s a d

TRABALHO DE BANCO DE DADOS

b o y s

Integrantes

dupla 1

Isadora Sophia	158018
Matheus Diamantino	156740

dupla 2

Arthur Costa Lopes	157699
Gabriel Souza Franco	155477

dupla 3

Lucas Alves Racoci	156331
Luiz Fernando Rodrigues da Fonseca	156475

dupla 4

Henrique Noronha Facioli	157986
Thiago Silva de Farias	148077

dupla 5

Lauro Cruz e Souza	156175
Willian Tadeu Beltrao	157595

Problema proposto

Objetivo

Investigar se há uma relação entre os dados estatísticos como a taxa de casos de suicídio, porcentagem de intenção de suicídio e porcentagem de depressão por estados americanos, e os tweets coletados nessas condições, com uma análise de sentimento realizada usando técnicas de Machine Learning. Também foi proposto saber se um usuário triste ou feliz tende a seguir usuários tristes ou felizes.

A verificação dessas relações, permitiria analisar possíveis sintomas, vocabulários ou comportamento de pessoas que sofrem de depressão e se os dados conferem com a quantidade de suicídios.

Linguagens

A linguagem de implementação escolhida foi Python, com o front-end em JavaScript.

Coleta de Dados Twitter

Script para extração

Por meio de um script em Python, usando a API tweepy conseguimos extrair tweets do Twitter.

A API apresenta algumas limitações como download de pequenos blocos (200 tweets) e possibilidade de baixar apenas os 3000 tweets mais recentes de um usuário.

Seleção de usuários

Para escolher os usuários começamos com alguns perfis de celebridades e contas famosas nos Estados Unidos. Depois disso incluímos num conjunto os seguidores dessa conta e seguidos por ela, continuamos o processo com um usuário qualquer do conjunto, repetindo a operação, e assim por diante.

ARTHUR COSTA LOPES
GABRIEL SOUZA FRANCO

Coleta de Dados Twitter

Quantidade de Tweets

Foram extraidos mais de 11.000.000 de tweets de mais de 9.000 contas diferentes.

Cerca de 60% em inglês e mais de 100 mil com localização.

Filtragem dos tweets

Ligação com dados de suicídio por meio da região (Estados Unidos).

Apenas Tweets em inglês para melhor resultado da classificação.

Poucos usuários habilitam a opção de mostrar localização.

Menos de 1% do dataset total é útil.

ARTHUR COSTA LOPES
GABRIEL SOUZA FRANCO

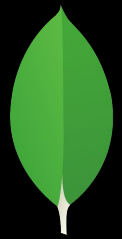
Coleta de Dados Twitter

Ambiente

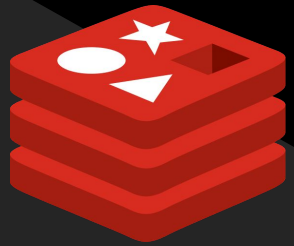
MongoDB para armazenar os objetos JSON que a API do Twitter retorna.

Redis para guardar o conjunto de usuários pendentes.

Máquinas emprestadas do LMCAD.



mongoDB®



redis

ARTHUR COSTA LOPES
GABRIEL SOUZA FRANCO

Coleta de Dados Sobre Depressão e Suicídio

Dados estatísticos

Os dados coletados são de caráter estatístico:

- Taxas de pessoas com depressão clínica
- Taxas de suicídio
- Dados sobre as respectivas pessoas:
  - Gênero
  - Idade
  - Região
- Data da análise

Obtenção dos dados

- Download de bancos de dados já organizados
- Dados obtidos por sistemas de query online
- Dados obtidos manualmente
- Dados obtidos em sua maioria no formato CSV

Limpeza dos dados

- Remoção de redundâncias após a união dos dados
- Verificação de consistência entre as diferentes fontes
- Remoção de atributos desnecessários para a análise das outras duplas
- Padronização de todos os dados para formato csv

LAURO CRUZ E SOUZA
WILLIAN BELTRAO

Coleta de Dados Sobre Depressão e Suicídio

Fotos dos dados

State	Position	Rate
Alabama	24	1446
Alaska	2	2197
Arizona	13	18
Arkansas	16	1725
California	43	1046
Colorado	7	1978
Connecticut	47	971
Delaware	33	132
Florida	28	1384
Georgia	38	1265
Hawaii	29	1364

State	18-	2016/12/17	18-25	26-
Total U.S.	6.62523112%	11.00820161%	9.00479441%	6.21911476%
Northeast	6.66412627%	10.6278766%	9.6072663%	6.17760971%
Midwest	6.81378919%	10.8144019%	9.38702199%	6.37590065%
South	6.47117198%	10.77088478%	8.30557691%	6.15991849%
West	6.66809491%	11.82479874%	9.30003853%	6.20249139%
Alabama	6.84833467%	10.73843459%	8.23833929%	6.61230305%
Alaska	6.56719375%	9.91567723%	9.19212915%	6.06878767%
Arizona	7.31909211%	13.229368%	8.86159114%	7.05135094%
Arkansas	7.31348903%	11.94965352%	9.52062423%	6.93944581%
California	6.29591055%	11.52641869%	8.79721417%	5.83844677%
Colorado	6.28786538%	11.68243072%	8.37681275%	5.9343425%

LAURO CRUZ E SOUZA
WILLIAN BELTRAO

Extração de sites

CDC

Para a extração de dados no site da CDC foi utilizado o WISQARS (Web-based Injury Statistics Query and Reporting System). Um banco de dados online interativo que fornece dados sobre ferimentos fatais dos Estados Unidos e seus estados.

