

MBA em Inteligência Artificial

Big Data e Visualização de Dados

Prof. Diego Nogare



Faculdade de Computação e Informática
Mackenzie

Diego Nogare



Former MVP Artificial Intelligence
Microsoft Regional Director
Diretor no PASS.org
Chief Data Officer @ Lambda3
Membro notável na I2AI
Mestre em IA

www.diegonogare.net
www.livrosdonogare.com.br





PROCESSAMENTO DE LINGUAGEM NATURAL

Como analisar texto e tomar decisões baseadas nisso

Introdução ao Big Data

Big Data deve aceitar a confusão dos dados

Visão Tradicional

Alta qualidade dos dados
Reduzir erros
Dados Precisos



Visão Big Data

Afrouxar a qualidade dos dados
para coletar mais dados



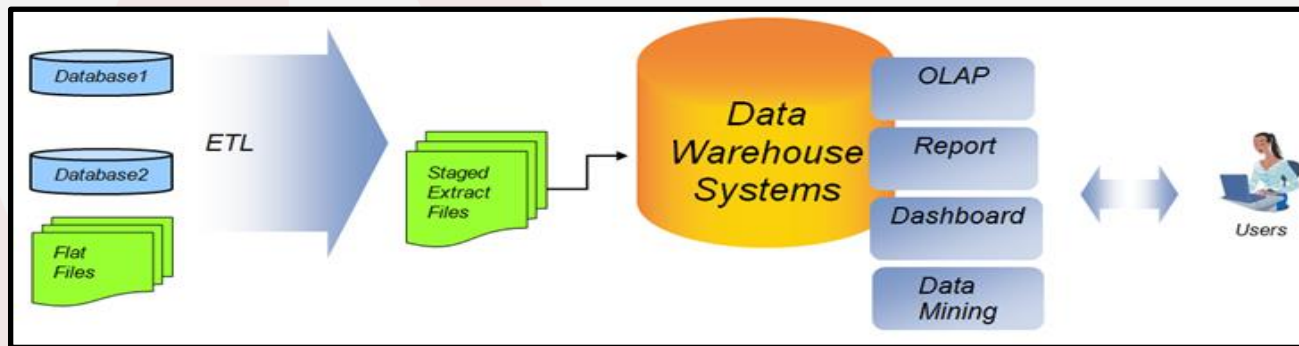
Introdução ao Big Data

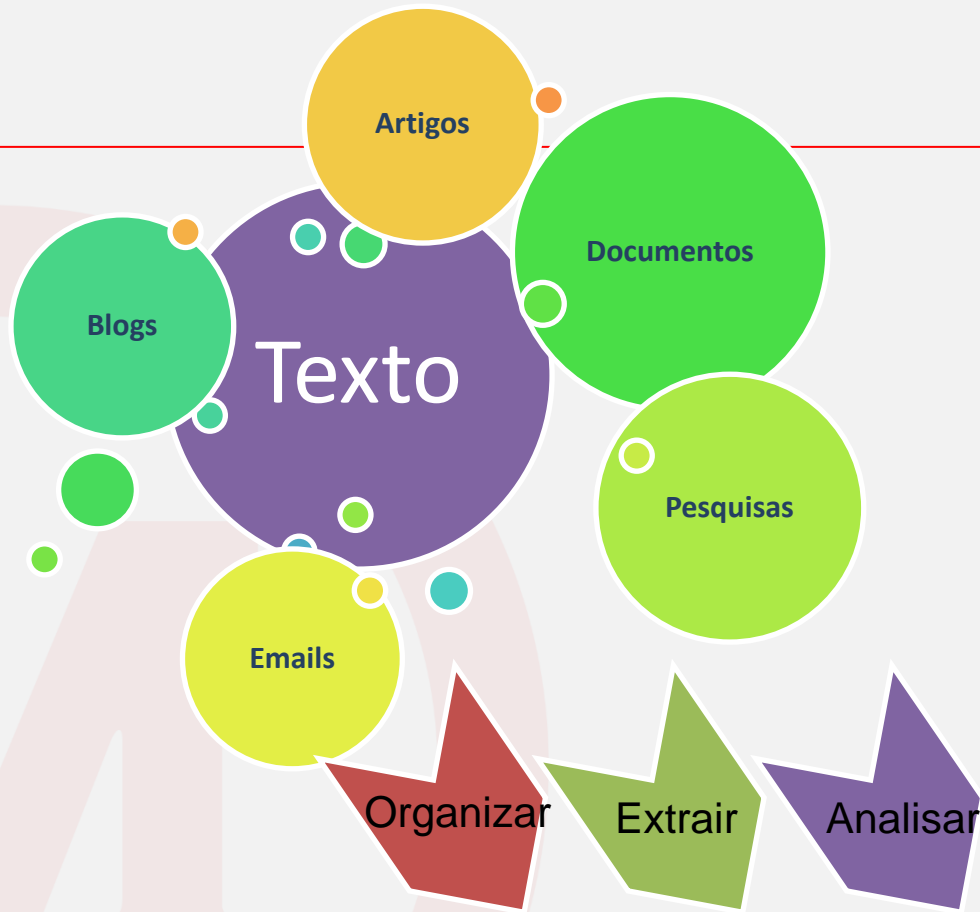
A arquitetura tradicional prioriza a precisão

