

Mineração de Texto

Visão panorâmica

Prof. Walmes M. Zeviani

walmes@ufpr.br

Laboratório de Estatística e Geoinformação

Departamento de Estatística

Universidade Federal do Paraná



Humanos como sensores do mundo



Figura 1. Dispositivos e humanos como sensores do mundo. Fonte: o autor.

Estrutura

- ▶ **Estruturados:** Numéricos, cronológicos, espaciais, categóricos.
- ▶ **Não estruturados:** Documentos, correspondências, jornais, blogs, imagens, áudios, vídeos.
- ▶ **Semi estruturados:** Arquivos de registro, notas fiscais eletrônicas, tweets, emails.

Dados na forma de texto

- ▶ A maioria dos dados é não estruturado.
- ▶ Crescem em abundância e variedade.
 - ▶ Tweets, laudos médicos, processos judiciais, avaliações de produtos, reclamações de clientes, transcrições de conversas.
 - ▶ Hashtags, marcações de usuários, localização, emojis, nota.
 - ▶ Linguagem coloquial, erros de grafia, siglas, ironia, negação.
- ▶ Contém informação de valor estratégico.
 - ▶ Expressam opinião, dados, fatos e ações.
 - ▶ Para indivíduos, governos, empresas e ciência.
 - ▶ Implicações no comércio, indústria, saúde, política, ciência e tecnologia.



Mineração de texto



Definição

Mineração ou análise de texto é o processo de extrair informação transformá-la de forma que possa ser aproveitada ou consumida.

Tecnologias

- ▶ Emerge do contato de disciplinas consideradas distantes.
 - ▶ Estatística.
 - ▶ Linguística computacional.
 - ▶ Ciência da computação.
- ▶ Abordagens com diferentes níveis de maturidade.
 - ▶ Algumas metodologias vem estabelecidas.
 - ▶ Várias em fase de descoberta e experimental.
 - ▶ Soluções impactantes com tanto com abordagens simples quanto complexas.
 - ▶ Em comum: todas são centradas em **transformar texto em números** para análise.
- ▶ Diferentes níveis de acesso para o praticante.
 - ▶ Disponibilidade de software.
 - ▶ Acesso aos detalhes da metodologia nos elementos computacionais, linguísticos e estatísticos.
 - ▶ **Importante:** idioma.



Disciplinas e áreas práticas da mineração de texto

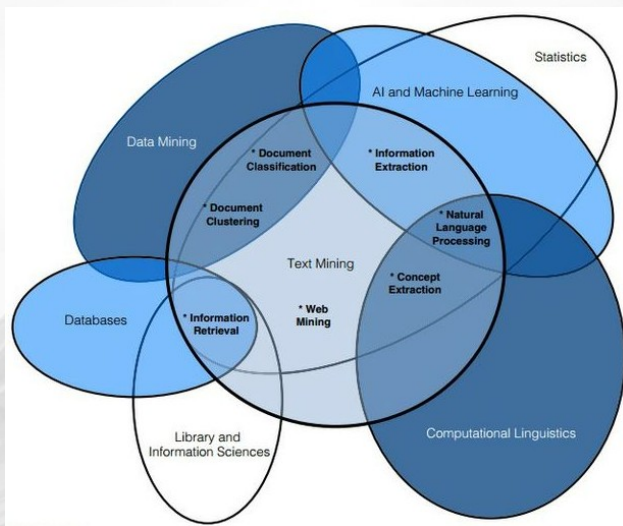


Figura 2. Disciplinas relacionadas com as 7 áreas da mineração de texto (??).

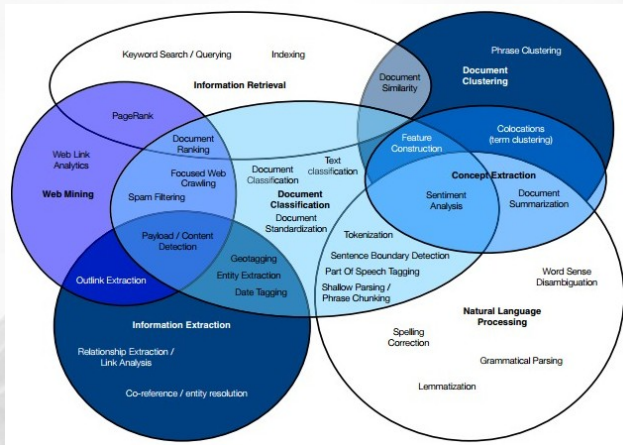


Figura 3. Áreas de mineração de texto e suas principais temas e ferramentas (??).

