Mineração de Texto

Visão geral das tarefas e métodos

Prof. Walmes Zeviani

walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Text Mining

Definição

Análise de texto é sobre extrair informação.

Text mining é o processo de analisar um texto **desestruturado**, extrair informação relevante e transformá-la em estruturada de forma que possa ser aproveitada de diversas formas (HURWITZ et al., 2016).

> *The practice of text mining is aimed at undestanding and applying insights from the most complex analytical processing system in the universe - the human brain - to the analysis of written language.*

Motivação e exemplos

Texto e informação

- **Somos sensores** sobre o mundo e **registramos o que percebemos com texto**.
- Quando lemos um livro, recordamos das sensações mas não da prosa.
- Tratamos a informação de texto na sociedade assim também.
- Acredita-se que a informação em texto sobre o mundo hoje é tão rica que as máquinas poderiam dominar o mundo.

Dados de texto são abundantes

Opinião do consumidor

1. http://www.carrosnaweb.com.br/opiniaolista.asp.
2. https://www.reclameaqui.com.br/.
3. https://www.consumidor.gov.br/.
4. http://www.macworld.co.uk/review/iphone/.

Descoberta de tópicos e tendências

1. https://twitter.com/search-advanced?lang=pt.
2. http://www1.folha.uol.com.br/mercado/.
3. http://www.valor.com.br/opiniao.
4. https://www.ncbi.nlm.nih.gov/pubmed.
5. http://apps.webofknowledge.com/.
6. http://www.sciencedirect.com/.
7. http://cnpq.br/projetos-pesquisa.

Dados de texto são abundantes

Oportunidades de emprego

1. http://www.catho.com.br/.
2. https://www.indeed.com.br/.
3. https://www.bne.com.br/.
4. https://www.infojobs.com.br/.

Similaridade e agrupamento

1. https://www.cifraclub.com.br/.
2. http://www.tudogostoso.com.br/.

Modelagem preditiva

1. http://www.infomoney.com.br/
2. https://www.webmotors.com.br/
3. http://www.imovelweb.com.br/.

Alguns casos de aplicação de análise de texto

1. Descoberta de ameaças terroristas.
2. Mapear focos de dengue (UFMG) e demais problemas de saúde pública.
3. Fornecer diagnóstico de doença pelo relato de caso (IBM Watson).
4. Melhorar qualidade de produto pelo relato dos consumidores.
5. Aproveitar conversas transcritas de telemarketing.
6. Registros de call center.
7. Escrita para aumentar sucesso no desfecho de petições/processos.
8. Classificação de documentos para busca em biblioteca.

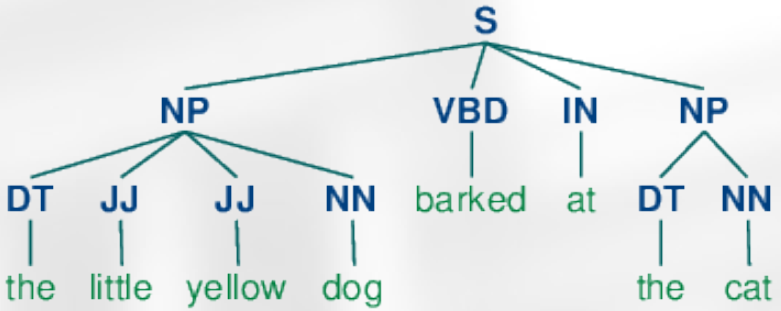
Tipos de formato de documentos com texto

- Dados não estruturados = estrutura imprevisível.
- Exemplos: texto, imagem, aúdio, vídeo, etc.

Nota fiscal	Notícia	Tweet
pré estrutura números e campos	organização língua formal	coloquial e curto abreviações e hashtags

Abordagens principais

Análise sintática



NLP

- Análise lexical/morfológica: formas da palavra.
- Análise sintática: estrutura gramátical, criar contexto.
- Análise semântica: determinar significado, eliminar ambiguidades.
- Análise do âmbito do discuso: significado além do discurso, inferência.
- É uma análise complexa que pode determinar: quem, o que, quanto onde e porquê.

