

Ciência de Dados e Big Data

Recuperação da Informação na Web e em Redes Sociais

PUC-Minas IEC | Pós-Graduação Lato Sensu

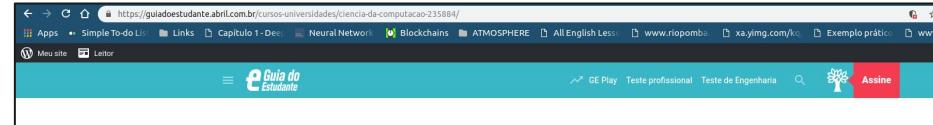
Zilton Cordeiro Jr.

Um pouco sobre a minha Formação...



Formação

- ★ Bacharel em Ciência da Computação: 2003 - 2008
 - Universidade Estadual de Santa Cruz - BA



The screenshot shows a web browser displaying the 'Guia do Estudante' website. The page is for the course 'Ciência da Computação' at 'Universidade Estadual de Santa Cruz - Uesc'. It includes sections for 'Sobre o curso' (Course Overview), 'Sobre a instituição' (About the Institution), and various statistics like 'Inscritos' (Registrants) and 'Vagas no último processo seletivo' (Places in the last selection process). The URL in the address bar is <https://guiadoestudante.abril.com.br/cursos-universidades/ciencia-da-computacao-235884/>.

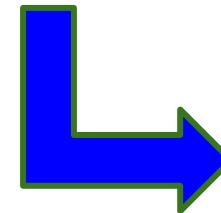




Formação

★ Bacharel em Ciência da Computação

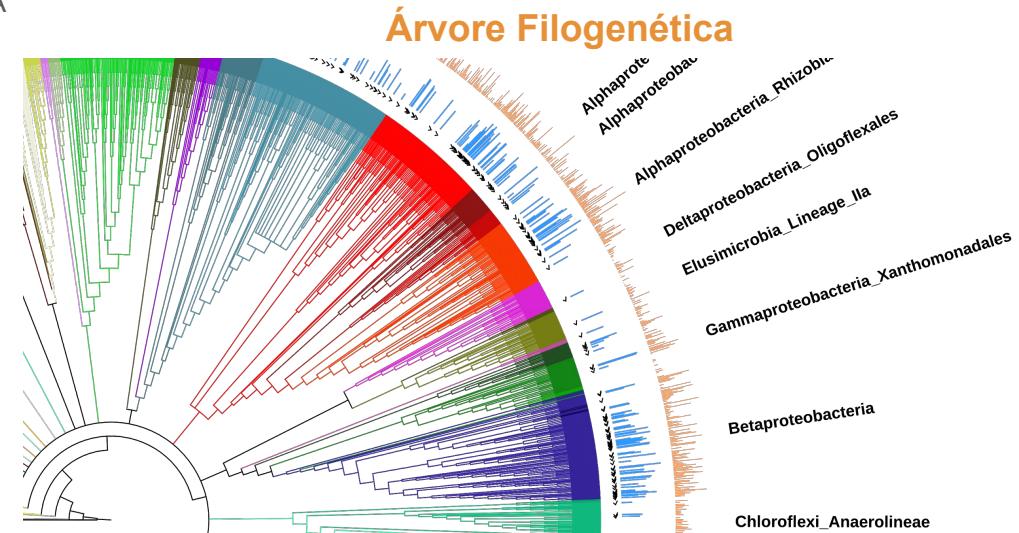
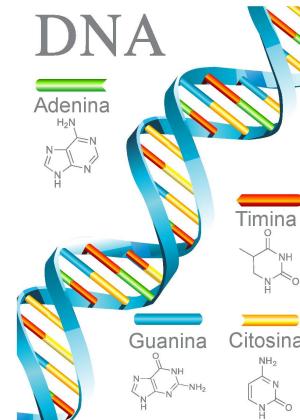
- Universidade Estadual de Santa Cruz - BA
- Iniciação Científica: Bolsista CNPq (2004 - 2008)