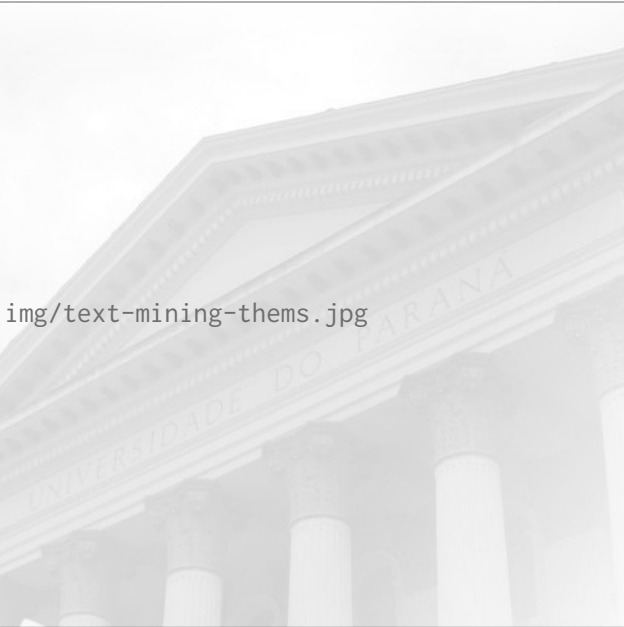


Figura 4. Áreas de mineração de texto e suas principais temas e ferramentas (??).



Abordagens principais em mineração de texto

As 2 abordagens principais

Bag of Words

- ▶ Baseada nos termos que o documento contém (e.g. palavras).
- ▶ Termo é uma característica.
- ▶ Não depende da estrutura ou ordem de escrita.
- ▶ Quase sempre é idioma agnóstico.
- ▶ Requer etapas de pré-processamento.
- ▶ Engenharia de características é fundamental (remoção de redundância, redução de dimensão, representação).
- ▶ Abordagens bem simples e escaláveis em termos estatísticos e computacionais.

As 2 abordagens principais

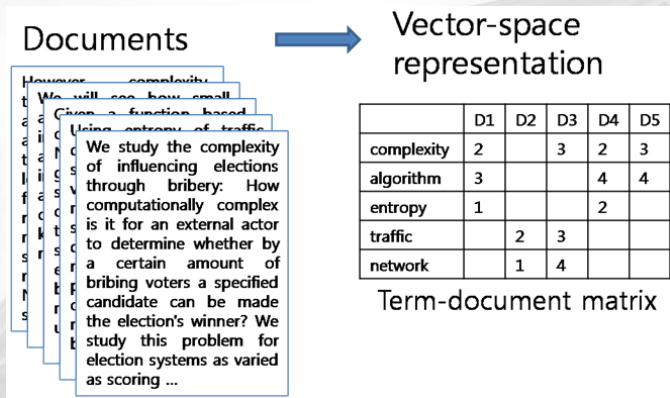


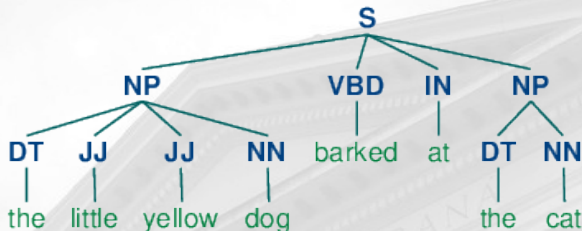
Figura 5. Matriz de documentos e termos gerada a partir da fragmentação do documento em palavras.

As 2 abordagens principais

Natural Language Processing

- ▶ Voltada a extração de significado do texto.
- ▶ Envolve conceitos linguísticos, estrutura gramatical.
- ▶ Quase sempre depende do idioma.
- ▶ Determina: quem, com quem, quando, onde, como e por quê.
- ▶ Bastante complexo.
- ▶ Chat bots, transcrição de áudio, assistente de voz.
- ▶ Uma das áreas que mais cresceu com o deep learning.

As 2 abordagens principais



Four score and seven years ago **our fathers** brought forth on this continent a **new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.**

Now **we are engaged** in a great **civil war**, testing whether that **nation**, or any **nation** so **conceived** and so **dedicated**, can long **endure**. **We are met** on a great **battlefield** of that **war**. **We have come to dedicate** a portion of that **field**, as a **final resting place** for those who here **gave their lives** that that **nation** might **live**. It is altogether fitting and proper that **we** should do this.

