Mineração de Texto

Visão geral das tarefas e métodos

Prof. Walmes Marques Zeviani

JUSTIÇA 4.0: INOVAÇÃO E EFETIVIDADE NA REALIZAÇÃO DA JUSTIÇA PARA TODOS
PROJETO DE EXECUÇÃO NACIONAL BRA/20/015

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, Nugent, Dr. Halper, and Kaufman (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

- Quando lemos um livro, recordamos das sensações mas não da prosa.
- Tratamos a informação de texto na sociedade assim também.
- Somos sensores sobre o mundo e registramos o que percebemos com texto.
- 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/.

Dados de texto são abundantes

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.

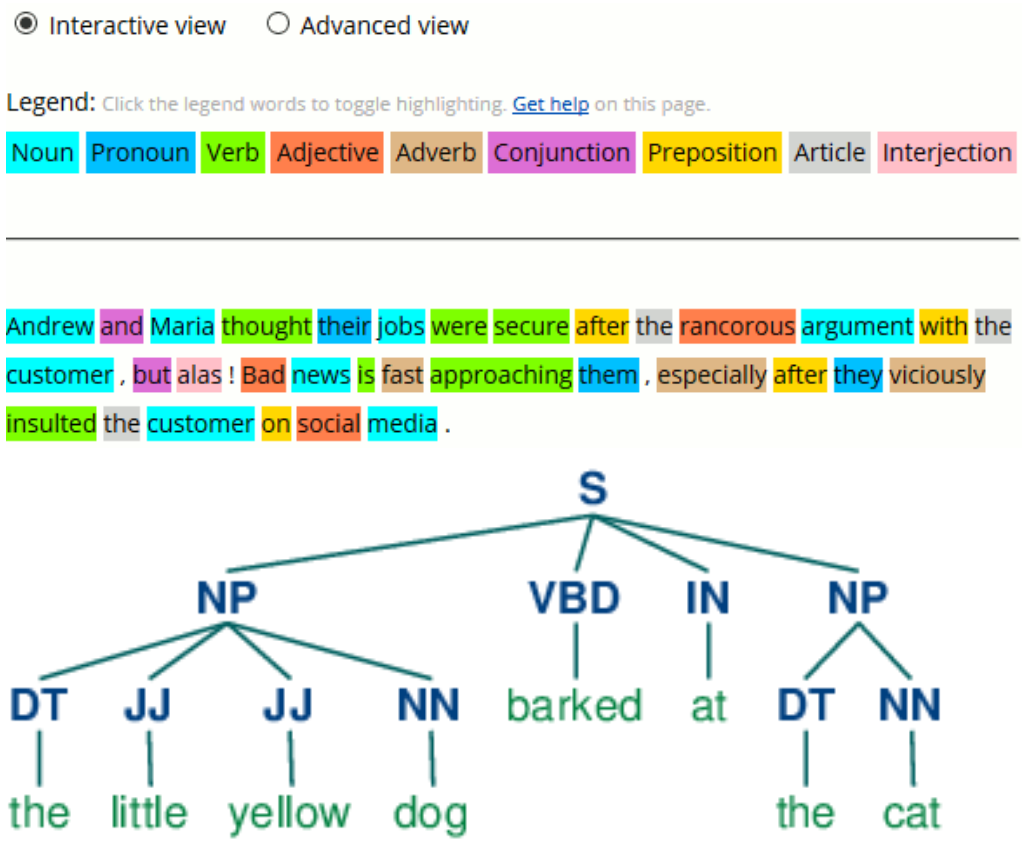
Tipos de formato

Dados não estruturados = estrutura imprevisível.