Formação

★ Bacharel em Ciência da Computação

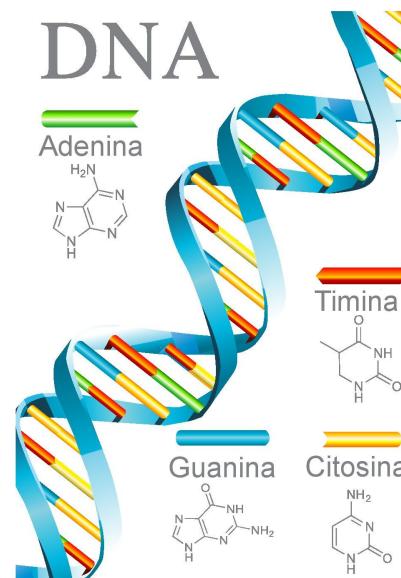
- Universidade Estadual de Santa Cruz - BA



Formação

★ Bacharel em Ciência da Computação

- Universidade Estadual de Santa Cruz - BA



GAATTCGCGTCCGCTTACCATGTGCCGTGGATGCCAACAGGAGGCC
CGTTGACGGCGAATGACTTACTCAAGGGAGTAGCCAATCTGCGGATACG
CCCGGATTGGAGCTGCCATGGAGGGCTACAAGAAAGCGGTGGAGGAT
TGCTCGCATACTGCAGAGACCGTTCTGAAGGAGATGGCTATGGAGTACC
TGCCTACGCTTGTGCCGCCGAGAAGTGGTAAGAAGAACGGAGCCCATA
CACCAAGGTGATATGGTCTTCGTCTGGATCCGCCCTGCCCCGGGAGA
GTGGTCAAGGCATCATGGAGGAAGTCTCCAGCAGAGCAGATGGAGCAA
CGGCCCTATAGAGGACACTGATGCTACCCGTCTAAAGCTTGCAGTTTG
ATTTAAGTGAATCGTTATTCA CGGGGTCGGGATGTCGGGATCGAACG
GTGCAATCGATAGGCGTAATCAGTATTCCAGATAGTGTATAAGATTGGT
GGATAAAATGTGCGGGCACACTAATGGCCGCATCGTAAGCCGCAAAA
GCTTAGCGTCATTGTCATCGAGAGTTGGAGGGCAAAGTGGGTAAGA
TAAGATTAATAATTGTACTGAATAATCTAAAGAATCTGTATGGAAAG
CGCCATGCAGTCACATATAATATGTCAGAGCTCCTCTCCATGCAGT
AGAACGACAGAGTTACTGTCATGCCAATCTGTGCAAATGGCTG
TGAAGTTGCAATGCCCGTAAGAAGGCCAGTCAAAATGATTATATTGCG
GATGCGAATCTTAGACTGCTAAAGTCTGGTAGGTGTTCCAGAAGGACA
GACGCTTTCTTCTCTGGACTTTAGTAAGCCCGTAATCTGTGCTG
CCAACACCACAGAATGGTCGGGCCAATTAGAGGGTCTCTGCCCTT
CCTGGCTAGGTTGTCGGCTAGCTATTCCGGATGTTGTTGTC
GGACCCACCTATTGTGACTTGTGACAGCTCCAAGTTGCTAGTGC
GTCCTTACTTCATTTACTAGCTTGTAGATTTATCTTGTAGTT
CTTCATGGCTGCTTGAATCAGACAGTATGCAATGTCCTGCCATGAT
AGTTCCCTTTAGATTAAACTCTGCACAGCGTCCAATAGCAGACACTTC
GCTTGAATGCTGGTGTATCTGCCATTGATTGCTGGTATTTCAACCTGG
GCCCACTTCCCTCGCGTGAGGGATCCGTCTGTATACCACTTTATTGTT
TGTGGTTTCATAGGGTGTACTTGGCCAGGGATGTCACACTTTTATT

