EMERGING CODE

Text Mining:
Conceitos e práticas usando R

JUNHO 2021

Charles Mendes de Macedo

Senior Software Engineer
MCSD, MCSA, MCTS, MSc at USP

Agenda

1. O que é Text Mining?
    - Objetivo
2. Tecnologias
3. Demo 1
4. Técnicas
    - Tokenização
    - Stop words
    - Bag of words
    - Nuvem de palavras
    - Análise quantitativa de texto
    - N-Gram
5. Demo 2

O que é Text Mining?

"É uma campo de estudo dentro Data Mining. É o processo para descobrir conhecimento em dados textuais (dados não estruturados)."

"80% das informações em empresas estão contidas em documentos textuais (e-mails, mensagens, etc)." [1]

– Ah-hwee Tan

Text Mining: Conceitos e práticas usando R

[1] Text Mining: the state of the art and the challenges, Ah-hwee Tan – 2000.

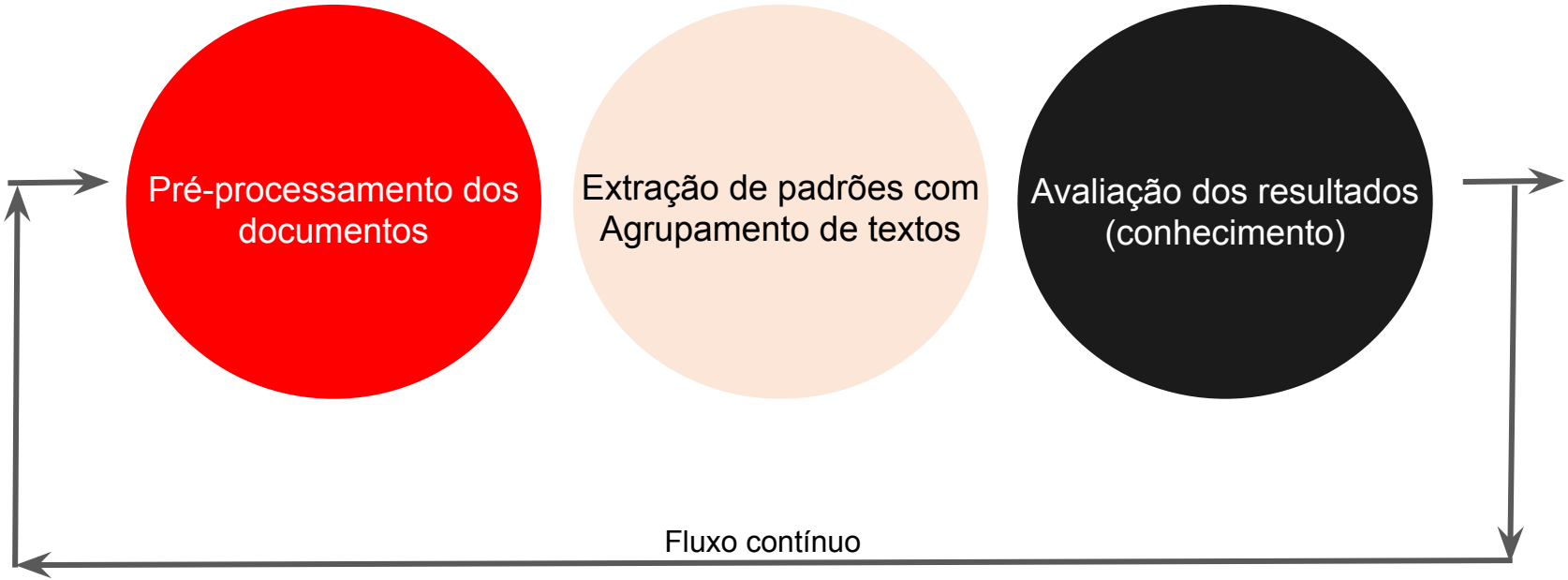
Objetivos

Principais tarefas feitas com Text Mining:

- Análise quantitativa de texto;
- Análise de sentimento;
- Agrupamento;
- Sumarização de textos;
- Reconhecimento de entidades.

Text Mining: Conceitos e práticas usando R

Processo de Text Mining

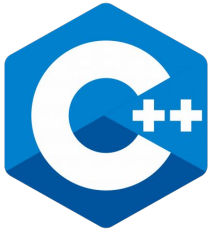


[2] REZENDE, S. O., MARCACINI, R. M., MOURA, M. F. / Revista de Sistemas de Informacao da FSMA n. 7 (2011) pp. 7-21.