Link: *https://www.cdc.gov/*

World Health Organization

Para a extração de dados no site da WHO foi utilizado o CoDQL, Cause of Death Query Online. Um sistema baseado na web para extrair dados de mortalidade por país a partir do banco de dados da WHO.

Link: *http://www.who.int*

LAURO CRUZ E SOUZA
WILLIAN BELTRAO

Extração de sites

A. F. for Suicide Prevention

Dados sobre suicídio que estavam todos em imagem ou tabelas interativas das quais não conseguimos retirar os dados do html. Todos os dados foram retirados manualmente e passados para uma tabela CSV.

Link: *https://afsp.org/about-suicide/suicide-statistics/*

Statista Statistics Portal

As estatísticas são separadas como artigos oferecidas para download em xlsx. Assim, juntamos todos os artigos que pareciam interessantes para o trabalho (em torno de 15 tanto para depressão quanto para suicídio) e baixamos as tabelas.

Link: *https://www.statista.com/*

LAURO CRUZ E SOUZA
WILLIAN BELTRAO

Banco de Dados

MongoDB para Neo4J

Foi usado MongoDB para armazenar temporariamente os tweets e usuários em formato de documento json.

Então, foi preciso extrair os dados dos documentos dos Tweets e Usuários, limpá-los e enviá-los para o Neo4J.

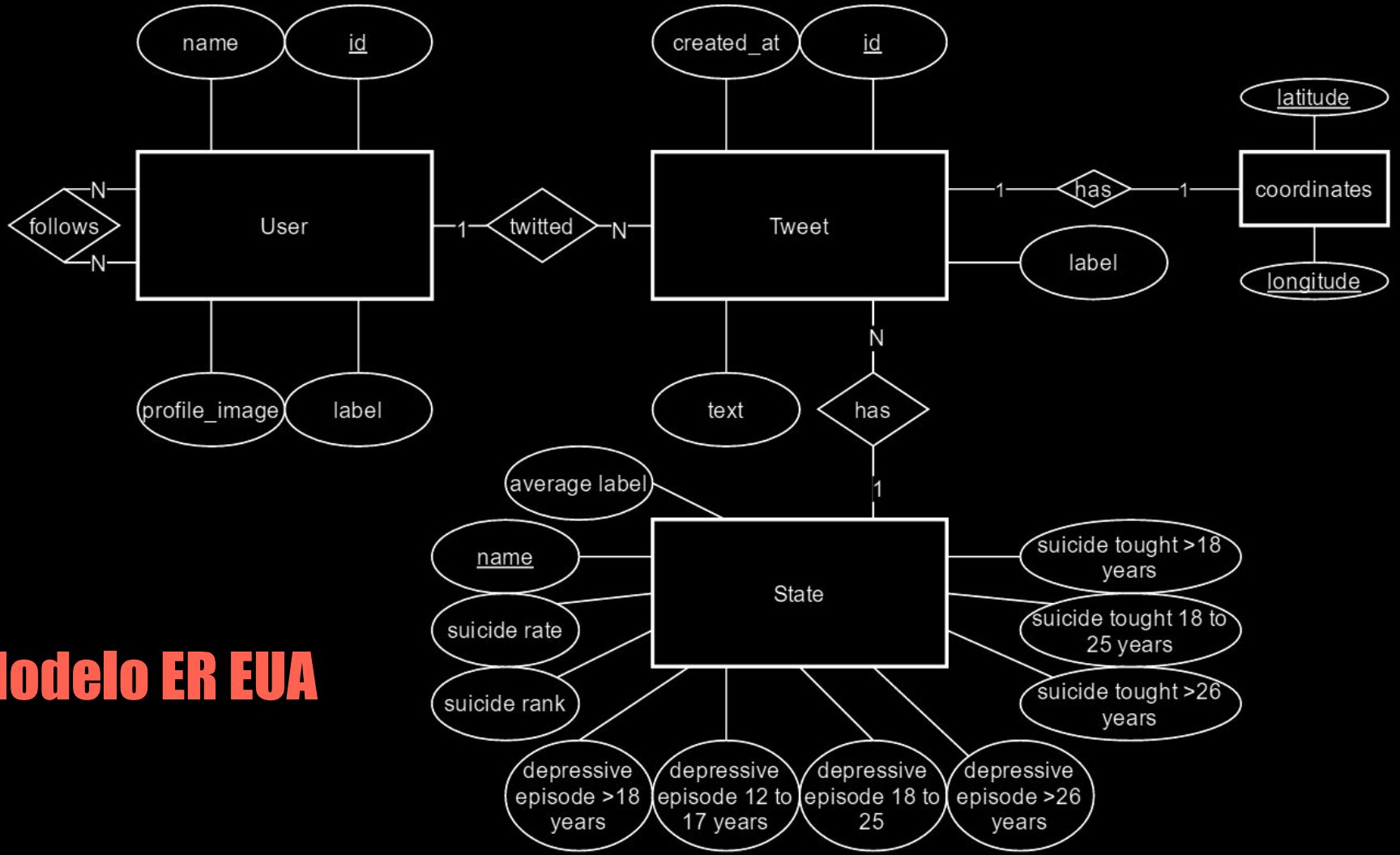
Também foram calculadas as médias de sentimentos dos tweets para cada usuário e a média do sentimento por estado.