Saco de palavras (*baf-of-words*)

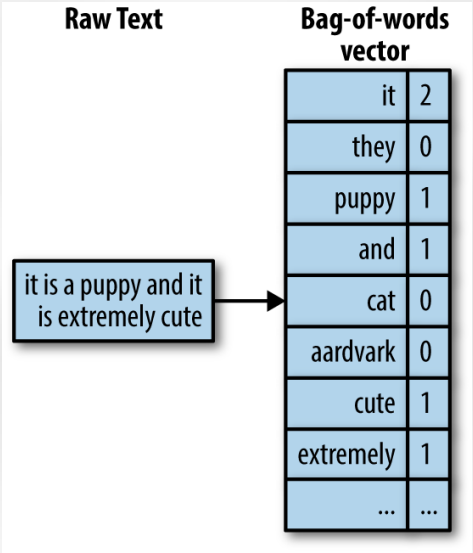
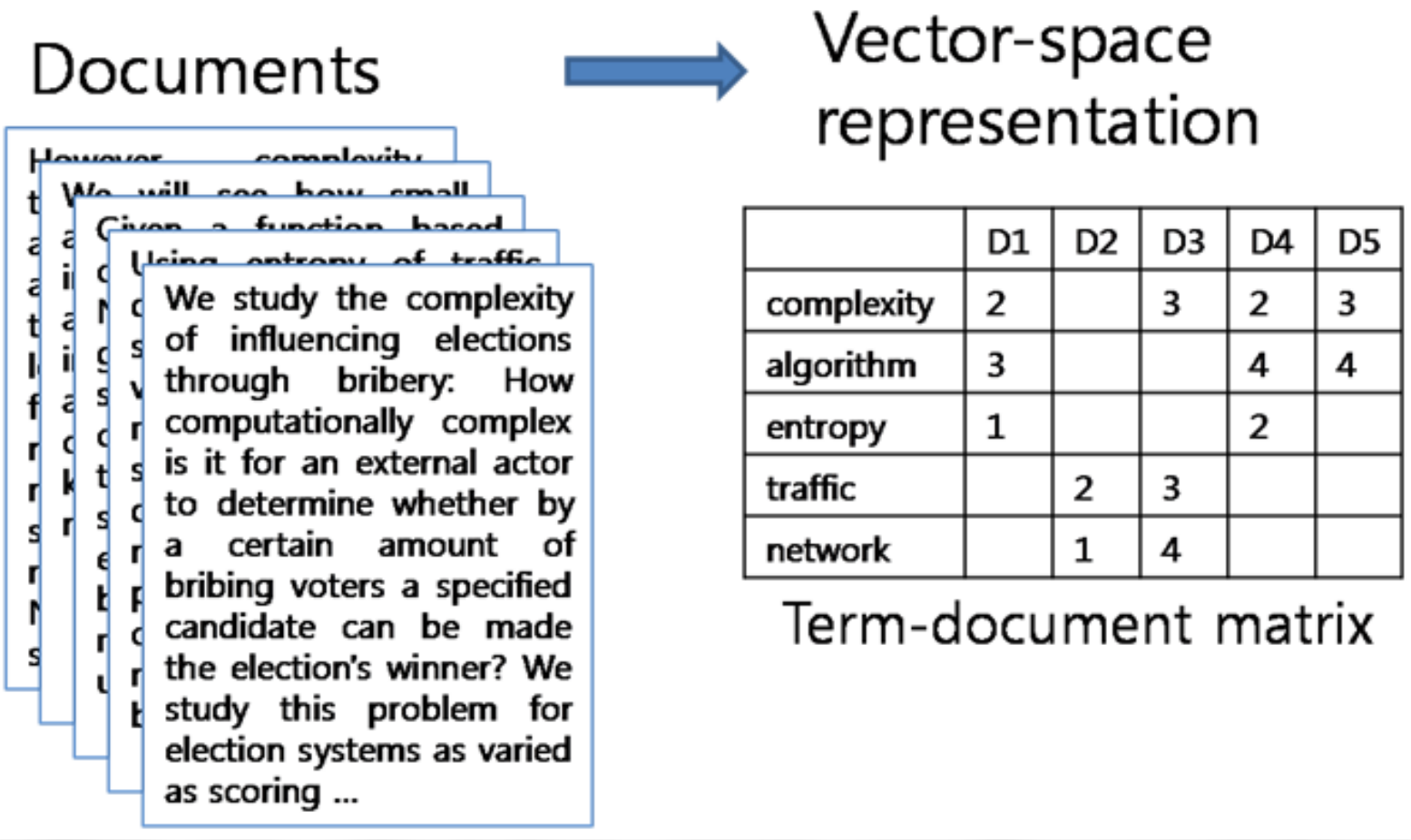


Figura 1. `http://uc-r.github.io/creating-text-features`.

- As frases são desfeitas.
- Cada palavra é um termo.
- Representa-se quantas vezes cada um ocorre no documento.
- Estrutura linguística é ignorada.
- Apesar de simples, é muito robusta e útil.