Carga e
Transformação

Armazenamento
e Organização

Analise

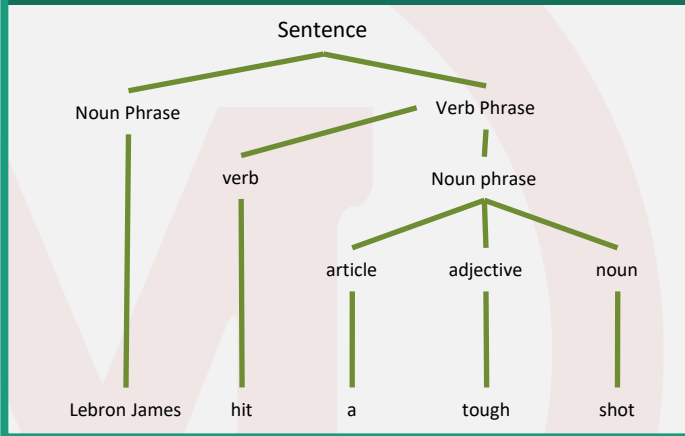




■ Abordagens para Text Mining

“Lebron James hit a tough shot.”

Semantic Using Syntactic Parsing



Bag of Words



■ Abordagens para Text Mining

Alguns desafios do Text Mining

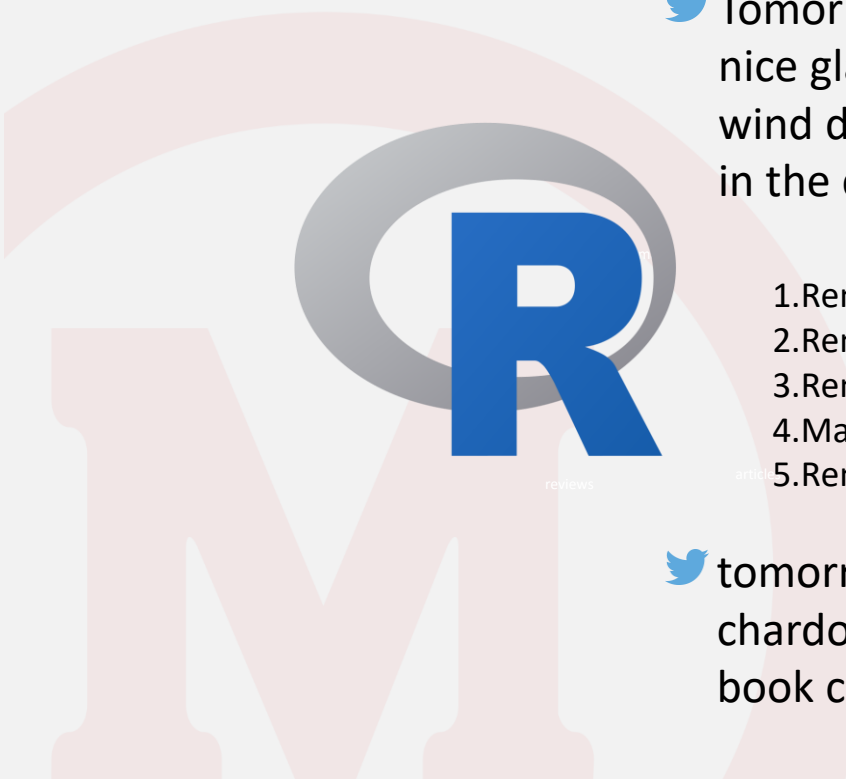
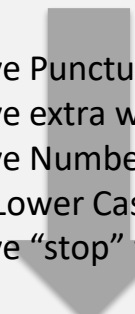
- Compound words (tokenization) changes meaning
 - “not bad” versus “bad”
- Disambiguation
- Sarcasm
 - “I like it...NOT!”
- Cultural differences
 - “It’s wicked good” (in Boston)

“I made her duck.”