But, in a larger sense, **we can not dedicate**, **we can not consecrate**, we can not **hallow** this **ground**. The **brave men, living and dead**, who **struggled** here, have **consecrated** it, far above our **poor power** to add or detract. The world will little note, nor long **remember** what **we** say here, but it can **never forget** what **they** did here. It is for **us the living**, rather, to be **dedicated** here to the **unfinished work** which **they who fought** here have thus far so **nobly advanced**. It is rather for **us** to be here **dedicated** to the great task remaining before **us**—that from **these honored dead** **we take increased devotion** to that **cause** for which **they gave** the last full measure of **devotion**—that **we** here highly **resolve** that these **dead shall not have died in vain**—that this **nation, under God**, shall have a new **birth of freedom**—and that **government of the people, by the people, for the people**, shall not **perish** from the earth.

Figura 6. Elementos de análise baseada em processamento natural da linguagem.

Níveis de análise

- ▶ Lexical/morfológica.
 - ▶ Grafia e morfologia dos termos.
 - ▶ Termos isolados.
- ▶ Sintática.
 - ▶ Combinação dos termos.
 - ▶ Estrutural e coletiva.
- ▶ Semântica.
 - ▶ Determinação de significado.
 - ▶ Estrutura linguística de contexto.
- ▶ No âmbito do discurso.
 - ▶ Com elementos específicos de contexto.



O que torna as coisas difíceis

Problemas linguísticos

- ▶ Homônimos.
 - ▶ Mesma grafia com significado dependente do contexto.
 - ▶ *Ele pediu a conta ao garçom* vs *Ele não conta a ninguém os erros que comete.*
- ▶ Sinônimos.
 - ▶ Grafias diferentes mas significado igual/similar.
 - ▶ *Casa com 3 dormitórios* vs *Residência com 3 quartos.*
- ▶ Ambiguidade
 - ▶ *Ele sentou na cadeira e quebrou o braço.*
 - ▶ *Ana encontrou o gerente da loja com o seu irmão.*
- ▶ Hiperonímia e hiponímia
 - ▶ Frutas: maçã, laranja, limão, melancia.
 - ▶ Eletrodomésticos: geladeira, micro-ondas, forno elétrico.
- ▶ Homófonos.
 - ▶ *Não soube me dizer onde era o acento.*
 - ▶ *Não soube me dizer onde era o assento.*



Ferramentas e Softwares

Softwares comerciais

- ▶ STATISTICA Text Miner.
- ▶ SAS Text Miner
- ▶ IBM SPSS Text Analytics .
- ▶ Clarabridge.

Mais em https://en.wikipedia.org/wiki/List_of_text_mining_software.

Softwares livres

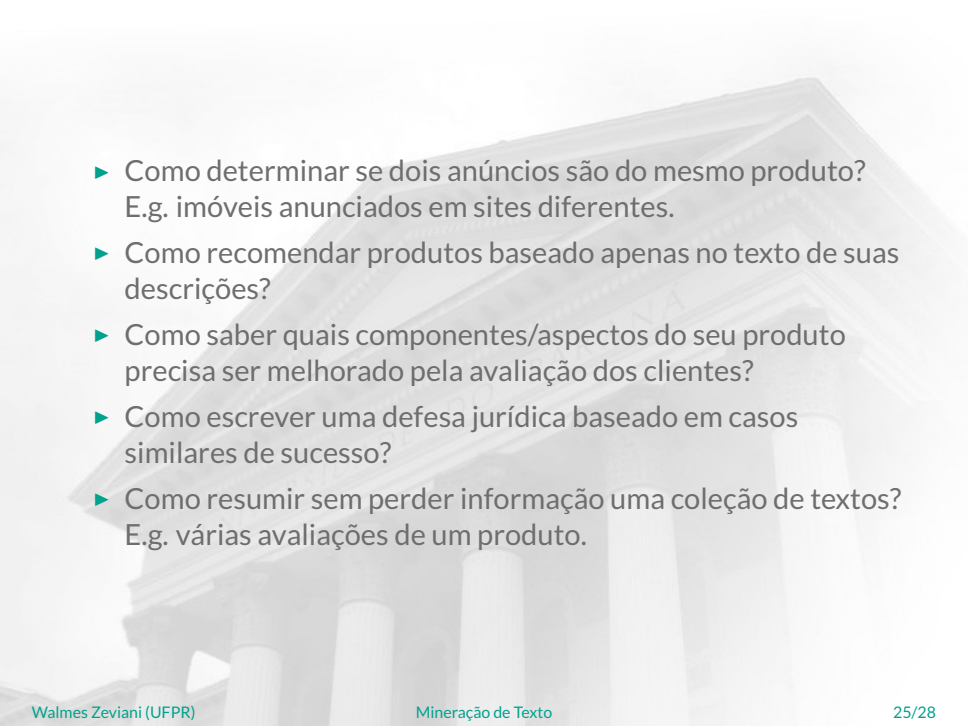
- ▶ R: pacote `tm` e *task view* de Natural Language Processing.
- ▶ Python: Natural Language Toolkit.
- ▶ Weka: Text mining in Weka cookbook.

