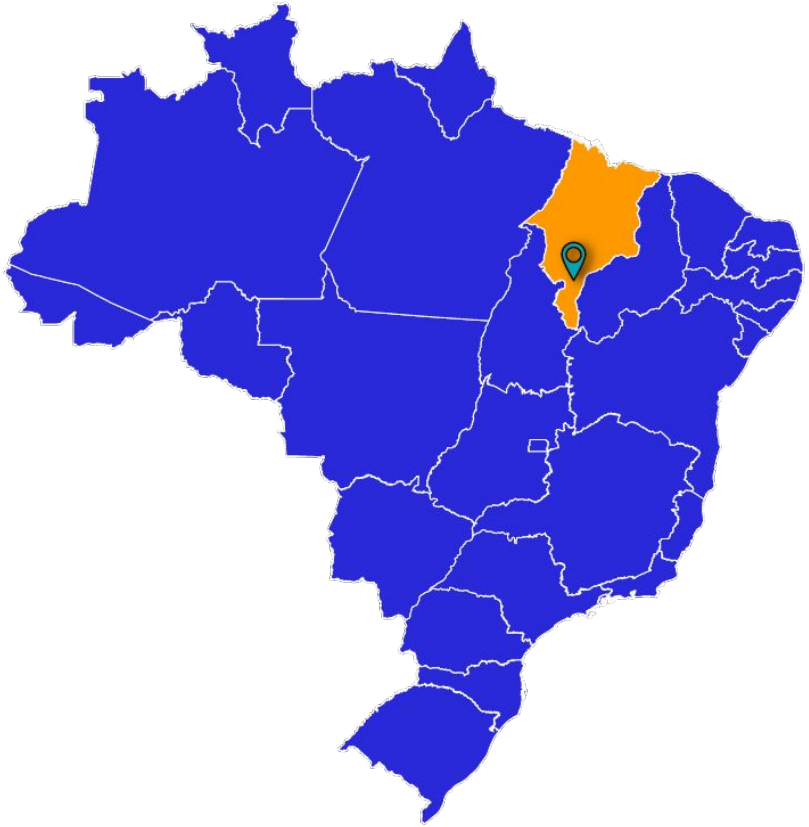
CAIPYRA

Indicadores Inteligentes para Detecção de Epidemias de Dengue através do monitoramento de Redes Sociais em Tempo Real

Palestrante: Jadson José Monteiro Oliveira
<jadsonjjmo@gmail.com>

Quem sou eu?

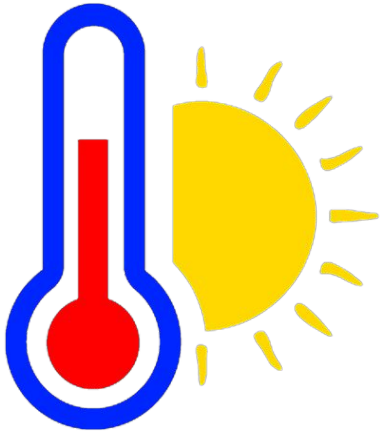
- **Mestrando em Ciências de Computação**
  - ICMC/USP
    - Mineração de dados em **Big Data** utilizando processamento massivamente paralelo.
- **Bacharel em Sistemas de Informação**
  - Faculdade de Balsas



Balsas – Maranhão

Temperatura média anual: ~ 27° C

** ~ 3 a 4 meses sem chuva!



~75%

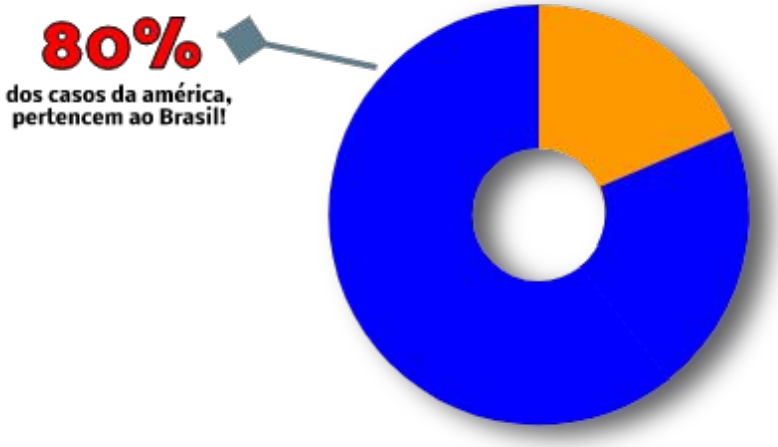
Pessoas como sensores!

