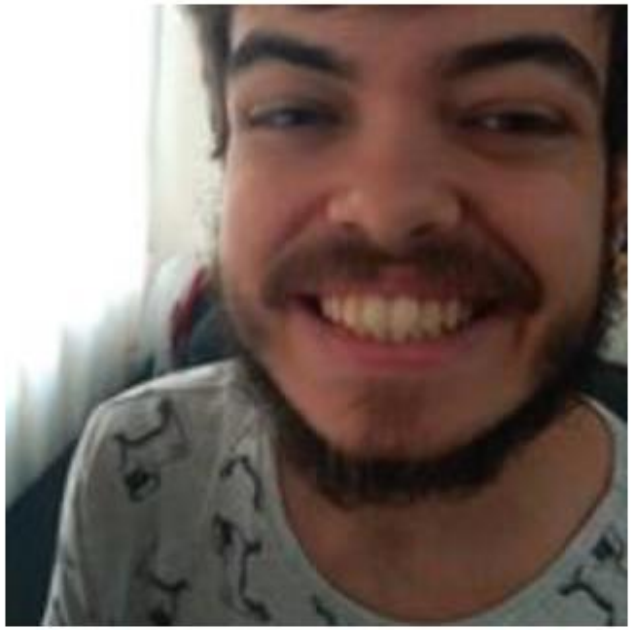
Introdução à Mineração de Textos

Andrei Martins e Charles Mendes

Quem somos – Andrei Martins



- **Cientista de Dados** na Genoa Performance, onde atua com o desenvolvimento de sistemas inteligentes para otimização do atendimento em Call Centers.
- Como parte de seu **mestrado na Universidade de São Paulo** conduz pesquisa na área de **Sistemas de Recomendação**.

Quem somos – Charles Mendes

- **Cientista de Dados** e **Analista Desenvolvedor** na Genoa Performance, atua com o desenvolvimento de sistemas inteligentes para otimização do atendimento para empresas de Telecomunicação.
- **Estudante de Mestrado pela Universidade de São Paulo**, com pesquisa na área de Sistemas Inteligentes aplicado na detecção automática de falhas no desenvolvimento de software.

Mineração de Textos – O que é?

- Mineração de texto é o processo de descoberta de conhecimento a partir de conteúdo textual (não estruturado).

- Segundo Ah-hwee Tan [1], **80% das informações** de uma companhia estão contidas em documentos textuais (registro de histórico de atividades, memorandos, documentos, e-mails, projetos, estratégias, etc).

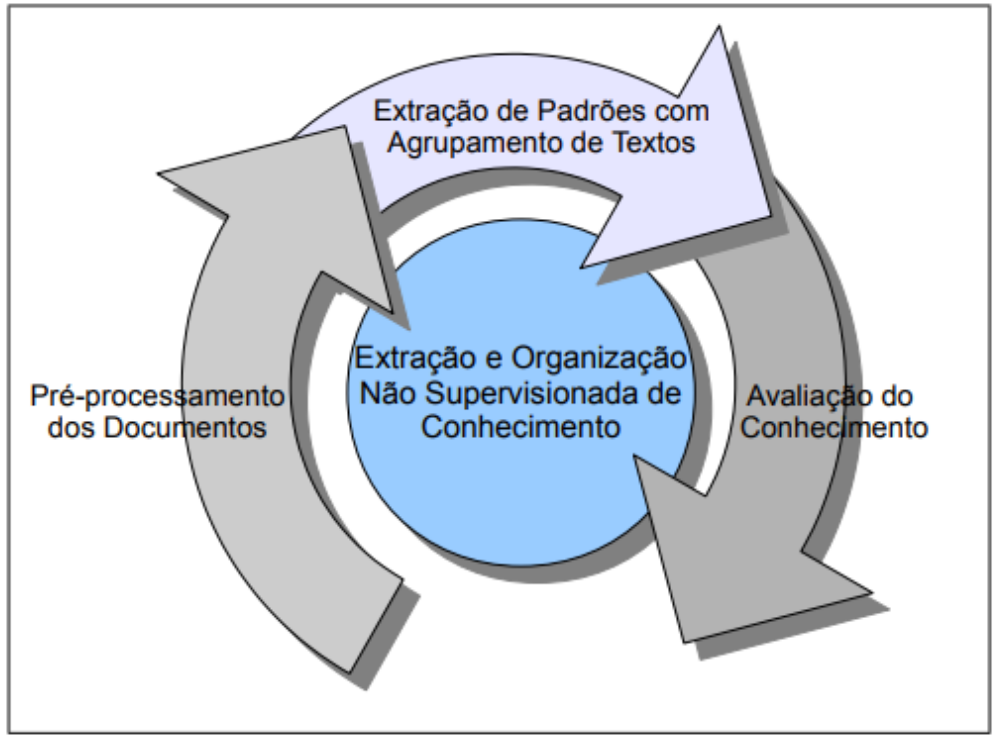
[1] Text Mining: the state of the art and the challenges, Ah-hwee Tan – 1999.