Índices no Neo4J

Todas as propriedades de todos os nós foram indexadas para tornar as queries mais rápidas.

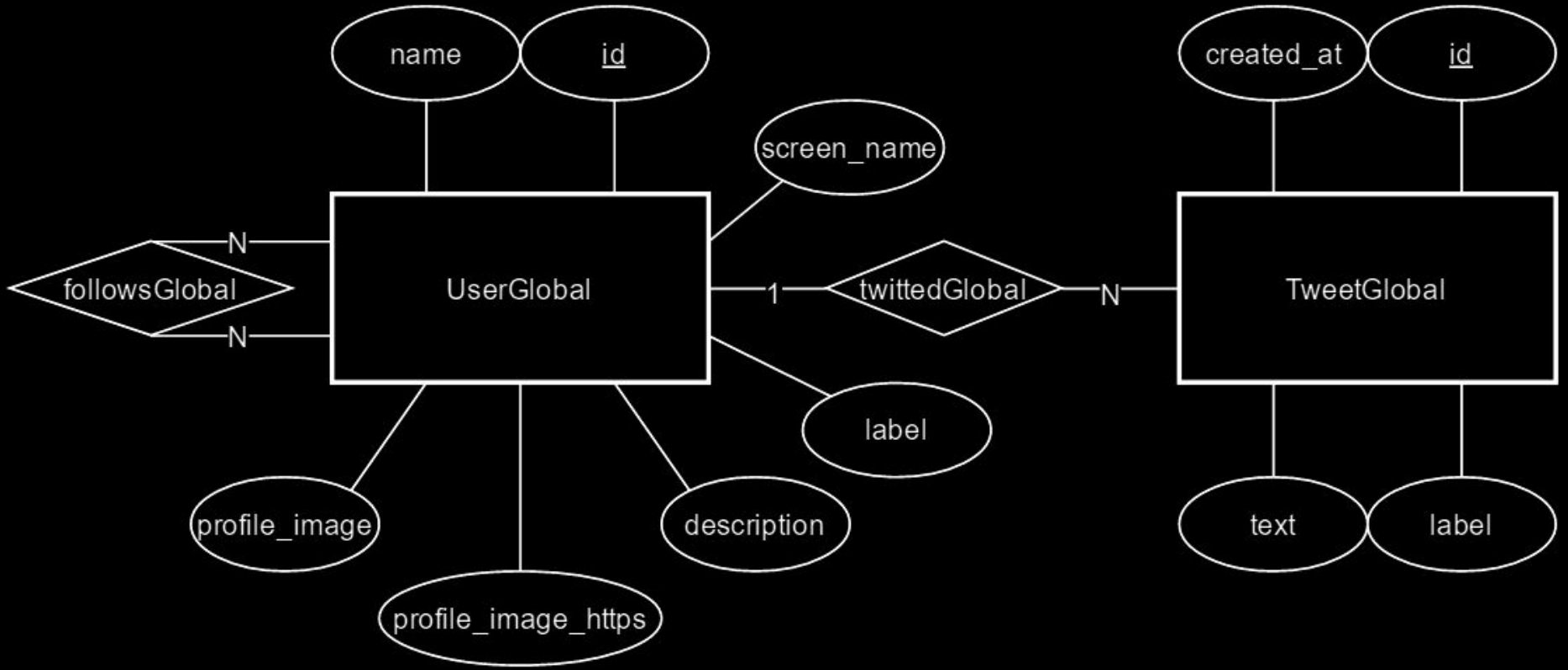
Pensou-se que isso já era feito automaticamente, mas notou-se uma diferença de performance após a criação dos índices.