- I cooked waterfowl to eat.
- I cooked waterfowl belonging to her.
- I created the (clay?) duck and gave it to her.
- Duck!!

■ Passos para limpeza

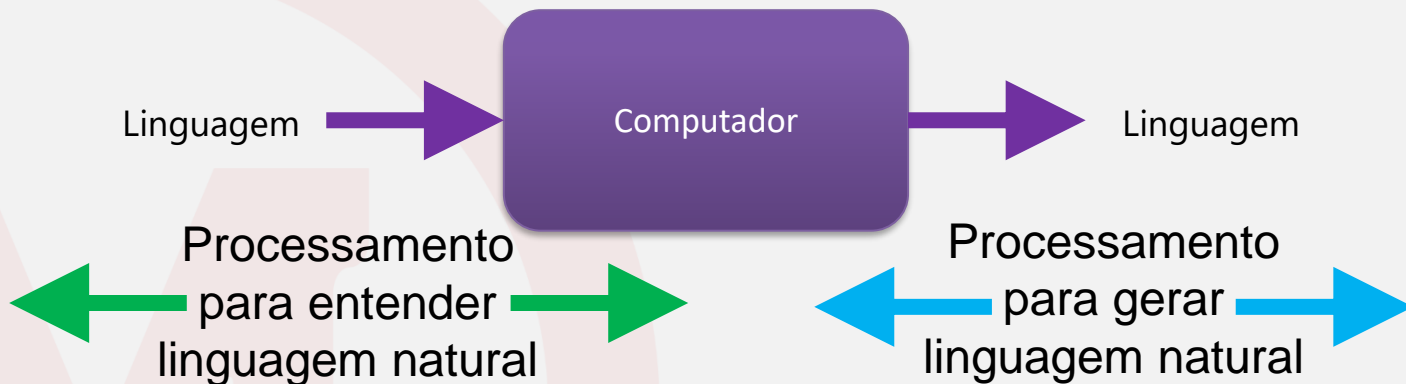
🐦 Tomorrow I'm going to have a nice glass of Chardonnay and wind down with a good book in the corner of the county :-)

- 
- 
- 1.Remove Punctuation
 - 2.Remove extra white space
 - 3.Remove Numbers
 - 4.Make Lower Case
 - 5.Remove “stop” words

🐦 tomorrow going nice glass chardonnay wind down good book corner county

■ Overview - NLP

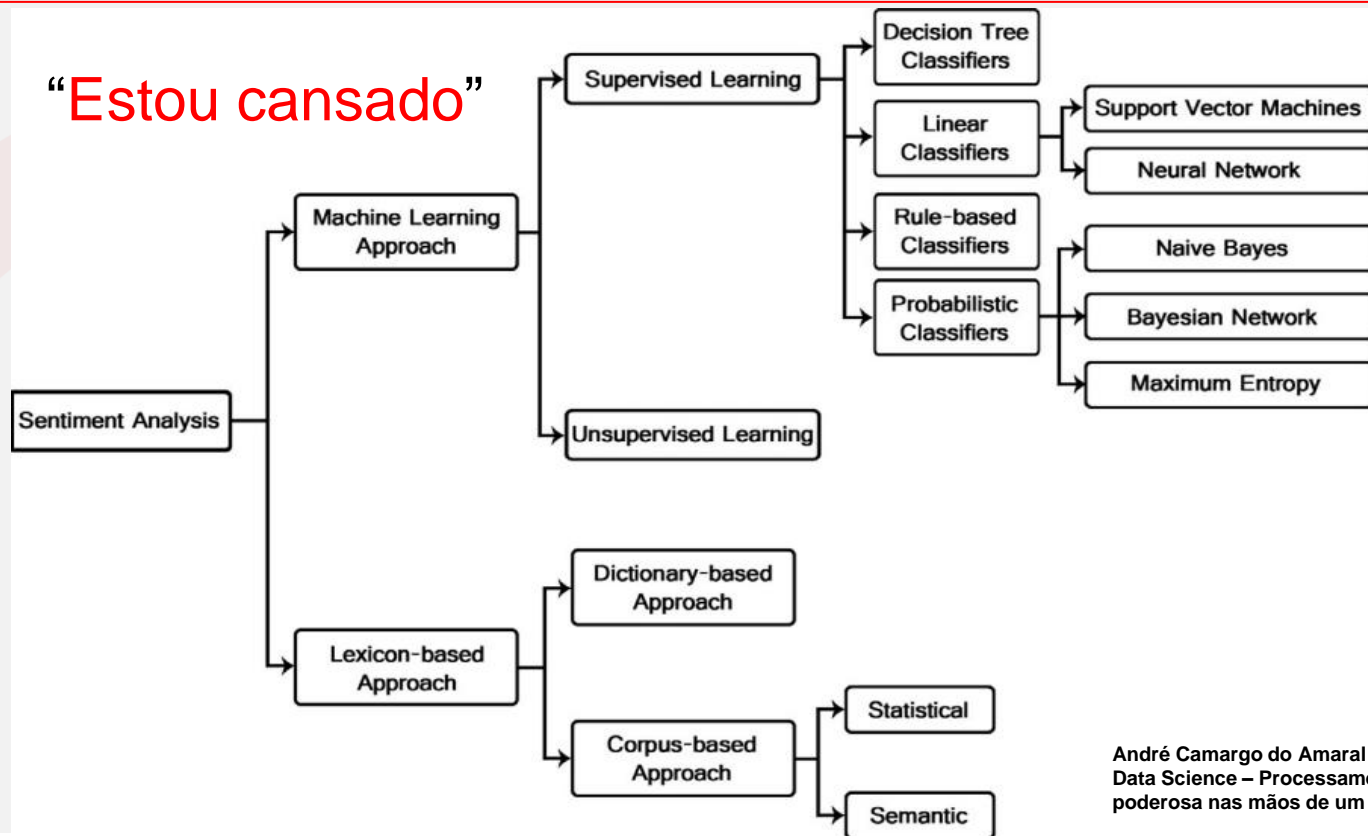
Computadores usam Linguagem Natural como input e/ou output