Tecnologias

Principais linguagens para fazer Text Mining

Microsoft SQL Server



Text Mining: Conceitos e práticas usando R

Tecnologias

Na demonstração serão usadas:


PRTG
NETWORK
MONITOR

Demo 1 : Pré-processamento

- O que a empresa Farfetch faz?

Texto Original

Hide

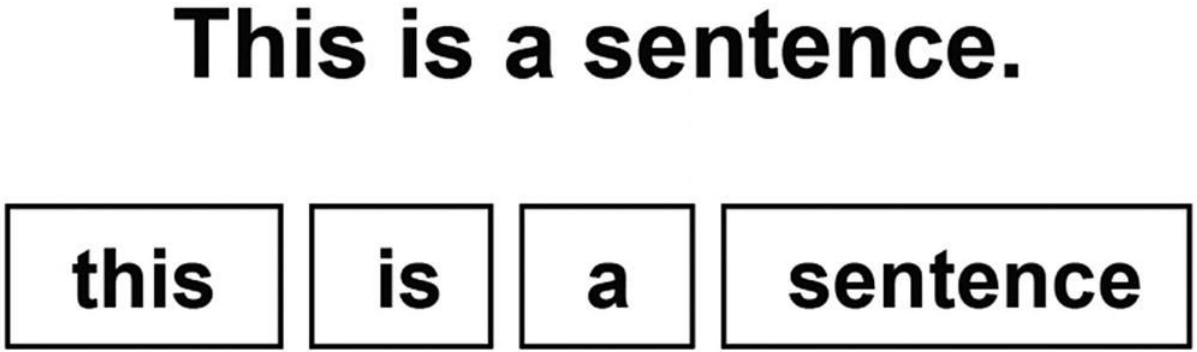
```
farfetch_wikipedia <- "Farfetch é uma empresa de e-commerce de moda de luxo nacional e internacional que vende marcas famosa
s e moda premium atual para comprar online. O site foi fundado em 2008 pelo empresário português José Neves, e a empresa tem
sede fiscal em Londres e principais filiais em Porto, Guimarães, Lisboa, Nova Iorque, Los Angeles, Tóquio, Xangai e São Paul
o. A Farfetch trabalha em parceria com boutiques independentes em um modelo baseado em comissão, vantajoso para ambas as par
tes, uma vez que as vendas online representam em média, 30% do total vendido pelas boutiques. A empresa opera em diversos me
rcados internacionais com sites no idioma local, como inglês, francês, japonês, mandarim, português, coreano, alemão, russo
 e espanhol. Desde maio de 2015, a Farfetch conta com mais de 1000 funcionários globalmente."

farfetch_revista_exame <- "A Farfetch foi lançada em outubro de 2008, no início da crise mundial causada pela bolha imobiliá
ria dos Estados Unidos, por José Neves, empreendedor português baseado em Londres. O timing não poderia ser pior, mas Neves
 desde o princípio orientou o negócio para atender os consumidores das marcas mais exclusivas do mundo – e, como costumam di
zer os que lidam com o mercado de luxo, nesse segmento não há crise. Podia ser verdade ou pensamento positivo, mas a Farfetc
h cresceu de maneira espetacular enquanto os países mais ricos do mundo entravam na pior crise em um século. A empresa deve
 vender 340 milhões de dólares neste ano, quase o dobro de 2013.O valor reflete as transações feitas no site – a Farfetch fi
ca com uma porcentagem a título de comissão e serviço. A Farfetch está mais próxima de uma plataforma de transações, como um
eBay ou um MercadoLivre, do que de uma pura loja virtual, como a Amazon, mas as coisas são um pouco mais complicadas do que
 isso."
corpus <- c(farfetch_wikipedia, farfetch_revista_exame)
corpus
```

Text Mining: Conceitos e práticas usando R

Técnicas: Tokenização

Dividir uma frase ou sentença na menor unidade, como uma palavra ou um termo.

Text Mining: Conceitos e práticas usando R



Técnicas: Stop Words

É um conjunto de palavras comuns e frequentes, usadas no idioma e que podemos removê-las.