LUIZ FERNANDO RODRIGUES DA FONSECA
LUCAS ALVES RACOCI

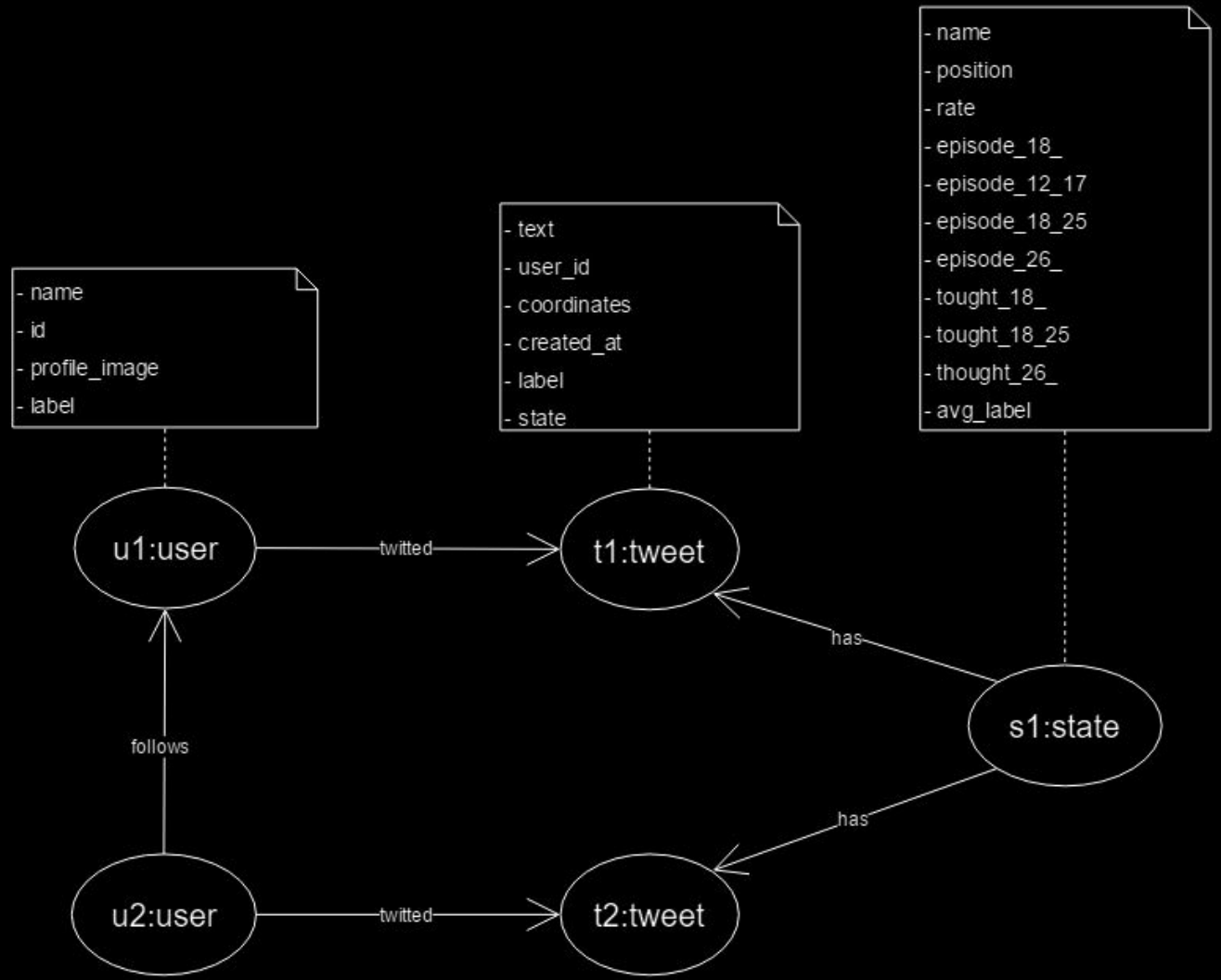


Modelo ER EUA

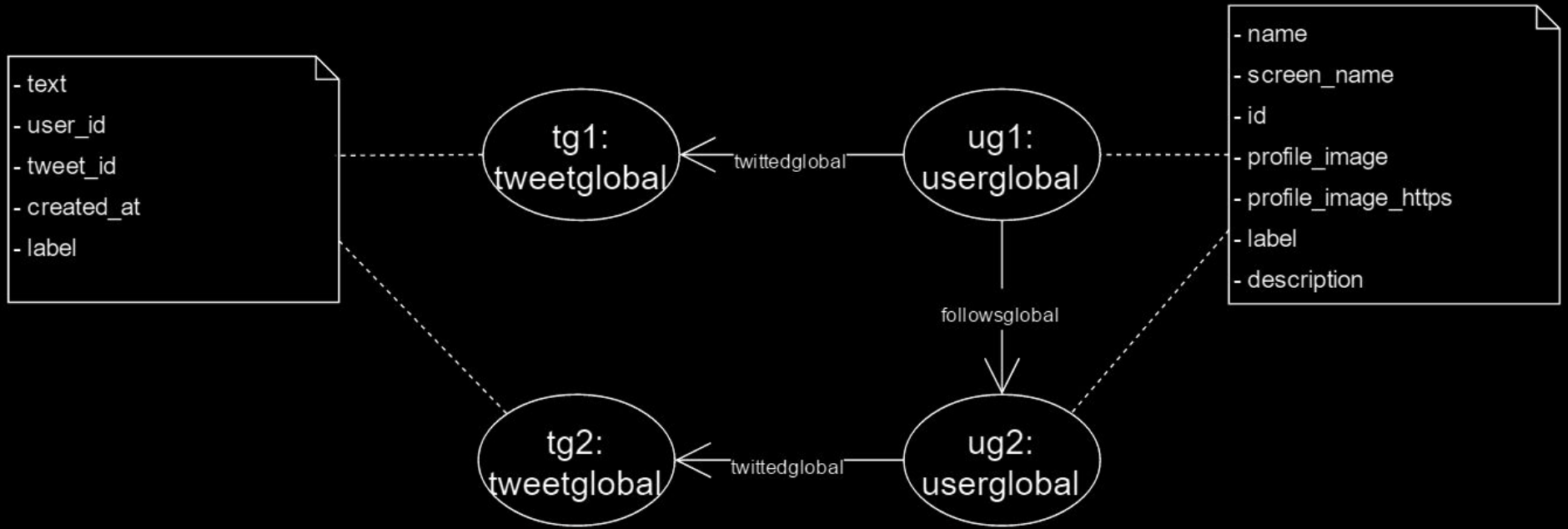
Modelo ER Global



Mapeamento EUA para Grafo



Mapeamento Global para Grafo



Banco de Dados

Geopy Nominatim

Para conseguir extrair os estados dos tweets, foi usado a API do OpenStreetMap para enviar uma coordenada geográfica e ela retornar as informações do local. Assim foram extraídos os tweets em estados americanos.

```
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)
```