André Camargo do Amaral - TDC SP 2016
Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Um exemplo básico

“Estou cansado”



André Camargo do Amaral - TDC SP 2016
Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

CTNBio - “Autorizações” para desenvolvimento de biotecnologia

COMISSÃO TÉCNICA NACIONAL DE BIOSSEGURANÇA

EXTRATO PRÉVIO Nº 5210/2016

O Presidente da Comissão Técnica Nacional de Biossegurança - CTNBio, no uso de suas atribuições e de acordo com o artigo 14, inciso XIX, da Lei 11.105/05 e do Art. 5, inciso XIX do Decreto 5.591/05, torna público que se encontra em análise na Comissão o processo a seguir discriminado:

Processo nº. 01200.001731/2016-54

Requerente: FuturaGene Brasil Tecnologia Ltda.

CQB: 325/11

Endereço: Av. Dr. José Lembo nº1010, sala A, Jardim Bela

Vista, Itapetininga - SP

Assunto: Liberação planejada no meio ambiente (RNO)

Ementa: A requerente solicita à CTNBio autorização para conduzir liberação planejada no meio ambiente de cucurbita geneticamente modificado - "Progenies provenientes de cruzamentos entre o evento geneticamente modificado TR679 com matrizes convencionais visando a seleção de clones". O ensaio será conduzido na Fazenda Fortaleza - Araraquara/SP. A CTNBio informa que, de acordo com a Portaria MCT nº 146/2006, fica mantido o sigilo para os genes e seus elementos regulatórios, constantes nos Anexos 1 e 2. A CTNBio esclarece que este extrato prévio não exime a requerente do cumprimento das demais legislações vigentes no país, aplicáveis ao objeto do requerimento. A CTNBio informa que o público terá trinta dias para se manifestar sobre o presente pleito, a partir da data de sua publicação. Solicitações de maiores informações deverão ser encaminhadas, via Sistema de Informação ao Cidadão - SIC, através da página eletrônica do Ministério da Ciência, Tecnologia, Inovações e Comunicações.

EDIVALDO DOMINGUES VELINI

Como Extrair e Processar?

- Regex
- Dicionários Léxicos
- Corpus Linguísticos
- Inteligência Artificial

André Camargo do Amaral - TDC SP 2016

Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Estruturação possibilita diversas aplicações

Extratos do Diário Oficial da União

[Veja esse email no seu navegador](#)



Diário Oficial da União

Comissão Técnica Nacional de Biossegurança

Terça-feira, 05 de Julho de 2016

Extrato Prévio

Requerente: Bayer SA

Número Processo: 01200.004010/1996-19

Número Extrato: 5.225/2016

Data da Publicação: 05/07/2016

Assunto: Extensão de CQB.