Formação

- ★ Mestre em Ciência da Computação: 2009 - 2011

 - Universidade Federal de Minas Gerais



Vida Profissional

2011



Instituto Nacional de Ciência e Tecnologia para a Web

observatório da web

GALILEU

HOME NOTÍCIAS GALERIAS VÍDEOS BLOGS COLUNISTAS REVISTA

Fobias podem ser genéticas

"A violência é uma doença contagiosa."

comente | envie por e-mail | comarilhe | imprima

tamanho do texto AA

De olho na web

Curtir 321

Tweetar

G+

Pesquisadores brasileiros criam ferramenta para vasculhar mensagens no Twitter, Orkut e Facebook. Acredite, isso pode salvar vidas

por Alexandre Rodrigues



Detective da web: o cientista da computação Wagner Meira cria algoritmos para investigar o que se diz na internet sobre temas como futebol e eleições

Vida Profissional

observatório da web



Vida Profissional

Coleta, Extração e Tratamento de Dados em tempo real



Acompanhar a dinâmica e a interação das pessoas

Prever tendências na sociedade

Orientar organizações que utilizem ou interajam por meio
da internet



Vida Profissional

observatório da dengue

observatório do brasileirão

observatório da copa

2010

**observatório
olímpico**

2012

**observatório
das eleições**

(Presidencial)

**observatório
das eleições**

(Municipal)