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

Abordagens principais · Análise sintática

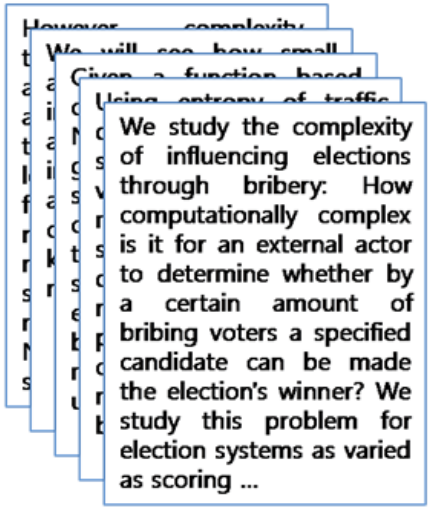
- Análise lexical/morfológica: formas da palavra.
- Análise sintática: estrutura gramátical, criar contexto.
- Análise semântica: determinar significado, eliminar ambiguidades.
- Análise do âmbito do discuso: significado além do discurso, inferência.
- É uma análise complexa que pode determinar: quem,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



Abordagens principais · Saco de palavras

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

Documents



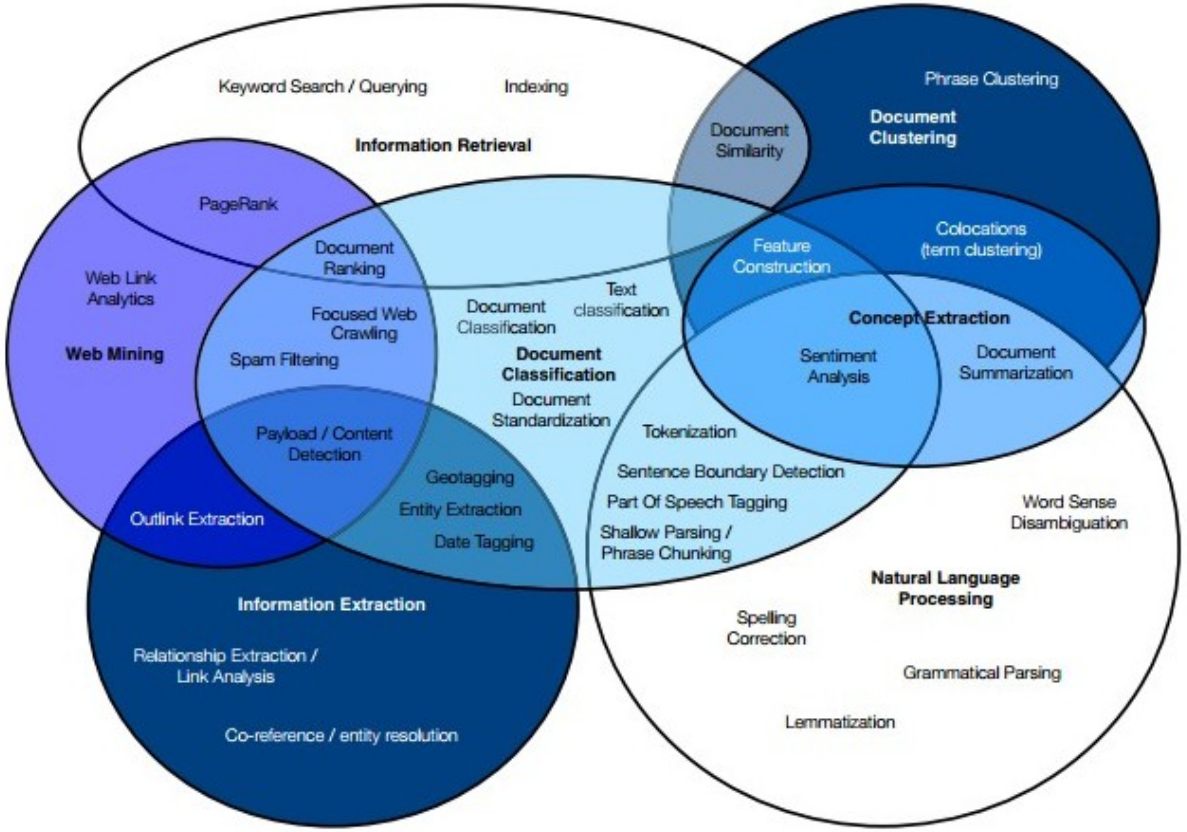
Vector-space representation

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

Term-document matrix

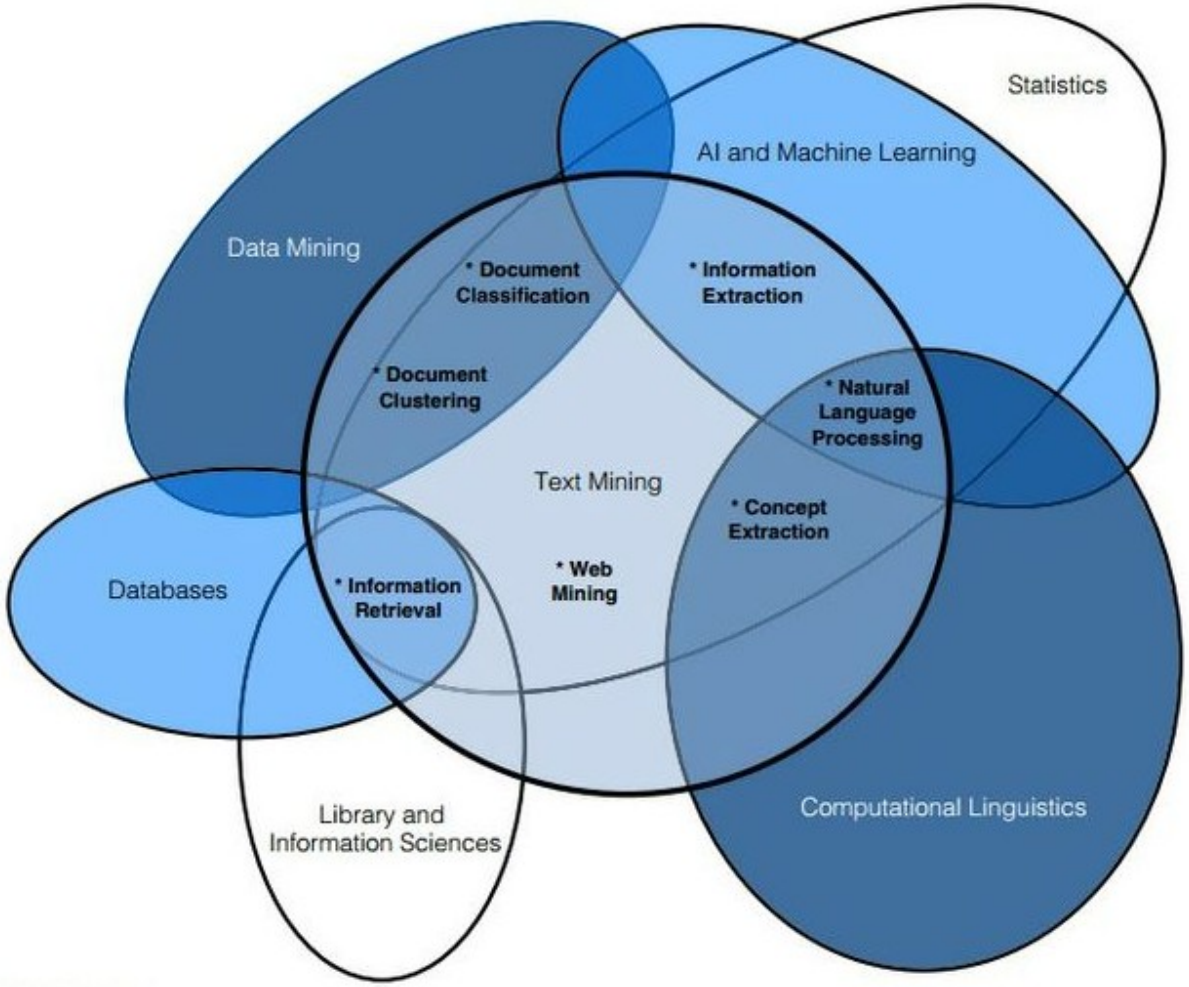
Áreas e disciplinas relacionadas

As 7 áreas da mineração de texto



Miner, Elder, and Hill (2012)

Disciplinas relacionadas



Miner, Elder, and Hill (2012)

Ferramentas de mineração de texto

Ferramentas online

- https://www.paperrater.com/.
- http://www.articlegeneratorpro.com/.
- http://articlegenerator.org.
- http://parts-of-speech.info/.
- https://iwl.me.
- http://textalyser.net/.