Ementa: Solicita extensão de CQB para inclusão da Sala de Desenvolvimento Técnico de Trait (DTT) localizada na Fazenda Ilha Bela II, Luis Eduardo Magalhães/BA. As atividades a serem desenvolvidas serão transporte, avaliação de produto, descarte e armazenamento de plantas classificadas na Classe de Risco I. A CTNBio esclarece que este extrato prévio não exime a requerente do cumprimento das demais legislações vigentes no país, aplicáveis ao objeto do requerimento. A CTNBio informa que o público terá trinta dias para se manifestar sobre o presente pleito, a partir da data de sua publicação. Solicitações de maiores informações deverão ser encaminhadas via SIC (Serviço de Informação ao Cidadão).

EDIVALDO DOMINGUES VELINI

COMISSÃO TÉCNICA NACIONAL DE BIOSSEGURANÇA

EXTRATO PRÉVIO Nº 5.225/2016

O Presidente da Comissão Técnica Nacional de Biossegurança - CTNBio, no uso de suas atribuições e de acordo com o artigo 14, inciso XIX, da Lei 11.105/05 e do Art. 5, inciso XIX do Decreto 5.591/05, torna público que encontra-se em análise na Comissão o processo a seguir discriminado:

Processo nº: 01200.004010/1996-19

Requerente: Bayer SA

Próton: 37.333/2016

CQB: 05/96

Assunto: Extensão de CQB.

Ementa: Solicita extensão de CQB para inclusão da Sala de Desenvolvimento Técnico de Trait (DTT) localizada na Fazenda Ilha Bela II, Luis Eduardo Magalhães/BA. As atividades a serem desenvolvidas serão transporte, avaliação de produto, descarte e armazenamento de plantas classificadas na Classe de Risco I.

A CTNBio esclarece que este extrato prévio não exime a requerente do cumprimento das demais legislações vigentes no país, aplicáveis ao objeto do requerimento.

A CTNBio informa que o público terá trinta dias para se manifestar sobre o presente pleito, a partir da data de sua publicação. Solicitações de maiores informações deverão ser encaminhadas via SIC (Serviço de Informação ao Cidadão).