Vida Profissional



2013

observatório da web
#automóveis



observatório do **Investimento**

2014

2015



Vida Profissional

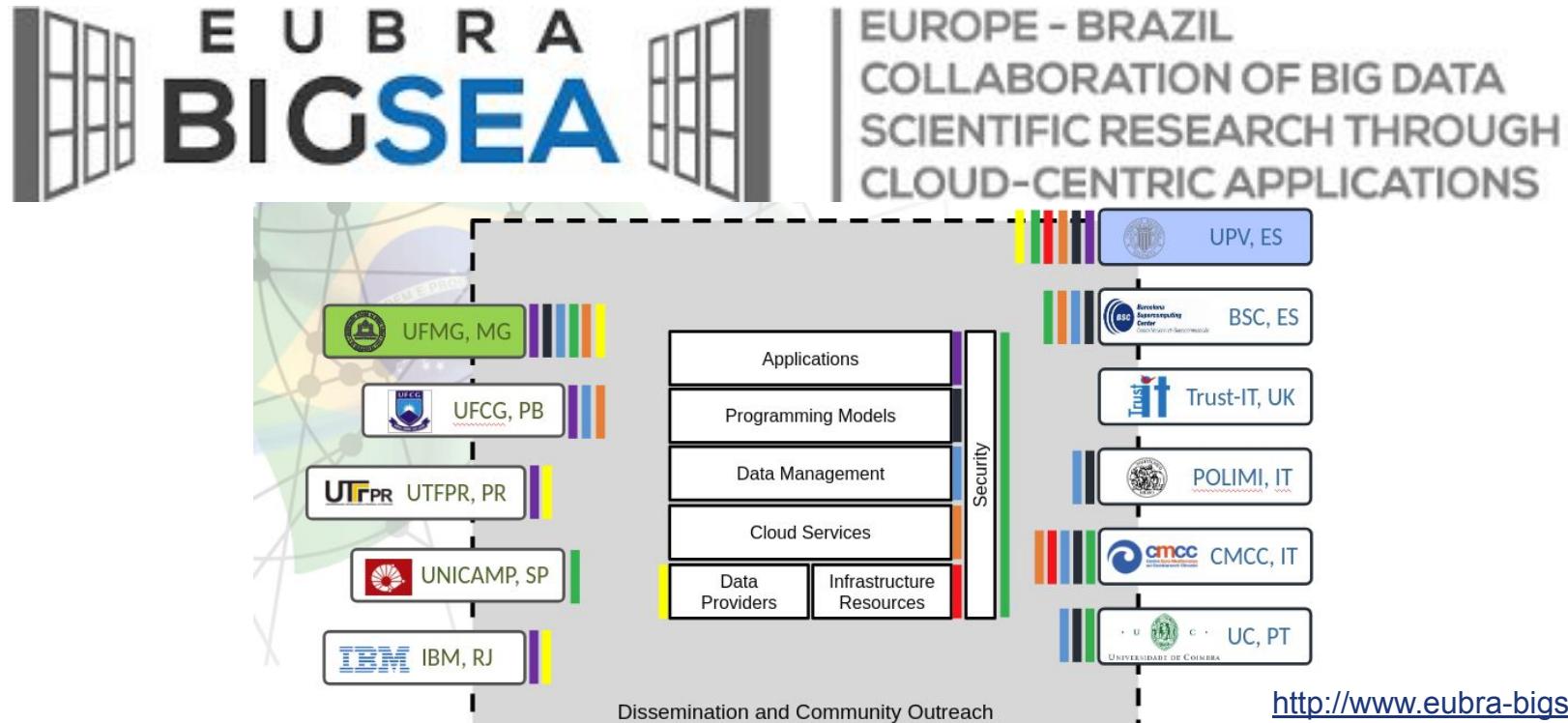


RECORD
IN PROCESS SIGNALS



Vida Profissional

- 2016 - 2018
 - Gerente de Projetos
 - Cientista de Dados



Vida Profissional



Lemonade



LEMONADE - Live Exploration And Mining Of A Non-Trivial Amount Of Data From Everywhere

HOME / TECHNOLOGY / TOOLBOX OF DESCRIPTIVE AND PREDICTIVE MODELS



EUBRA BIGSEA
EUROPE - BRAZIL
COLLABORATION OF BIG DATA
SCIENTIFIC RESEARCH THROUGH
CLOUD-CENTRIC APPLICATIONS

About Features, Benefits & Downloads Users

Lemonade (Live Exploration and Mining Of a Non-trivial Amount of Data from Everywhere) is an analytics platform that supports intuitive definition of tasks for knowledge discovery, mining, and learning from large amounts of data that come from a wide spectrum of scenarios. The platform interface is a web application in which users may define analytics workflows visually by dragging and dropping operations and data sources, and connecting them. Lemonade is being developed by UFMG as part of the EUBra-BIGSEA project and targets users who do not want to learn a programming language, but need to develop analytics workflows. It supports the creation of a processing workflow, import, export or management of datasets, executing and managing workflows and the data visualisation.

Next >

Vida Profissional

- 2018 - 2021
 - Gerente de Projetos
 - Cientista de Dados

ATMOSPHERE

Adaptive, Trustworthy, Manageable, Orchestrated, Secure Privacy-assuring Hybrid, Ecosystem for REsilient Cloud Computing



KUNUMI

UFMG



UnB



UNICAMP



UFAM



SUPRA OMNES LUX LUCIS



UNIVERSIDADE DE
COIMBRA



Trust-IT Services
Communicating ICT to markets



Πανεπιστήμιο Πειραιώς
University of Piraeus



POLITECNICO
MILANO 1863

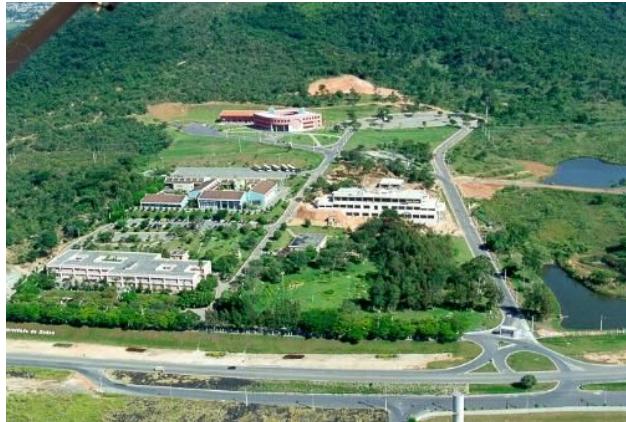


TECHNISCHE
UNIVERSITÄT
DRESDEN



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Vida Profissional (Professor)



2012 - 2017

- Graduação:
 - Ciência da Computação
 - Engenharia Civil



2014 - 2016

- Curso de Formação de Oficiais



2018...