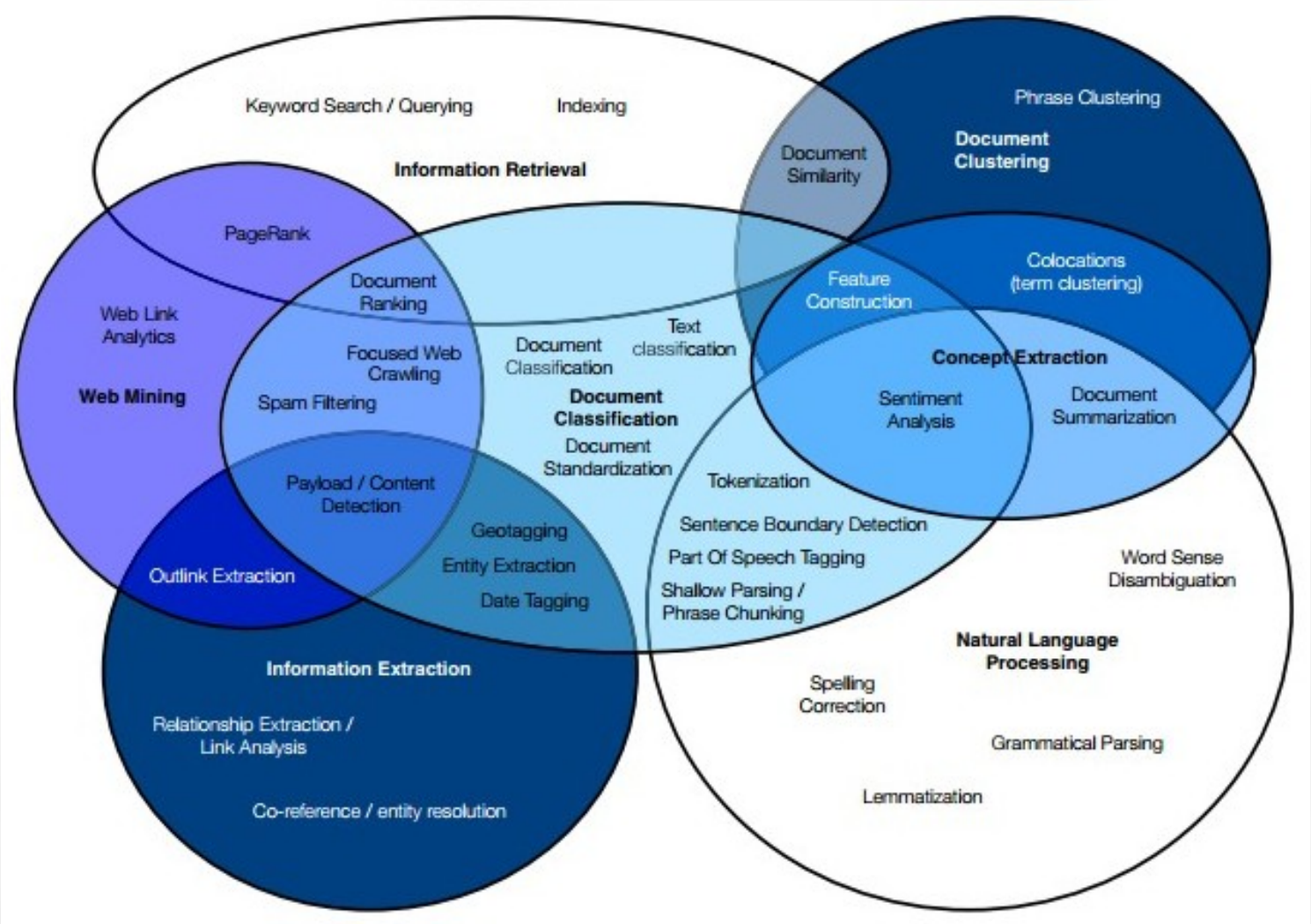
Saco de palavras (*baf-of-words*)