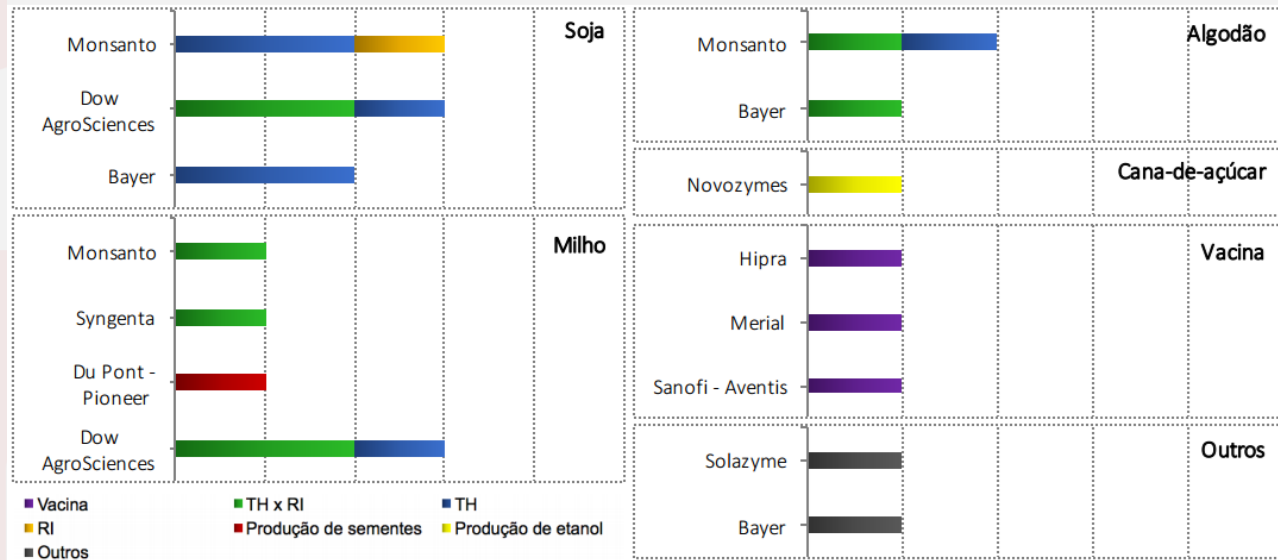
EDIVALDO DOMINGUES VELINI

André Camargo do Amaral - TDC SP 2016

Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Relatório por cultura

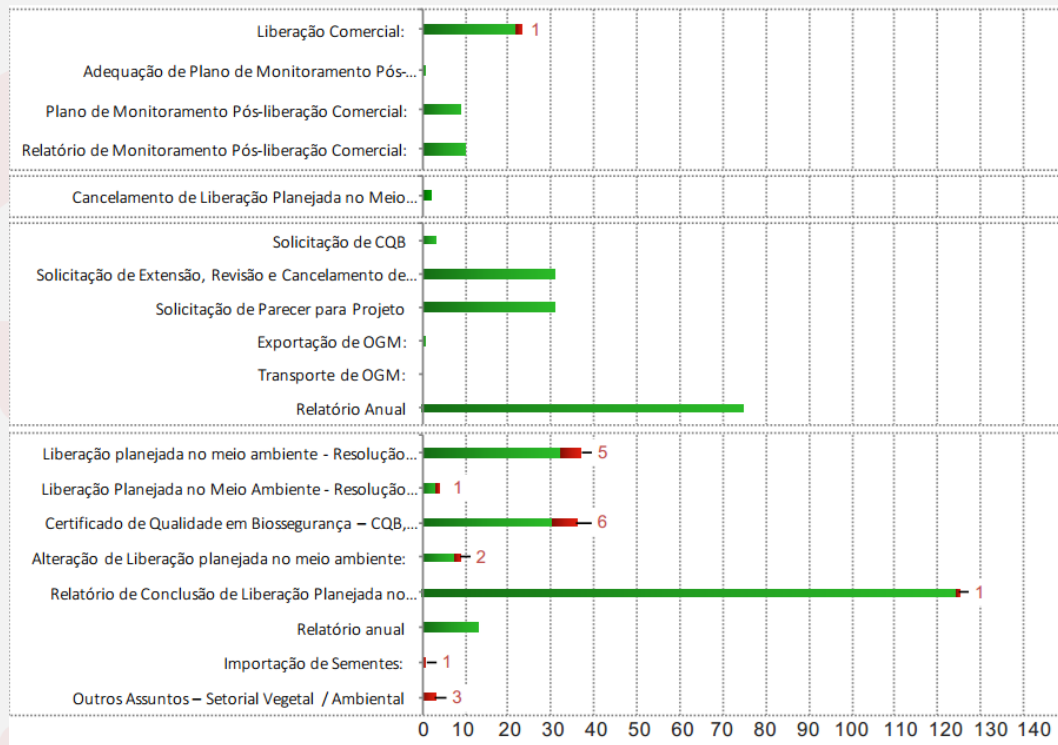
LIBERAÇÃO COMERCIAL



André Camargo do Amaral - TDC SP 2016
Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Relatório por periodo