- Pós-graduação

Informações sobre a disciplina

Recuperação da Informação na Web e em Redes Sociais



A disciplina - RI

❖ Ementa

- Apresentar conceitos relacionados à *Web Crawling*, *Web Mining* e Redes Sociais
- Apresentar algoritmos e ferramentas para coleta e análise de dados extraídos da Web e de Redes Sociais
- Apresentar propriedades e comportamentos de redes complexas

A disciplina - RI

❖ **Plano de Ensino**

- **Unidade 01:** Conceitos de inteligência competitiva e coletiva, crowdsourcing e redes sociais. Recuperação da informação e Máquinas de busca. Desafios da Mineração na web e nas redes. Exemplos de Projetos da disciplina.
- **Unidade 02:** Algoritmos e soluções para problemas de busca e extração de informação da WWW. Ferramenta e prática de processamento textual e recuperação de informação.
- **Unidade 03:** Tipos de coleta, arquitetura e componentes de coletores Web. Ferramenta e prática de coleta de dados na Web.

A disciplina - RI

❖ **Plano de Ensino**

- **Unidade 04:** Aprofundando na mineração de texto e linguagem natural. Algoritmos e soluções para a análise da informação presente nas redes sociais online e em sites de conteúdo. Ferramenta e prática de mineração de texto.
- **Unidade 05:** Caracterização de redes sociais: Tipologia, características e representações gráficas. Algoritmos estocásticos, análise de redes complexas. Ferramenta e prática de mineração de redes complexas.
- **Unidade 06:** Indexação, Busca e Mineração em plataforma de Big Data.

A disciplina - RI

❖ Teórico e Prático

- O conteúdo estudado será exercitado em práticas utilizando ferramentas de mineração de texto e busca.
- As aulas práticas serão avaliadas e em cada prática uma tarefa deverá ser realizada de maneira autônoma.
 - **40 pontos.**
- O Projeto Final será formado por conceitos discutidos e aplicados nas aulas, com adaptações individuais para um caso de uso real. O resultado das tarefas práticas poderá ser reaproveitado.
 - **60 Pontos**

A disciplina - RI

❖ Teórico e Prático



A disciplina - RI

❖ Projeto Final

- ❖ Consiste em realizar um estudo da Web para um assunto real e de livre escolha.
 - Exemplos: Automóveis, moda, música, imóveis...

❖ Será necessário

- Coletar dados em texto de redes sociais e/ou sites da Web
- Analisar o conteúdo textual obtido e/ou
- Analisar dados de relacionamentos entre usuários (i.e. nas redes)
- Relatório final

❖ Data de Entrega

- Até o 15º dia após a última aula - às 23h59m
 - A forma de entrega será definida em breve

