia.org/wiki/Cross_entropy/"