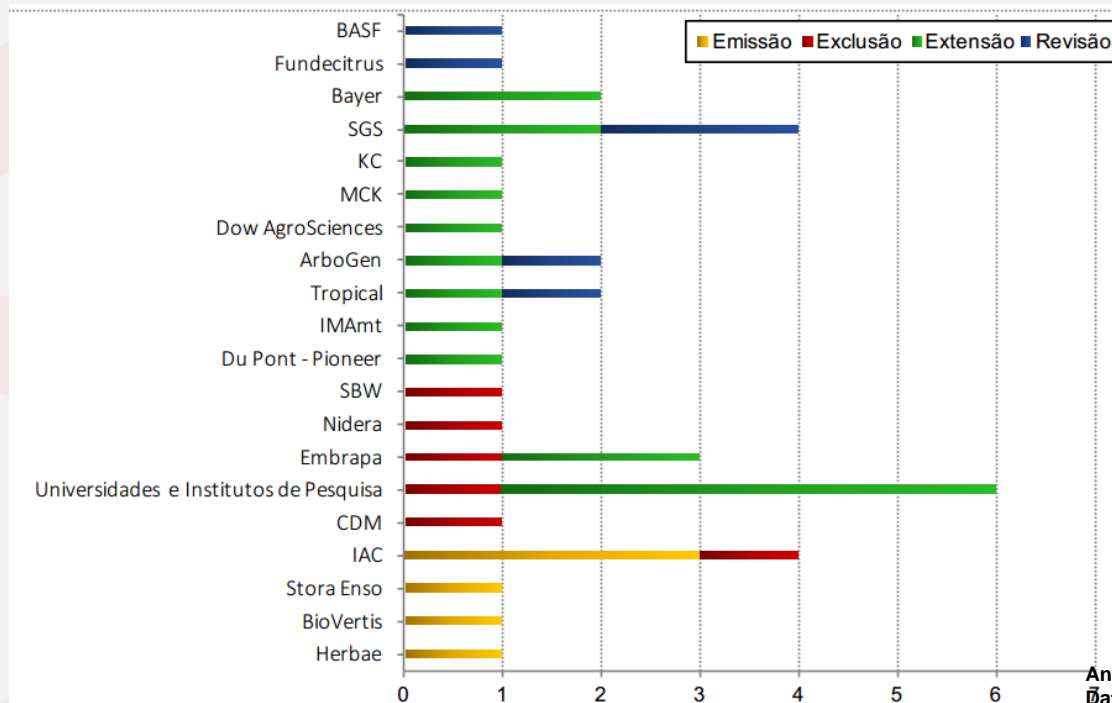
VISÃO GERAL



André Camargo do Amaral - TDC SP 2016
Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Relatório estratégico

CQB – AMBIENTAL / VEGETAL



André Camargo do Amaral - TDC SP 2016

Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist

Introdução ao Big Data

Como traduzir? “**Bonjur**” → “Bom dia”, “olá”, “passe bem”, “oi”

Google

Tradutor

português

inglês

espanhol

Detectar idioma

▼



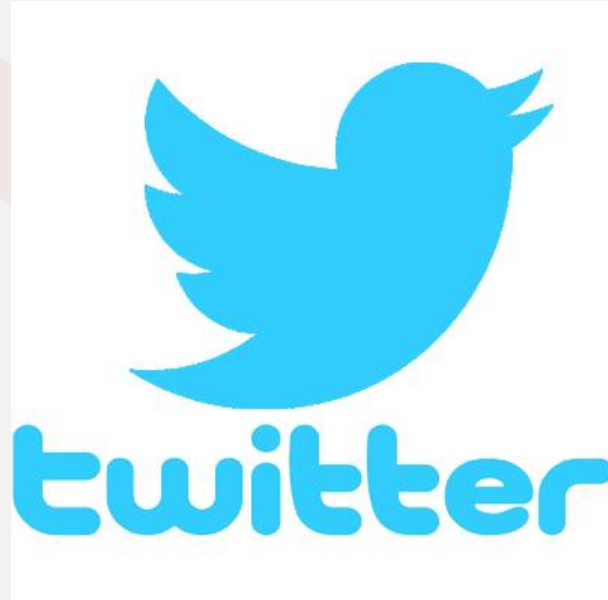
O sistema do Google funciona bem não porque é o algoritmo mais inteligente mas foi alimentado com mais dados

Mais é melhor que menos, e as vezes, melhor que mais inteligente.

■ Alguns “probleminhas”

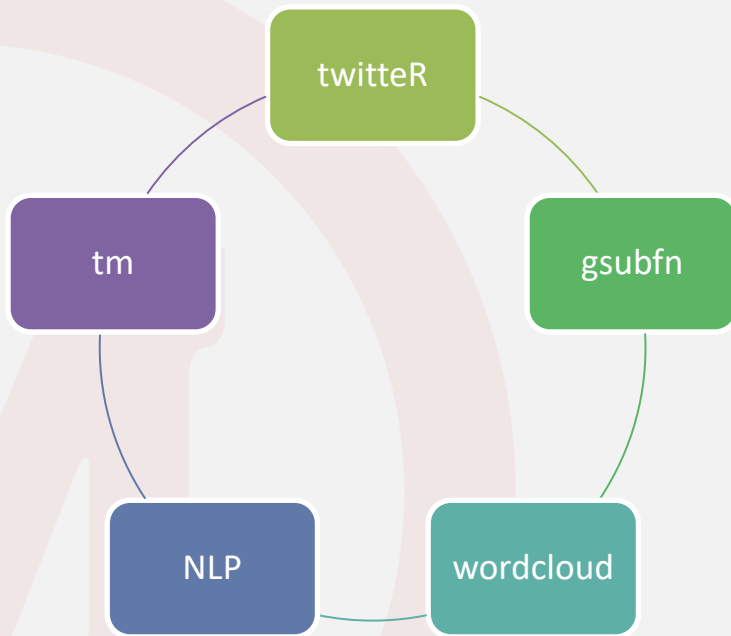
- Dificuldade em processar textos extremamente técnicos em português (dependendo da Abordagem)
- Necessidade de especialistas nos assuntos dos conteúdos publicados
- Em algumas abordagens estatísticas, conjunto de treinamento é necessário
- Falta de estruturação de dados em algumas fontes
- Grande quantidade de PDFs que precisam ser convertidos e nesse processo perdem a formatação original
- Necessidade de Dicionários técnicos sobre determinados assuntos para ajudar o processamento do texto
- *Dificuldades específicas para determinado objetivo