A disciplina - RI

- ❖ Como interagir e acessar os materiais da disciplina?
-

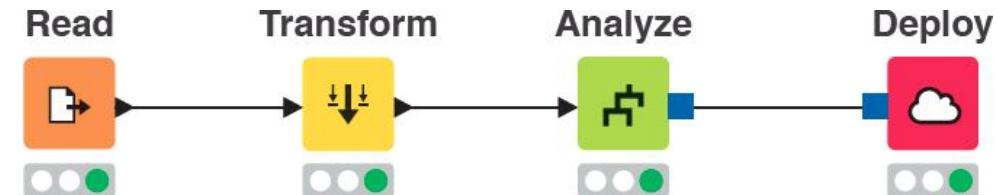
<https://goo.gl/ps88N4>



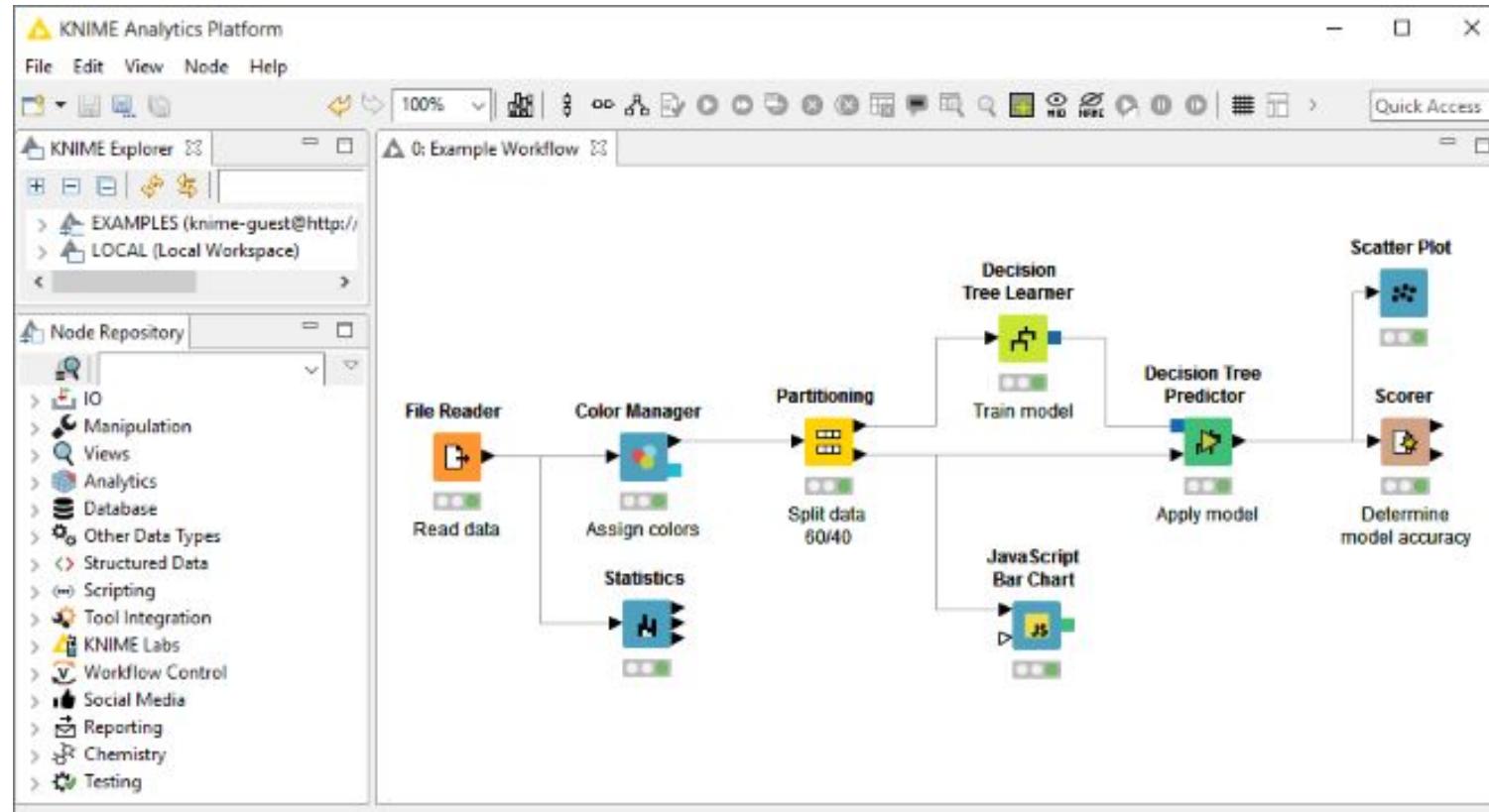
A disciplina - RI

❖ Konstanz Information Miner - KNIME

- Análise de dados ágil com workflows
- Componentes modulares facilitam a criação, modificação e manutenção dos fluxos de análise
- Requer o mínimo de programação (embora programadores possam construir módulos personalizados, incluindo outras linguagens como R, Python, Java)
- Opções de importação/exportação de dados e conexões com outros sistemas



A disciplina - RI



Recuperação da Informação na Web e em Redes Sociais



A Web

❖ Qual o tamanho aproximado da web?