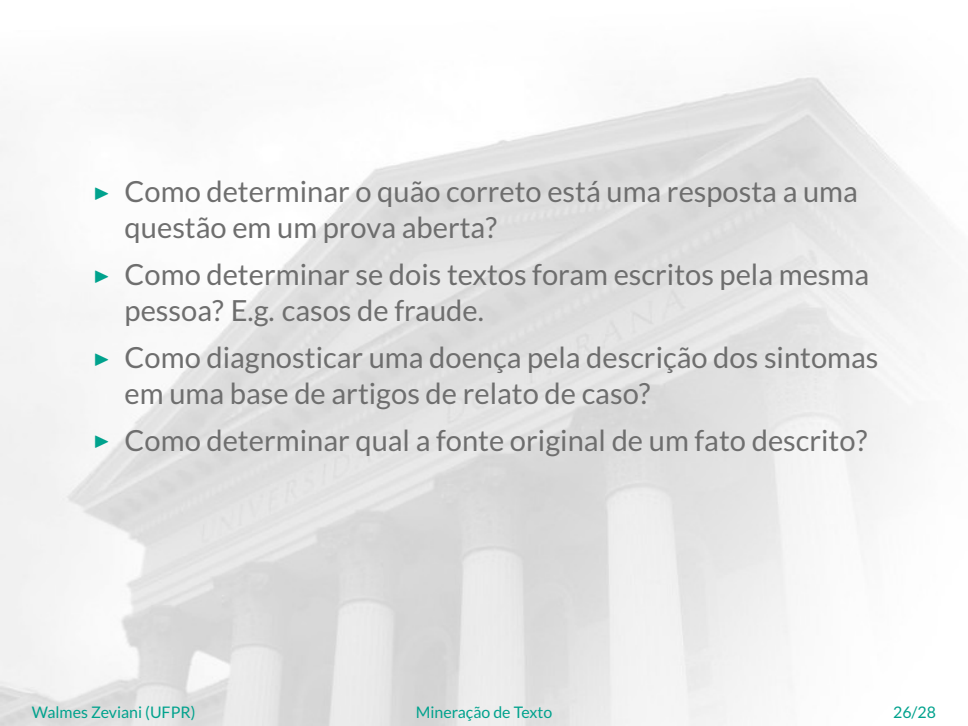
Ferramentas online

- ▶ <http://news-explorer.mybluemix.net/>.
- ▶ <https://www.paperrater.com/>.
- ▶ <http://www.articlegeneratorpro.com/>.
- ▶ <http://articlegenerator.org>.
- ▶ <http://parts-of-speech.info/>.
- ▶ <https://iwl.me>.
- ▶ <http://textalyser.net/>.



Alguns problemas práticos interessantes

- 
- ▶ Como determinar se dois anúncios são do mesmo produto?
E.g. imóveis anunciados em sites diferentes.
 - ▶ Como recomendar produtos baseado apenas no texto de suas descrições?
 - ▶ Como saber quais componentes/aspectos do seu produto precisa ser melhorado pela avaliação dos clientes?
 - ▶ Como escrever uma defesa jurídica baseado em casos similares de sucesso?
 - ▶ Como resumir sem perder informação uma coleção de textos?
E.g. várias avaliações de um produto.

- 
- ▶ Como determinar o quão correto está uma resposta a uma questão em um prova aberta?
 - ▶ Como determinar se dois textos foram escritos pela mesma pessoa? E.g. casos de fraude.
 - ▶ Como diagnosticar uma doença pela descrição dos sintomas em uma base de artigos de relato de caso?
 - ▶ Como determinar qual a fonte original de um fato descrito?



Complementos, dúvidas, reclamações?

Obrigado!

Referências bibliográficas