IBM Watson News Explorer

http://news-explorer.mybluemix.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.

O que vamos usar

Recursos no R

Task Views

- Natural Language Processing
- Web Technologies and Services*

Pacotes R

Text mining	Web scraping*	Outros*
tm	XML	d3Network
Rweka	rvest	leafletR
SnowballC	RCurl	googleVis
wordcloud	jsonlite	lattice
topicmodels	twitteR	latticeExtra
RTextTools	Rfacebook	ggplot2
lsa	Rlinkedin	
openNPL		
koRpus		
tidytext		

Referências

Hurwitz, J., A. Nugent, F. Dr. Halper, et al. (2016). *Big Data Para Leigos:* . ALTA BOOKS. ISBN: 9788576089551. URL: https://books.google.com.br/books?id=j8hYCwAAQBAJ.

Miner, G., J. Elder, and T. Hill (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press. ISBN: 9780123869791. URL: https://books.google.com.br/books?id=-B6amxqygTMC.

Mineração de Texto

Criação e Processamento de um Corpus

Prof. Walmes Marques Zeviani

JUSTIÇA 4.0: INOVAÇÃO E EFETIVIDADE NA REALIZAÇÃO DA JUSTIÇA PARA TODOS
PROJETO DE EXECUÇÃO NACIONAL BRA/20/015

Justificativa

- Em mineração de texto, o preprocessamento requer muito esforço/tempo.
- Bom preprocessamento contribui em muitro para o sucesso da análise.
- É importante conhecer as técnicas e sabem como aplicá-las.

Objetivos

- Descrever as principais etapas de processamento de texto.
- Criar e processar um Corpus.
- Descrever classes de objetos e métodos.
- Construir matriz de documentos e termos.
- Gerar algumas estatísticas e visualizações.

Etapas de preprocessamento

Padronização de caixa

- Elimina variações de escrita das palavras.
- $\{\text{Caio}, \text{UFPR}, \text{USA}, \text{latticeExtra}\} \rightarrow \{\text{caio}, \text{ufpr}, \text{usa}, \text{latticeextra}\}$.

Remoção de pontuação

- Retira os sinais de pontuação.
- Ex: `. , : - / ! * ( ) [ ] { } + ~ $ &`, etc.
- $\{\text{vende-se}, \text{olá}!, \text{o que}?\} \rightarrow \{\text{vende se}, \text{olá}, \text{o que}\}$

Elimina espaços

- Apara os espaços em branco nas extremidades.
- Elimina o excesso de espaços entre palavras.
- Substitui tabulações e similares por espaços.

Remoção dos números

- Elimina caracteres numéricos: 0 a 9.
- Ex: Bebeu 2 xícaras de café às 14h00 por R$ 4,50. → Bebeu xícaras de café às h por R$,.

Remoção das *stopwords*

- Elimina palavras que são muito frequentes.
- São palavras de ligação e sem contexto específico.
- Principalmente: artigos, preposições, conjunções e pronomes.
- Ex: o, a, os, as, com, sem, aos, mas, portanto, eu, ela, nós.
- E também verbos predominantes: ser, ter, haver.
- https://gist.github.com/alopes/5358189.

Desbaste de afixos de extremidade

- Elimina prefixos e sufixos.
- Reduz variações de gênero, número, tempo, etc.
- A tranform**ação** do materia**l** em produ**to** fin**al** pela indústria. $\rightarrow$ A transform do materia em produ fin pela indústria.

Remoção de caracteres acentuados

- Caracteres acentuados são comuns em português.
- Ex: ç ã â à á é ê í ó õ ú.
- A refeição foi macarrão à carbonara. → A refeicao foi macarrao a carbonara.

O que as etapas têm em comum?

- Diminuir o número de variações gráficas.
- Eliminar termos frequentes que não são de contexto específico.
- Tudo isso **faz reduz dimensionalidade**.

Vamos praticar