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

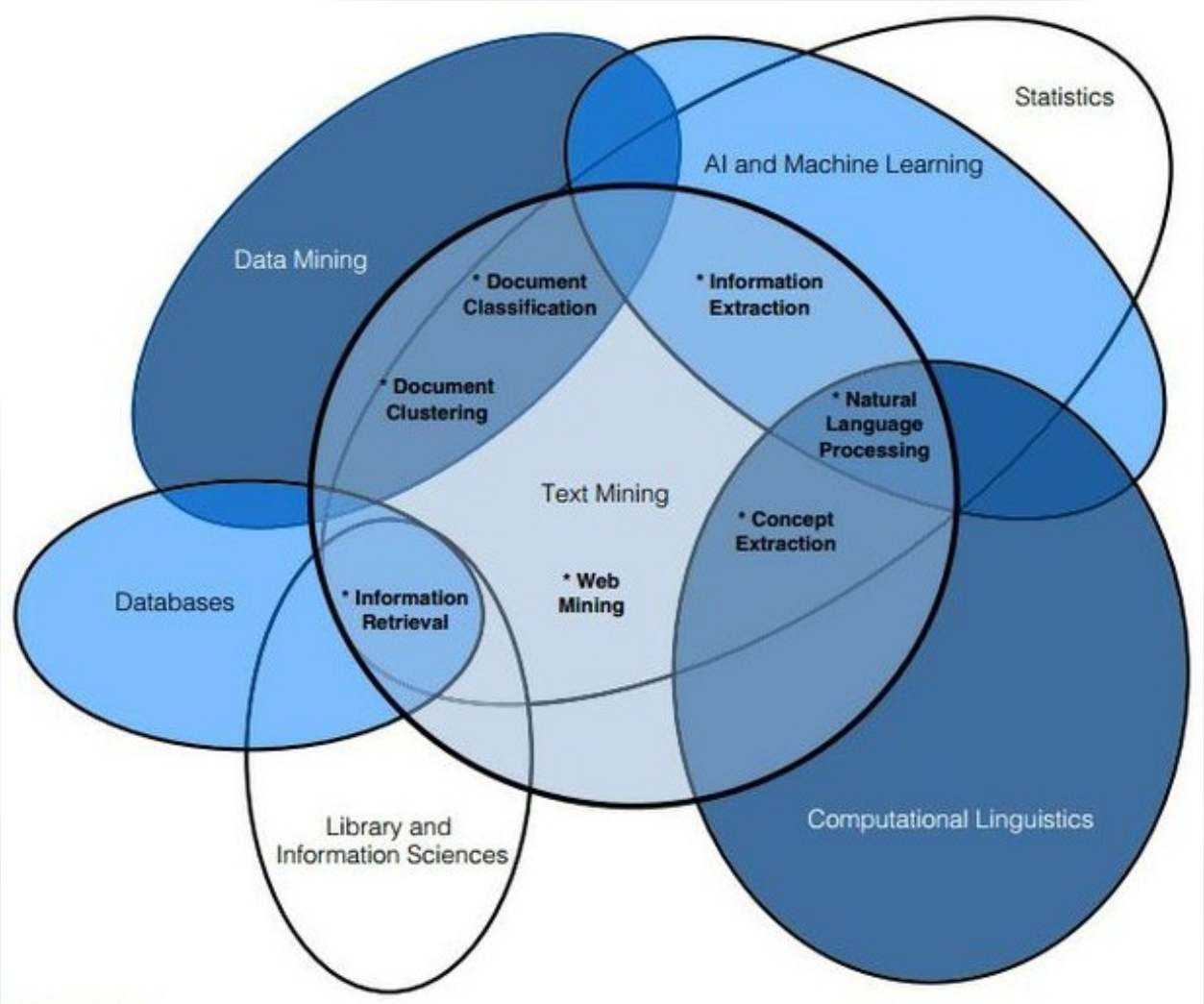
Áreas e disciplinas relacionadas

As 7 áreas da mineração de texto



Visão por tarefas (MINER et al., 2012).

Disciplinas relacionadas



Visão por disciplinas (MINER et al., 2012).

Ferramentas de mineração de texto

Ferramentas online

1. https://www.paperrater.com/.
2. http://www.articlegeneratorpro.com/.
3. http://articlegenerator.org.
4. http://parts-of-speech.info/.
5. https://iwl.me.
6. http://textalyser.net/.

Softwares comerciais

1. STATISTICA Text Miner.
2. SAS Text Miner.
3. Clarabridge.
4. IBM SPSS Text Analytics.
5. IBM News Explorer.

Mais em list of text mining software.

Recursos no R

Task Views relevantes

- Natural Language Processing.
- Web Technologies and Services.

Pacotes R

Text mining	Web scraping
tm, Rweka	XML, xml2
topicmodels, lsa	RCurl, httr
text2vec	rvest
tokenizers, udpipe	jsonlite
NLP, openNPL	twitteR
koRpus, lexiconPT	Rfacebook
RTextTools, tidytext	Rlinkedin

Complemento importantíssimo

Web scraping

Como extrair ou consumir dados da Web?

- XML
- HTML
- JSON

Referências

HURWITZ, J.; NUGENT, A.; DR. HALPER, F.; KAUFMAN, M. **Big data para leigos:** ALTA BOOKS, 2016.

MINER, G.; ELDER, J.; HILL, T. **Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications**. Academic Press, 2012.