- Seria possível criar indicadores a partir desses dados?
- O quão útil isso seria?
- E se fosse possível auxiliar a tomada de decisão para o setor da saúde do nosso país?

Qual a importância de monitorar a dengue no Brasil?

Qual a importância de monitorar a dengue no Brasil?

- A dengue é uma das maiores preocupações na saúde brasileira.



- Em **2016**, foram registrados **1.5 milhões** de diagnósticos da doença.
  - **700 pessoas** vieram a óbito!

Qual a importância de monitorar a dengue no Brasil?

- Pesquisas sobre surtos e epidemias são **importantes**!
  - Implementação de políticas mais eficazes.

- Tempo de acesso aos dados de diagnósticos é **ALTO**.
  - O que acontece hoje, será reconhecido pelo Ministério da Saúde somente **depois de alguns dias** ou até **semanas**.

Existem vários pontos ao nosso favor!

Pontos ao nosso favor!

01	**Grande quantidade de dados!**	• Dados gerados de aplicações comerciais e científicas. • Dados provindos de ínumeras fontes.
02	**Fácil aquisição de dados das redes sociais, como: Twitter.**	• https://developer.twitter.com • https://developers.facebook.com
03	**Tendência dos usuários é favorável!**	• Os usuários tendem a postar seu estado físico e emocional nas redes sociais.

Pronto!

#SQN

Necessidade de tratamento dos dados!

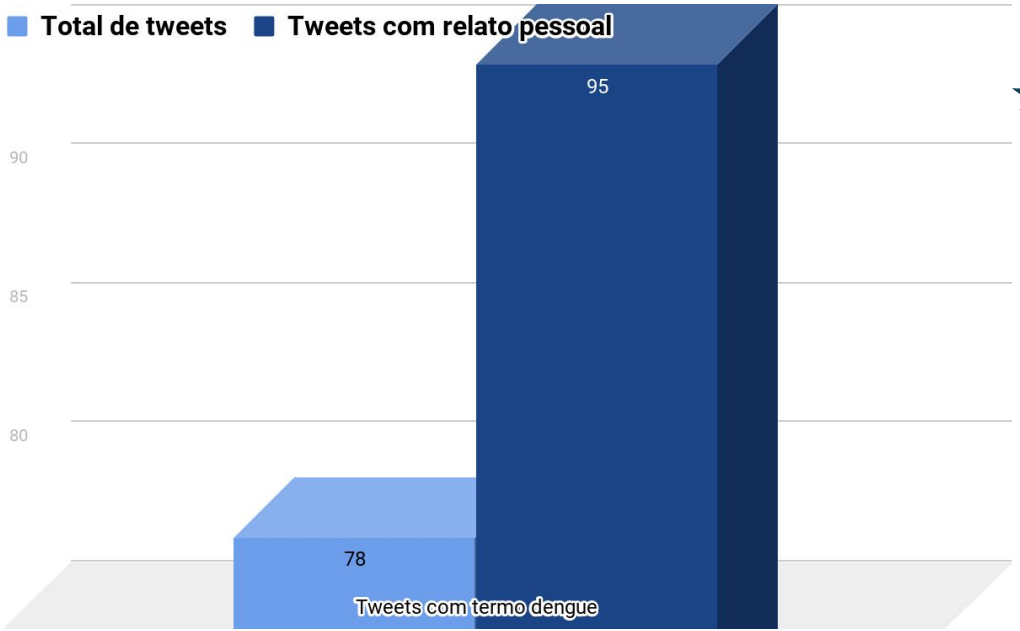
- Grande parte dos dados extraídos, **não estão estruturados**!
  - É necessário realizar um pré-processamento nos dados.

O que nós fizemos para criar esses indicadores inteligentes?