LUIZ FERNANDO RODRIGUES DA FONSECA
LUCAS ALVES RACOCI

Banco de Dados

API para a Apresentação Visual

Foram criadas APIs para facilitar o acesso da Dupla 4 aos dados armazenados no grafo que seriam necessários na apresentação visual.

A princípio todas as duplas poderiam ter lido diretamente do grafo o que precisavam, mas da forma como foi feito, as equipes puderam se concentrar mais em seus objetivos.

API para a Análise de Sentimento

Foram criadas APIs para facilitar a leitura e atualização de dados referentes a classificação de sentimento das frases de cada tweet, que foi usado pela Dupla 1 na Análise de Sentimento.

LUIZ FERNANDO RODRIGUES DA FONSECA
LUCAS ALVES RACOCI

Machine Learning

Objetivos

Nosso objetivo foi classificar sentimentos de tweets a partir do treinamento de um modelo em Deep Learning utilizando o dataset de Movie Reviews do IMDB para treinamento.

Contudo, para atingi-lo, foi necessário um estudo teórico de diferentes modelos e recorrência à literatura, de modo a obter o melhor resultado possível.

ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

Desafios

O primeiro desafio encontrado foi o de busca e estudo de um modelo eficiente para implementação. Após várias pesquisas pela literatura, chegamos em dois possíveis modelos: **LSTM**, um modelo de recurrent neural network (RNN) e um novo modelo com convolutional neural net (CNN) em 1D.

O próximo desafio foi a implementação e busca de frameworks e libraries que facilitariam este trabalho.

ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

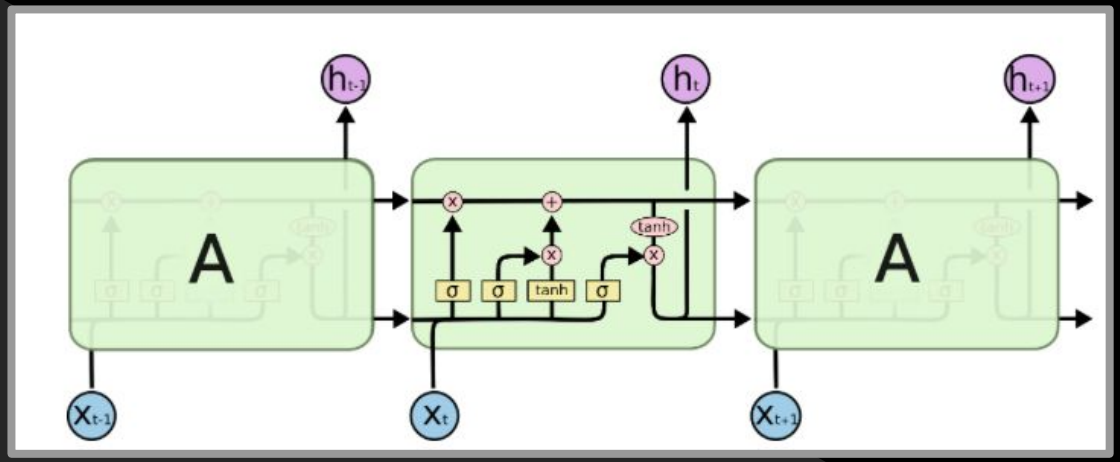
Estudo Teórico

RNN com LSTM

Uma primeira tentativa envolveu uma RNN com LSTM, que consiste em uma estrutura baseada em redes neurais estáticas em conjunto com uma estrutura de memória para contextualização das entradas.

Como nossas entradas são palavras em um texto, é desejável que nosso modelo contextualise as palavras baseado nas que vem antes dela.

Ao lado, uma imagem ilustrativa de um modelo RNN utilizando LSTM.



ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

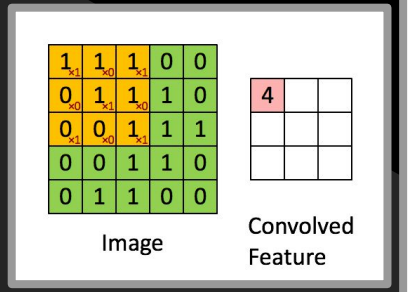
Estudo Teórico

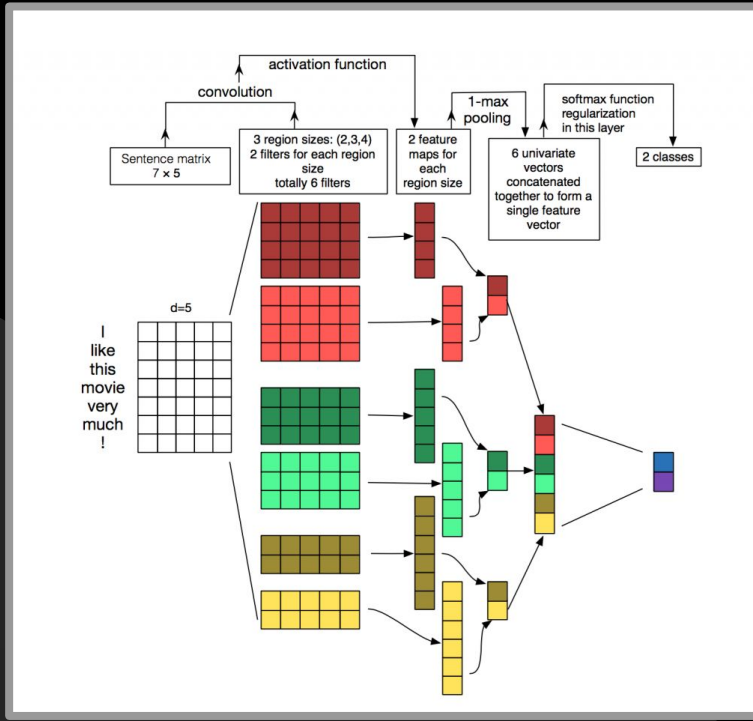
Rede Neural Convolucional

Nossa segunda tentativa envolveu o modelo de **Rede Neural Convolucional 1D**, que consiste em separar a entrada (vetor de palavras) em pequenas amostras (filtros) que vão ser avaliadas para extrair características determinadas em conjunto com o método de **Max Pooling**.

Após esta fase, o resultado deste modelo é passado por duas camadas de redes neurais fully connected para interpretação do resultado e predição do sentimento do texto dado. A estrutura final do modelo foi:

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





ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

Implementação

Keras

Para a implementação do modelo, foi utilizado o Keras, uma library de redes neurais em alto nível que facilita a definição e treinamento de modelos.

Com ela, foi possível uma rápida iteração entre os diferentes modelos testados, facilitando chegar em um resultado ótimo.

Tensorflow

O Tensorflow se trata de uma API em python que otimiza diversas estruturas de deep learning, permitindo, por exemplo, a paralelização em CPU e GPU.

É um software **open source** do google! Facilita bastante o desenvolvimento de deep learning.

ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

Resultados

```
["Today I typed my bank account password in the microwave keypad, and waited. My
wife told me I was trying to heat up the economy."]
Happy! :)
```

```
["Overwatch: the game I play to relax, but end up getting even more stressed than
 I was before I started"]
Sad.. ಠ_ಠ
```

```
["I have luck to be currently working with some of the most amazing individuals I
 met in a long time."]
Happy! :)
```

```
["Happy Thanksgiving to everyone. We will, together, MAKE AMERICA GREAT AGAIN!"]
Happy! :)
```

ISADORA SOPHIA
MATHEUS DIAMANTINO

Machine Learning

Fontes e referências

[1]: “Sentiment Analysis with Deeply Learned DistributedRepresentations of Variable Length Texts” , https://cs224d.stanford.edu/reports/HongJames.pdf

[2]: “LSTM Networks for Sentiment Analysis”, http://deeplearning.net/tutorial/lstm.html

[3]: "Learning LSTM", http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4]: "Understanting Convolutional Neural Nets, http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

ISADORA SOPHIA
MATHEUS DIAMANTINO

Apresentação de Dados

Flask

"Flask is a micro web framework written in Python"

Será usada para passar os dados para as bibliotecas de frontend em javascript

HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados

vis.js

“A dynamic, browser based visualization library.”