Mineração de Textos – Principais Tarefas

- As principais tarefas que são feitos dentro da Mineração de Texto:
    - Reconhecimento de entidades;
    - Análise de sentimento;
    - Análise quantitativa do texto;
    - Classificação;
    - Agrupamento;
    - Sumarização de textos;
    - Outras...

Mineração de Textos - Principais Etapas

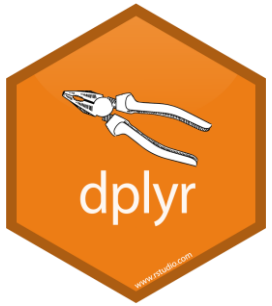
- “Um processo de Mineração de Textos para extração e organização não supervisionada de conhecimento pode ser dividido em três fases principais: Pré-Processamento dos Documentos, Extração de Padrões com Agrupamento de Textos e Avaliação do Conhecimento.” [2]



[2] REZENDE, S. O., MARCACINI, R. M., MOURA, M. F. / Revista de Sistemas de Informacao da FSMA n. 7 (2011) pp. 7-21. http://www.fsma.edu.br/si/edicao7/FSMA_SI_2011_1_Principal_3.pdf

Tecnologias que utilizaremos

- Para o desenvolvimento deste *meetup*, vamos utilizar:



Entre outros...