1

Validação da hipótese

1 Validação da hipótese

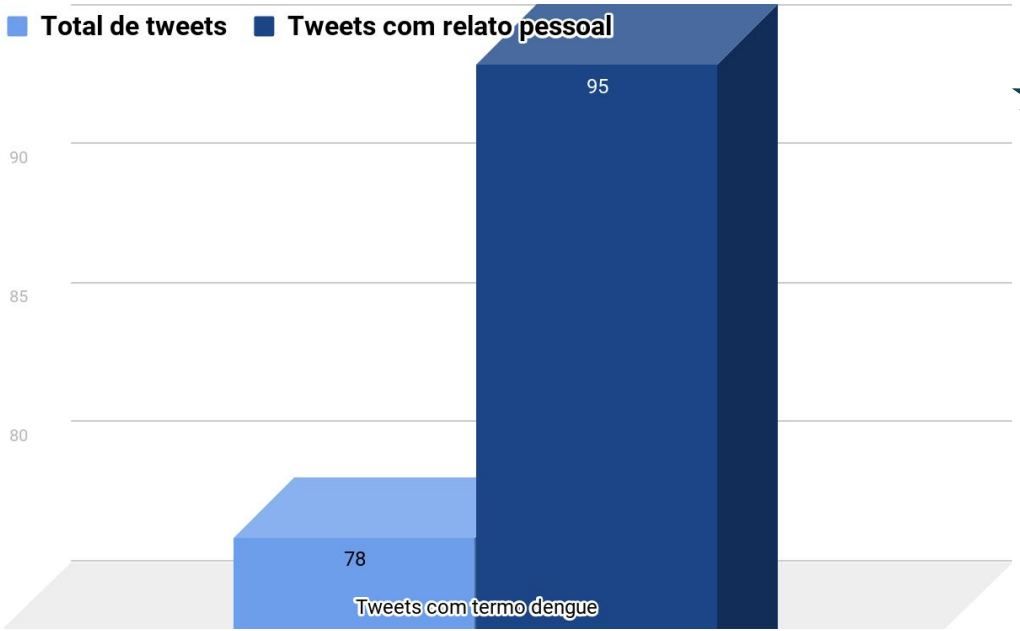


★ Gomide et al. (2011) – UFMG

- Tweets de 2009 e 2011.
- ~ **78%** de correlação com total de tweets que tinha o termo dengue.
- ~ **95%** de correlação com o total de tweets que relataram uma experiência pessoal.


Em resposta a
E eu q estou com **dengue**
1
1
1

1 Validação da hipótese



★ Gomide et al. (2011) – UFMG

- Tweets de 2009 e 2011.
- ~ **78%** de correlação com total de tweets que tinha o termo dengue.
- ~ **95%** de correlação com o total de tweets que relataram uma experiência pessoal.

Como fazer isso de forma rápida?

(visando agilidade no desenvolvimento)

Python

- ★ Linguagem simples
- ★ Sintaxe intuitiva
- ★ Documentação farta
- ★ Inúmeras bibliotecas abertas
- ★ Multiplataforma
- ★ ...

https://www.python.org

Como fazer isso de forma escalável?

(visando agilidade no processamento)

- ★ Ferramenta para Big Data
- ★ Processa os dados de forma paralela e distribuída
- ★ Conjunto de dados resilientes e distribuídos (RDD)
- ★ Processamento em memória principal
- ★ MLlib – biblioteca de Machine Learning
- ★ ...

https://spark.apache.org

- ★ Spark Streaming
  - Facilidade de desenvolvimento de aplicações que suportam fluxo de dados.
  - Tolerante a falhas
  - Escalável
  - ...

https://spark.apache.org

2 Coleta de dados para análise

2 Coleta de dados para análise

- Crawler (feito em python)
  - Captura de tweets dos anos de 2014 até 2017.

3 Rotular o *tweets* capturados!

3 Rotular o *tweets* capturados

meu wpp vai dar **dengue** de tão parado que tá

1 1 1

minha vida tá igual água de **dengue** - PARADA

1 3 6

Tô achando q meu whats vai dar **dengue**

1 2

3 Rotular o *tweets* capturados

Veeii impossivel que tô com **dengue** Mano, pqp n vou ficar sem passar mal 1 dia n jesus?