Text Mining: Conceitos e práticas usando R

Técnicas: Bag of words

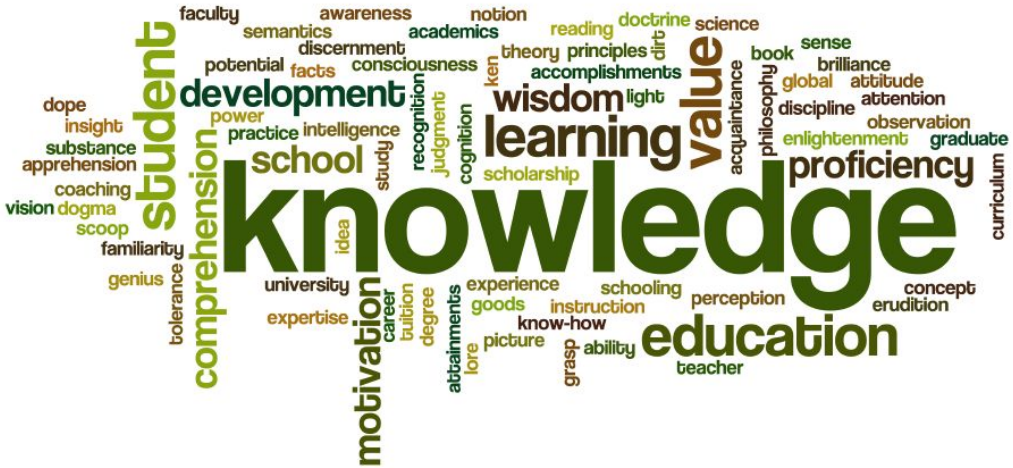
É uma representação descreva a ocorrência das palavras dentro de um documento.

Text Mining: Conceitos e práticas usando R

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Técnicas: Nuvem de palavras

É uma representação gráfica da frequencia de palavras, destacando os termos mais frequentes.



Text Mining: Conceitos e práticas usando R

Técnicas: Análise quantitativa de texto

É uma representação da frequencia e relevância de palavras de um documento ou dentro de uma coleção de documentos.

Text Mining: Conceitos e práticas usando R

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

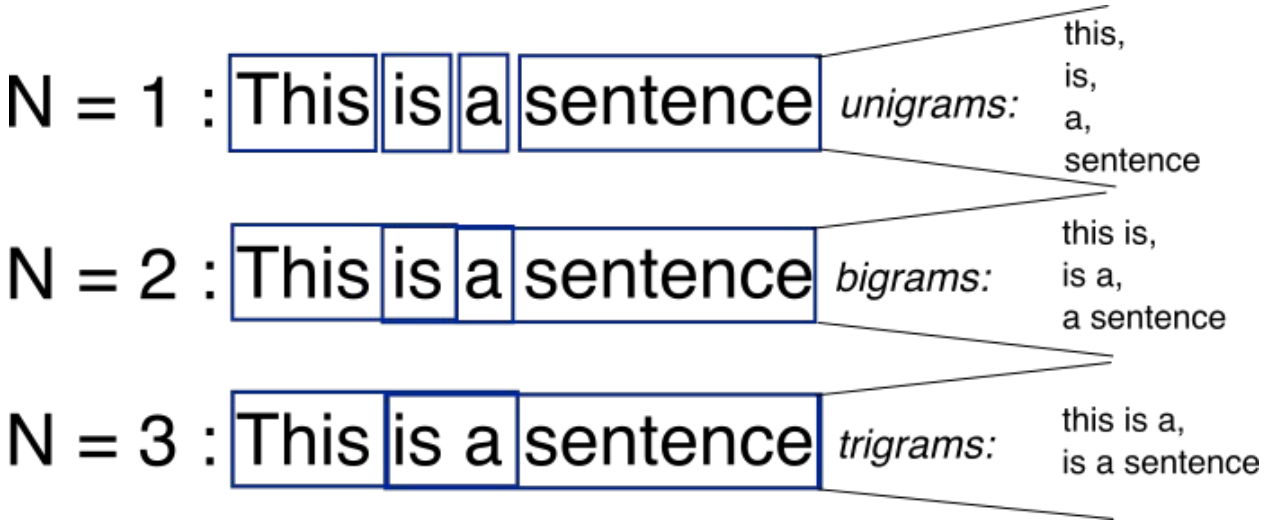
$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

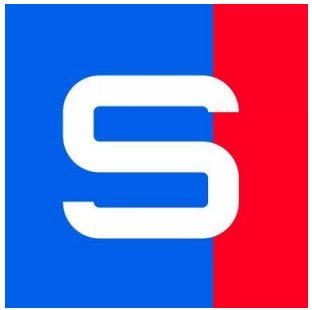
Técnicas: N-gram

Junção de N termos em uma unidade.



Demo 2

Aplicar essas técnicas básicas de Text Mining em tweets dos quatro grandes canais de Futebol do Brasil.