Demo 1 – Pré-processamento

Pré-processamento

Code

Andrei Martins e Charles Mendes

04 de Novembro de 2017

Hide

```
if(!require(quanteda)) install.packages("quanteda")
```

Texto Original

Hide

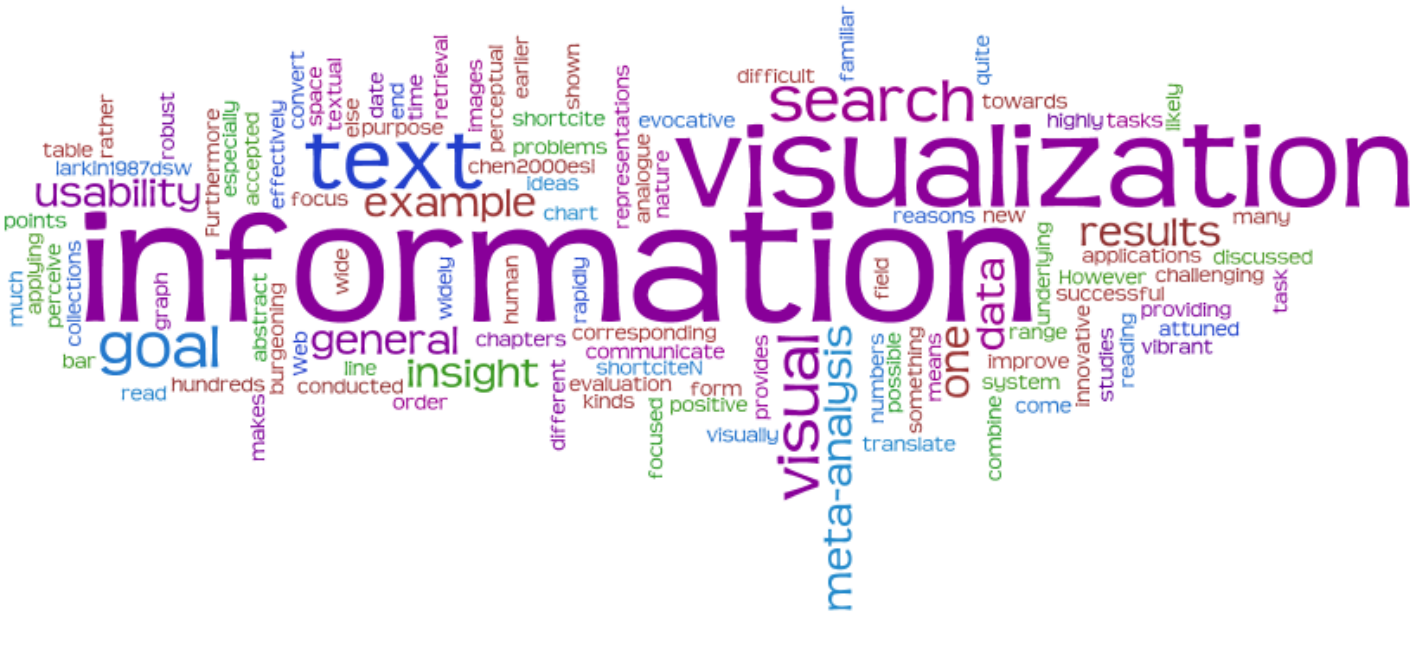
```
andrei <- "Cientista de Dados na Genoa Performance, onde atua com o desenvolvimento de sistemas inteligentes para otimização
 do atendimento em Call Centers. Como parte de seu mestrado na Universidade de São Paulo conduz pesquisa na área de Sistemas
 de Recomendação."
charles <- "Cientista de Dados e Analista Desenvolvedor na Genoa Performance, atua com o desenvolvimento de sistemas intelig
entes para otimização do atendimento para empresas de Telecomunicações. Estudante de Mestrado pela Universidade de São Paul
o, com pesquisa na área de Sistemas Inteligentes aplicado na detecção automática de falhas no desenvolvimento de software."
corpus <- c(andrei, charles)
corpus
```

```
[1] "Cientista de Dados na Genoa Performance, onde atua com o desenvolvimento de sistemas inteligentes para otimização do at
endimento em Call Centers. Como parte de seu mestrado na Universidade de São Paulo conduz pesquisa na área de Sistemas de Re
comendação."
[2] "Cientista de Dados e Analista Desenvolvedor na Genoa Performance, atua com o desenvolvimento de sistemas inteligentes p
ara otimização do atendimento para empresas de Telecomunicações. Estudante de Mestrado pela Universidade de São Paulo, com p
esquisa na área de Sistemas Inteligentes aplicado na detecção automática de falhas no desenvolvimento de software."
```

Demo disponível: https://github.com/MackMendes/MineracaoTexto-Nerdzao/blob/master/preproc.Rmd

Nuvem de palavras

- É uma representação gráfica de frequência de palavras, dando maior destaque para os termos mais frequentes.



Análise quantitativa do texto

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

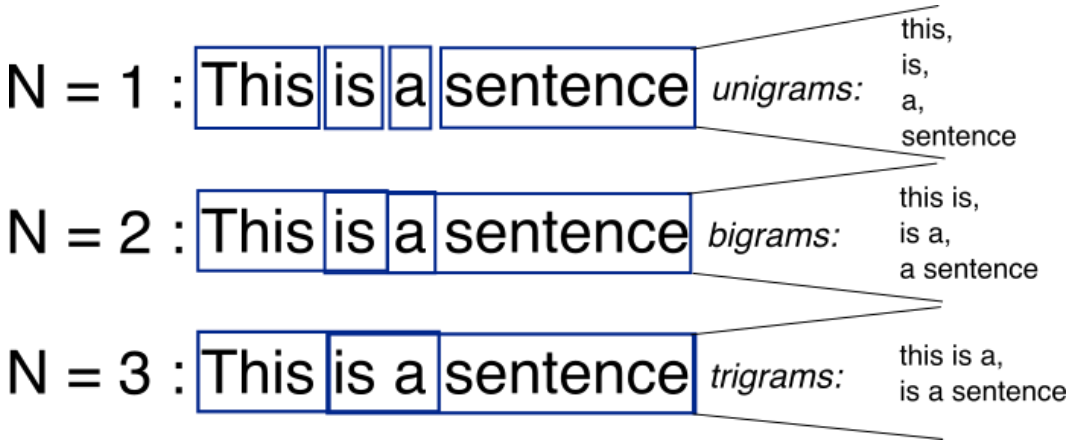
$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- Considere um corpus de 10 documentos (N = 10), o termo “Nerd” está contido em 1 documento (Doc1) e esse documento contém 100 palavras, então:
  - **Frequência das Palavras (TF)**: TF(“Nerd”, Doc1) = 4 / 100 = 0.04
  - **“Raridade da palavra” (IDF)**: IDF(Nerd) = log10(10/1) = 1
  - **TF * IDF** = 0.04 * 1 = 0.04