Ótima para visualização de grafos

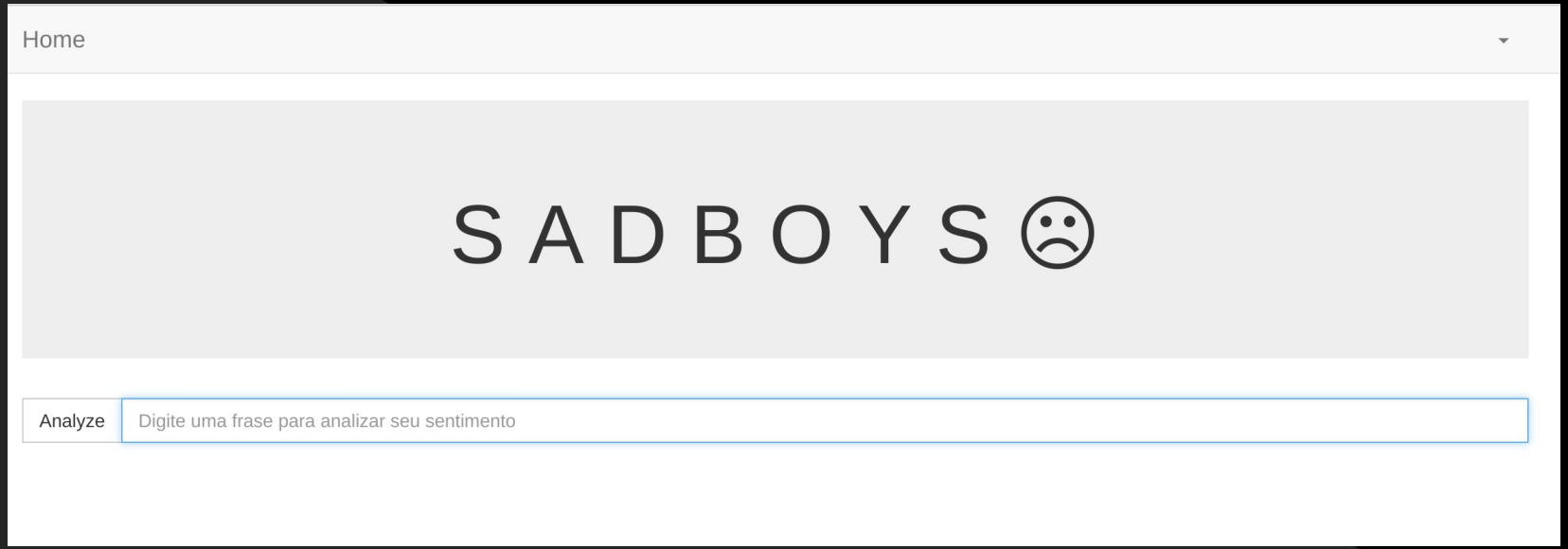
Google Maps API

“Customize maps with your own content and imagery.”

Uma API para apresentação de mapa de calor

HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados



HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados

"Asdrubal is a nice dinosaur. I'm very happy with him"

0.729924

Understanding your score:

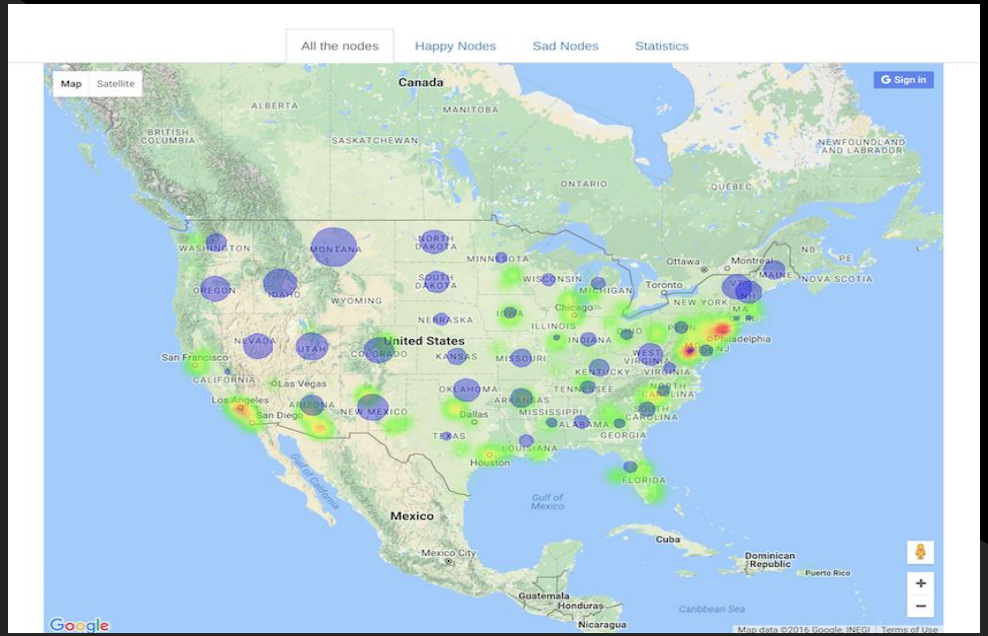
The number you see there is a gradient of happines, where:

Total Sadness	Neutral	Total Happiness
0	0.5	1

keep in mind that the model might not know the existance of some words, resulting in wrong predictions of the sentiment.