3 2

Achei que tava boa, aí amanheci vomitando. **Dengue** é um inferno, meus amigos

2 1

Governo emite alerta para 86 cidades do Maranhão com risco de surto de **dengue**, zika e chikungunya dlvr.it/QWYSNP

3 Rotular o *tweets* capturados

- Rotulação em 700 Tweets.
  - Dengue
    - Informação
    - Combate
    - Relato pessoal
  - Não dengue

Como fazer um sistema que entenda os *tweets* sobre dengue?

4

Estruturação dos dados

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras

- Prefeitura de São Carlos desenvolve ações educativas para o combate da dengue.
- Vamos combater a dengue!
- Eu estou com dengue.

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras
    a. Stemming

- Prefeitura de São Carlos desenvolve ações educativas para o combate da dengue.
- Vamos combater a dengue!
- Eu estou com dengue.

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras

- Prefeitura de São Carlos desenvolve ações educativas combate dengue.
- Vamos combater dengue!
- Eu estou com dengue.

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras

- Prefeitura de São Carlos desenvolve ações educativas combate dengue.
- Vamos combater dengue!
- Eu estou com dengue.

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras
    a. Stemming

- prefeitura de sao carlos desenvolve acoes educativas combate dengue
- vamos combater dengue
- eu estou com dengue

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras
   a. Stemming

- prefeit de sao carl desenvolv aco educ combat deng
- vam combat deng
- eu est com deng

4 Estruturação dos dados

1. Remoção de *stop-words*
2. Padronização de caracteres
3. Normalização das palavras
   a. Stemming

Natural language toolkit - nltk
https://www.nltk.org

- prefeit de sao carl desenvolv aco educ combat deng
- vam combat deng
- eu est com deng

4 Estruturação dos dados

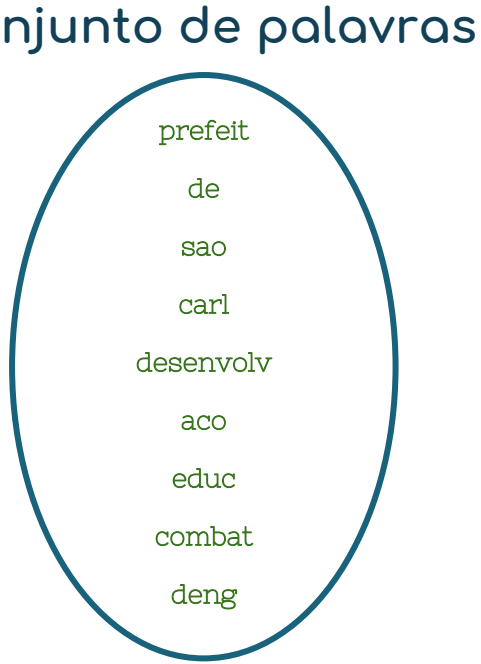
4. Geração de *n-grams*.

- prefeit de sao carl desenvolv aco educ combat deng
  - prefeit_de
  - de_sao
  - sao_carl
  - carl_desenvol
  - desenvolv_aco
  - ...

4 Estruturação dos dados

5. Bag-of-words

Representação utilizada para estruturar as palavras, geralmente em um formato de tabela.



4 Estruturação dos dados

5. TF ou TF-IDF

Utilização de uma medida estatística para representar a importância de uma determinada palavra.

- ★ Term-Frequency
  - Frequência do termo em cada documento
- ★ Inverse Document Frequency
  - Frequência inversa do termo nos documentos

4 Estruturação dos dados

5. TF ou TF-IDF

Utilização de uma medida estatística para representar a importância de uma determinada palavra.

- ★ Term-Frequency
  - Frequência do termo em cada documento
- ★ Inverse Document Frequency
  - Frequência inversa do termo nos documentos

4 Estruturação dos dados

5. TF ou TF-IDF

Utilização de uma medida estatística para representar a importância de uma determinada palavra.