N-gram – (Bigrama, N=2)

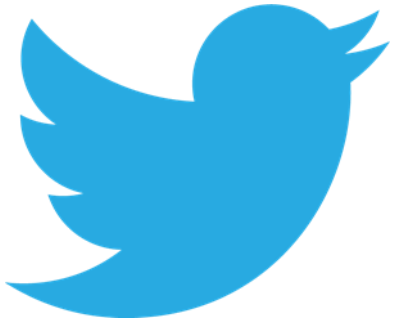
- Input: “Hoje vamos falar sobre Mineracao de Texto”.
- Output:

	bigrama
1	hoje vamos
2	vamos falar
3	falar sobre
4	sobre mineracao
5	mineracao de
6	de texto



Demo 2:

- Iremos aplicar técnicas básicas de Mineração de Texto em tweets dos quatros times paulistas com as maiores torcidas.



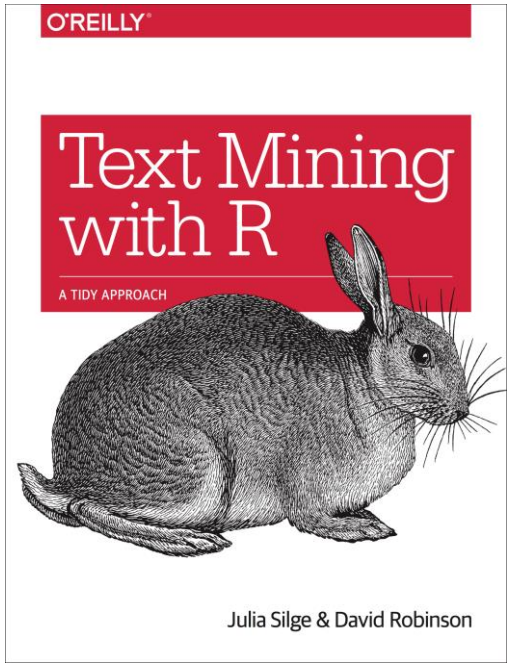
Demo disponível: https://github.com/MackMendes/MineracaoTexto-Nerdzao/blob/master/preproc.Rmd

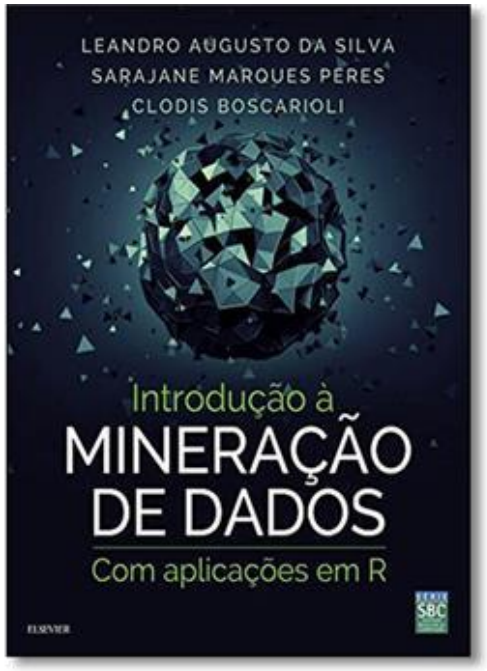
Começando os estudos...