Mais aplicações

- Dissertação: *Aplicação de algoritmos de agrupamento para descoberta de padrões de defeito em software JavaScript.*

  Link: *https://teses.usp.br/teses/disponiveis/100/100131/tde-29012019-152129/pt-br.php*

(i)

Watcher doesn't restart a process which is being stopped

Browse files

master (#1102) v2.1.4 ... 0.12.8

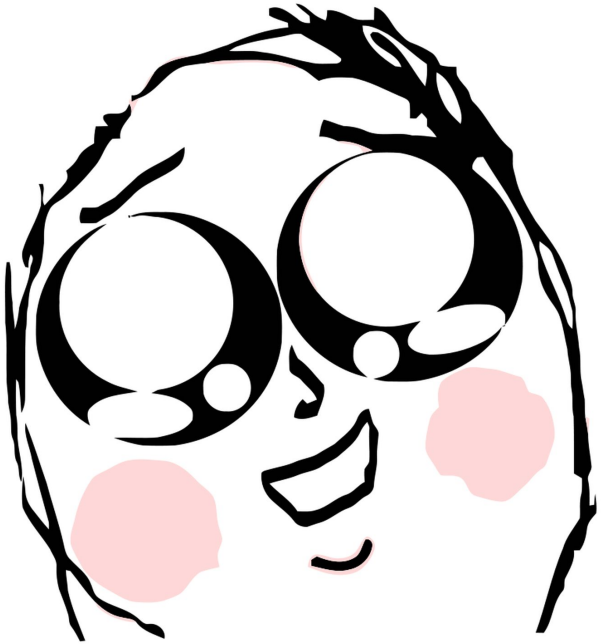
(ii)

6 lib/God/DeprecatedCalls.js View file

```
@@ -112,10 +112,12 @@ module.exports = function(God) {
112 112         return cb(God.logAndGenerateError('Unknown process'), {});
113 113
114 114       async.eachLimit(processes, cst.CONCURRENT_ACTIONS, function(proc, next) {
115     -       if (proc.pm2_env.status == cst.ONLINE_STATUS)
    115 +       if (proc.pm2_env.status === cst.ONLINE_STATUS)
116 116           return God.restartProcessId({id:proc.pm2_env.pm_id}, next);
117     -       else
    117 +       else if (proc.pm2_env.status !== cst.STOPPING_STATUS)
118 118           return God.startProcessId(proc.pm2_env.pm_id, next);
    119 +       else
    120 +         return next();
119 121       }, function(err) {
120 122         if (err) return cb(God.logAndGenerateError(err));
121 123         return cb(null, God.getFormatedProcesses());
```

RE_C_R_eq	RE_C_R_sheq	RE_C_I_sheq	RE_E_I_return	RE_C_I_return
1	0	2	1	0

Que legal! Gostei de Text Mining!



Text Mining: Conceitos e práticas usando R

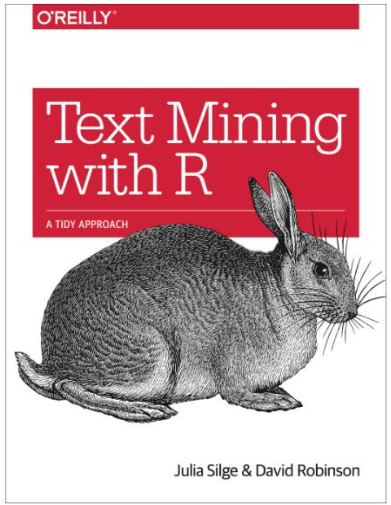
Iniciando os estudos - Cursos

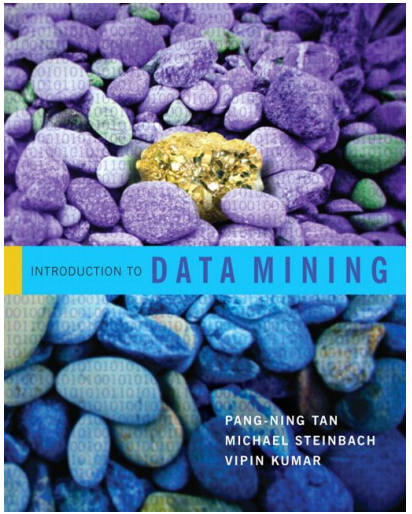
- Coursera:
  - Text Mining and Analytics;
  - Machine Learning, by Andrew Ng;
  - Data Science.

- EDX:
  - Data, Analytics and Learning.

Text Mining: Conceitos e práticas usando R

Iniciando os estudos - Livros





https://www.tidytextmining.com/

Text Mining: Conceitos e práticas usando R

Thank you

Contatos

- **GitHub**: MackMendes
- **Twitter**: @CharlesMendesMa
- **Linkedin:** linkedin.com/in/charles-mendes-de-macedo/
- **Blog**: http://charlesmms.azurewebsites.net/