André Camargo do Amaral - TDC SP 2016
Data Science – Processamento de Linguagem Natural como uma ferramenta poderosa nas mãos de um data scientist



<https://developer.twitter.com/en/apps>

■ Pacotes usados



Instalar e
Carregar

Autenticação no Twitter

#Autenticando no Twitter

```
options(httr_oauth_cache=T)
consumer_key <- " "
consumer_secret <- " "
access_token <- " "
access_secret <- " "

setup_twitter_oauth(
  consumer_key,
  consumer_secret, access_token,
  access_secret)
```

Integracao com R

[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)

Consumer Secret (API Secret)

Access Level Read and write ([modify app permissions](#))

Owner AlunosDoNogare

Owner ID 866308421636718592

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token

Access Token Secret

Access Level Read and write

Owner AlunosDoNogare

Owner ID 866308421636718592

twitterR

■ Limpeza de texto

`gsub(pattern="[:punct:]", texto, replacement="")` #Remove Pontuação

`gsub("\\s?(f|ht)(tp)(s?)(:|/)([^\\.]*)[\\.|/](\\S*)", texto, replacement="")` #Remove link

`gsub("\\n", texto, replacement=" ")` #Remove pulo de linha

`tolower(texto)` #Mantém tudo minúsculo

`removeWords(texto, stopwords('portuguese'))` #Remover Stopwords em Português

Montar a Nuvem de Palavras

#Cria um vetor a partir da origem

```
VectorSource(texto)
```

```
docs <- c("Este é um texto", "Este aqui é outro")  
vs <- VectorSource(docs)  
vs$content[1]  
vs$content[2]
```

#Cria os elementos a partir dos conteúdos

```
Corpus(VectorSource(texto))
```

Montar a Nuvem de Palavras

#Cria uma matrix a partir do documento Corpus

```
m <- as.matrix( TermDocumentMatrix(docCorpus) )
```

#Cria uma estrutura ordenada decrescente

```
v <- sort(rowSums(m),decreasing=TRUE)
```

```
x <- cbind(x1 = 3, x2 = c(4:1, 2:5))  
rowSums(x);
```

#Cria um DataFrame com o termo e a frequencia

```
d <- data.frame(word = names(v), freq=v)
```


■ Autenticando no Twitter

#Conectando e autenticando no Twitter

```
usuario <- "DiegoNogare"
```

```
perfil <- getUser(usuario)
```

```
location(perfil)
```

O resultado do «location» deve ser "São Paulo"

Isso vai garantir que estamos conectados ao twitter

twitterR

■ Lendo posts de um determinado termo

#Recupera os tweets

```
searchTwitter(termo, n=totalTweets, since=desde, lang =  
idioma)
```

#Converte o objeto em um Data Frame

```
twListToDF(posts)
```

#Não esqueça de
limpar o texto!

twitterR

■ Criar a nuvem de palavras

```
wordcloud(  
  words = d$word,  
  freq = d$freq,  
  min.freq = 1,  
  max.words=150,  
  random.order=FALSE,  
  colors=brewer.pal(8, "Dark2")  
)
```

wordcloud

■ Salvando a nuvem em um arquivo

#Salvando em um arquivo físico

```
png(nomeArquivo, width=1900, height=1900, units="px",  
res=300)
```

#Aqui vai o código da sua nuvem

ou de um gráfico

```
dev.off()
```

#Postar no twitter com a nuvem

```
updateStatus("Seu post"  
  ,mediaPath = SeuArquivoDeMidia )
```

A mídia é opcional

O arquivo pode ser a nuvem
de palavras ou um gráfico



TWEETS
1

MOMENTS
0

Tweets

Tweets e respostas



Alunos do Nogue

@AlunosDoNogare · 58 seg

Postando diretamente de dentro do R





Alunos do Nogue

@AlunosDoNogare

Este perfil tem o objetivo de compartilhar atividades de aulas. Todos os tweets são gerados por alunos e são de suas responsabilidades...

 [Sao Paulo, Brazil](#)

 Participa desde maio de 2017

 Nasceu em 20 de setembro de 1984

Pós-Graduação em **Inteligência Artificial** – Disciplina **Big Data e Visualização de Dados**



Referências

Pacote: TwitterR

<https://cran.r-project.org/web/packages/twitterR/twitterR.pdf>

Limpeza de Textos:

```
gsub(pattern="[:punct:]", texto, replacement="")
```

<https://stat.ethz.ch/R-manual/R-devel/library/base/html/regex.html>





Faculdade de Computação e Informática
Mackenzie