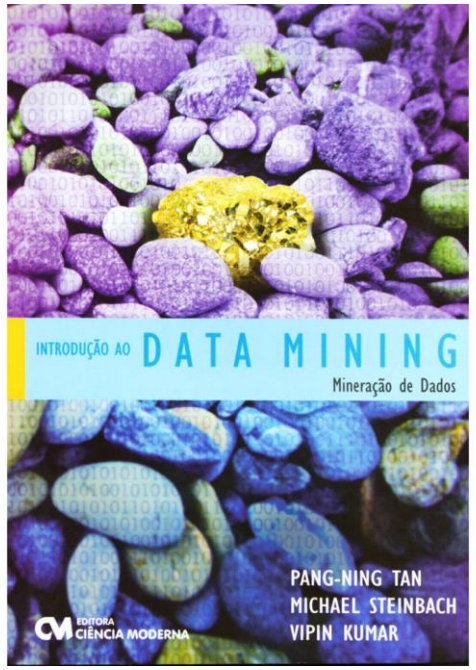
- **Cursos:**
  - **Coursera** (É possível assistir as aulas destes cursos gratuitamente! Se quiser obter o certificado de conclusão do curso, é necessário pagar o valor do curso):
    - Curso de "*Machine Learning*" do Andrew Ng: https://pt.coursera.org/learn/machine-learning
    - Uma série de cursos para a carreira "*Data Science*": https://pt.coursera.org/specializations/jhu-data-science
    - Curso de "*Text Mining and Analytics*": https://pt.coursera.org/learn/text-mining
  - **EDX**:
    - Curso "Data, Analytics and Learning": https://www.edx.org/course/data-analytics-learning-utarlingtonx-link5-10x

Começando os estudos...

- **Livros:**







http://tidytextmining.com/index.html

Começando os estudos...

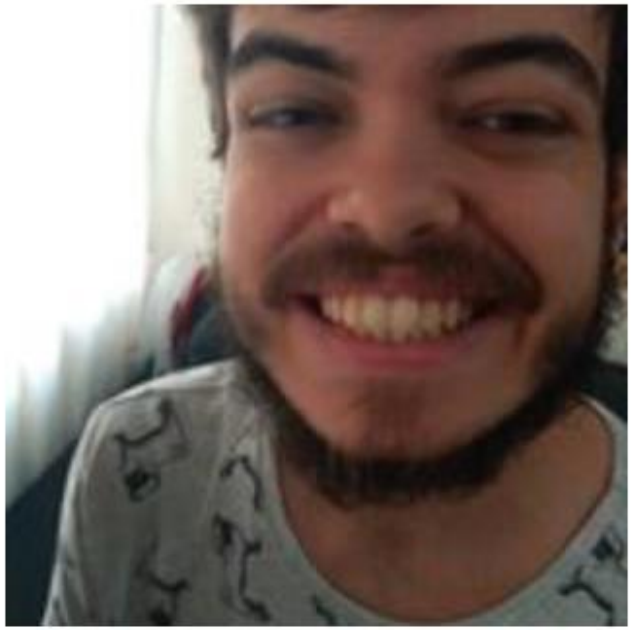
- **Mestrado (stricto sensu) - Universidade Pública**:
  Programa de Pós-graduação na USP:
  - **USP IME:** https://www.ime.usp.br/dcc/pos
  - **USP PPgSI:** http://ppgsi.each.usp.br/
  - **USP ICMC (São Carlos):** http://icmc.usp.br/pos-graduacao/ppgccmc/

Aluno Especial:

- **USP IME (Aluno especial):** https://www.ime.usp.br/dcc/pos/ae
- **USP PPgSI (Aluno especial):** http://ppgsi.each.usp.br/solicitacao/
- **USP ICMC (Aluno especial):** http://icmc.usp.br/pos-graduacao/ppgccmc/ingresso

Dúvidas?

Contatos



GitHub: github.com/a-n-d-r-e-i

Twitter: @andreisnitram

Linkedin: linkedin.com/in/amartins13

GitHub: github.com/MackMendes

Twitter: @CharlesMendesMa

Blog: charlesmms.azurewebsites.net

Linkedin: linkedin.com/in/charles-mendes-de-macedo