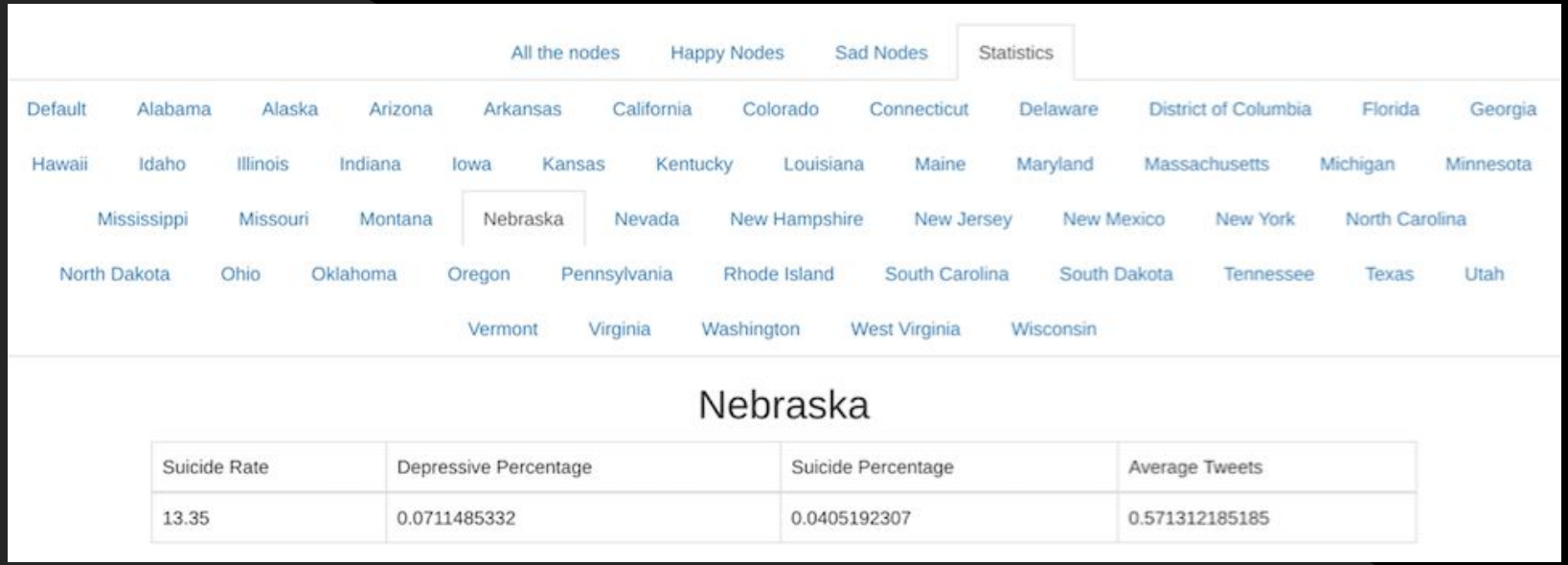
HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados



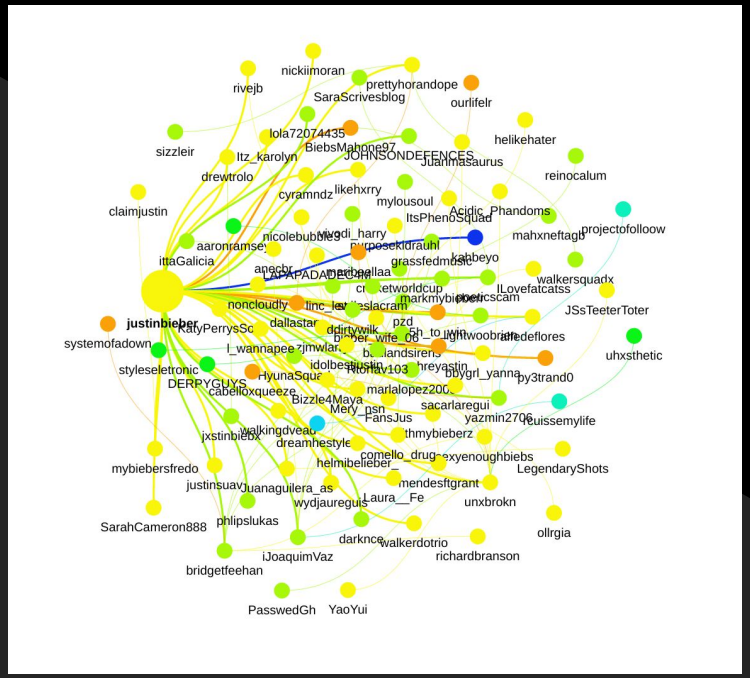
HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados



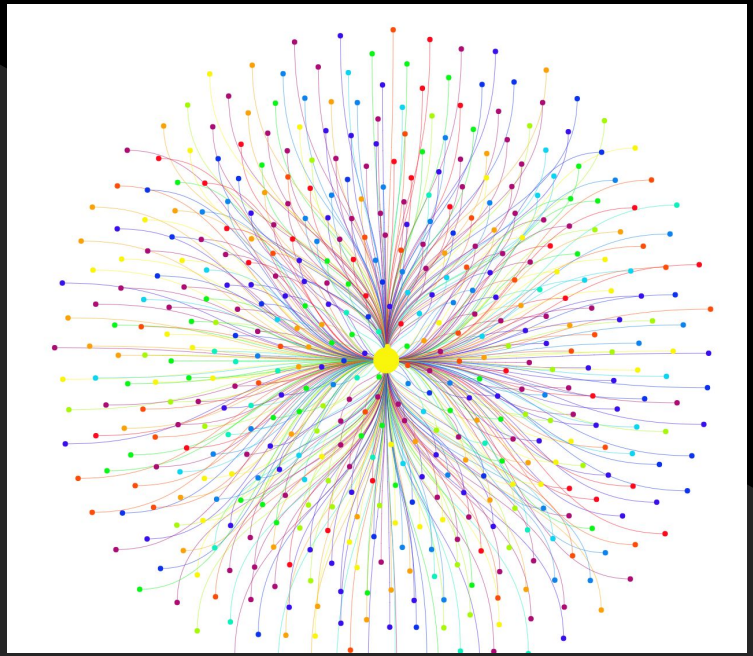
HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados



HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

Apresentação de Dados



HENRIQUE NORONHA FACIOLI
THIAGO SILVA DE FARIAS

DATA IS

THE KEY