- ★ Term-Frequency
  - Frequência do termo em cada documento
- ★ Inverse Document Frequency
  - Frequência inversa do termo nos documentos

5

Treinamento de um classificador

5 Treinamento de um classificador

- ★ Árvores de decisão
- ★ Naive Bayes Multinomial

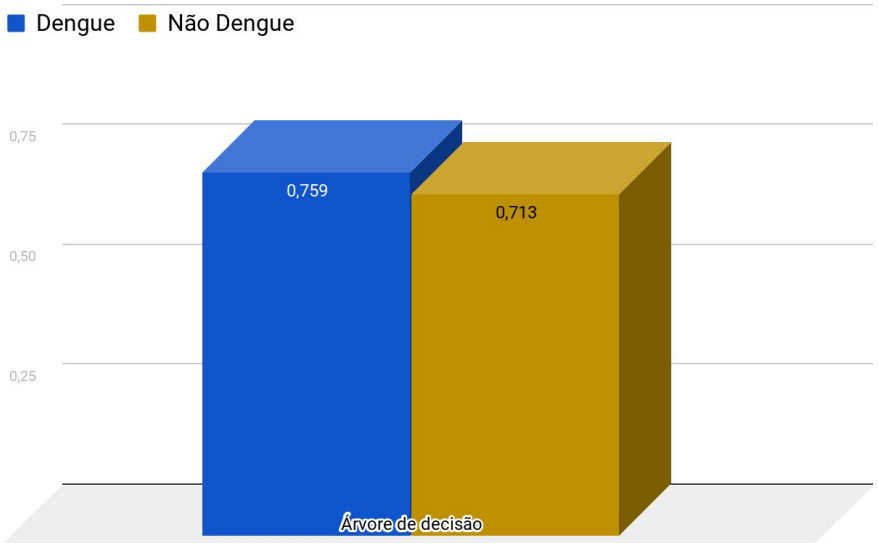
5 Treinamento de um classificador

- ★ Árvores de decisão
- ★ Naive Bayes Multinomial

- F-Measure
  - Medida objetiva que utiliza a precisão e o revocação para gerar um valor entre 0 e 1.
- Precisão
  - Quantos elementos selecionados são relevantes?
- Revocação
  - Quantos elementos relevantes foram selecionados?

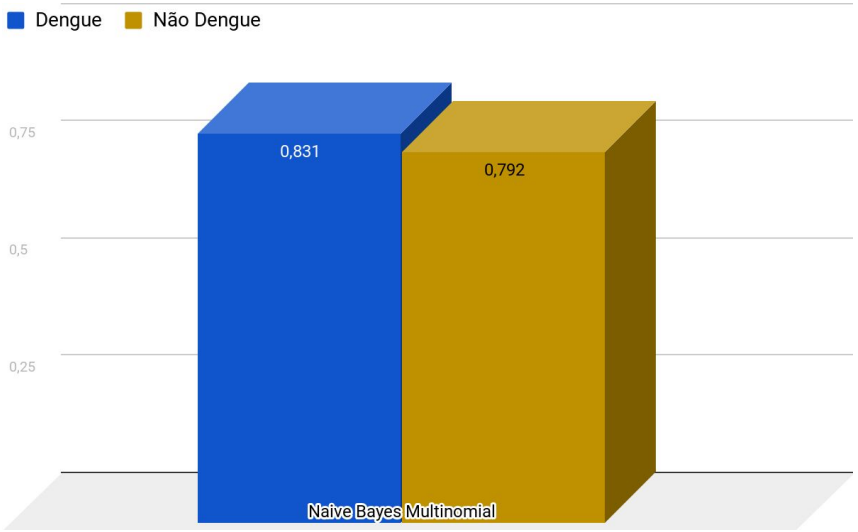
5 Treinamento de um classificador

- Árvores de decisão
- Naive Bayes Multinomial



5 Treinamento de um classificador

- Árvores de decisão
- Naive Bayes Multinomial



6

Persistir os dados em um banco de baixa latência

6 Persistir os dados em um banco de baixa latência

elasticsearch

- ★ ElasticSearch
  - NoSQL
  - Motor de busca distribuído
  - Baixa latência

https://www.elastic.co/products/elasticsearch

Próximo passo...

Como visualizar os tweets processados em tempo real?

6 Persistir os dados em um banco de baixa latência

- ★ Kibana
  - Plugin de visualização do ElasticSearch
  - OpenSource
  - Facilidade de uso

https://www.elastic.co/products/kibana

6 Persistir os dados em um banco de baixa latência

kibana

https://www.elastic.co/products/kibana

Obrigado!

Github: https://github.com/jadsonjjmo
Email: jadsonjjmo@gmail.com
Linkedin: https://www.linkedin.com/in/jadsonjjmo