- Bilhões de páginas Web
- Centenas de milhões de tweets diários
- Bilhões de consultas realizadas no Google diariamente
- Milhões de servidores e petabytes de dados



Experimento: busque pelo termo “a” no Google e observe o número de respostas.

A Web

Turbilhão de Dados, Informações e Conhecimento?



Dados

❖ O que são?

- Sucessões de fatos brutos, que **NÃO** foram:
 - Organizados,
 - Processados,
 - Relacionados,
 - Avaliados ou
 - Interpretados
- Representam apenas **partes isoladas de eventos, situações ou ocorrências.**
- É a forma primitiva que compõe os sistemas de informação.

Porém, se...

Dados

❖ ... os dados passarem por algum tipo de

- Relacionamento,
- Avaliação,
- Interpretação ou
- Organização.

Geramos...



Informação

❖ Pode ser definida como:

❖ Um dado acrescido de **contexto, relevância e propósito.**

 Röbs 🇨国旗 - @robstciola - Jan 17
Replying to @carlosed_18 @FiatBR
Da hotwheels

 Diego Alves 🇧🇷 - @DHzZz - Jan 17
Replying to @tbb @FiatBR
Ihh, que isso?
Michael Douglas! 🎭

 Ana Laura Caboclo Ribeiro @RitmoCaboclo - Jan 17
Replying to @tbb @FiatBR
Pippa agradice

 MariPen 🇧🇷 - @Mari_Pen - Jan 17
Replying to @boninho @FiatBR @tbb
Olike omic

 Hellen 🌸 - @hellenbeattriz - Jan 17
Replying to @tbb @FiatBR
17h baby

 Filipa Ferreira 🇵🇹 - @FilipaPSF - Jan 17
Replying to @boninho @FiatBR @tbb
Vol little fish ❤️

 andy XIII @andy229y - Jan 17
Replying to @boninho @FiatBR @tbb
amo as winners

 Jubs 🎉 - @jubscamemou - Jan 17
Replying to @boninho @FiatBR @tbb
GIRL POWERRR

 lingua preta @Fabcia777 - Jan 17
Replying to @CaraComentava @tbb and 2 others
Uma chatec 😢

 Felipe Grote 🎉 - @FelipeGrote - Jan 17
Replying to @FiatBR @tbb
BBB = Ito.

 Flá_Diretoria @Cambimba_B1 - Jan 17
Replying to @FiatBR @tbb
Block.

 Mendigo mil grau @Kcnwar - Jan 17
Replying to @FiatBR @tbb
Sal do BBB Fiat

Ferraz #34 FC conquistou o Prêmio Fiat Cinqucento da categoria profissional por fazer 500 pontos em 2 turnos <http://t.co/ayTzGRpp>
 Ferraz #34 FC conquistou o Prêmio Fiat Cinqucento da categoria profissional por fazer 500 pontos em 2 turnos <http://t.co/ayTzGRpp>
 SE Peixudas conquistou o prêmio Fiat Freemont na categoria base, por ter 7 jogadores que fizeram 7 pontos na rodada <http://t.co/Ipg9sJV8>
 SE Peixudas conquistou o prêmio Fiat Freemont na categoria base, por ter 7 jogadores que fizeram 7 pontos na rodada <http://t.co/Ipg9sJV8>
 SE Peixudas conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/Hpg9sJV8>
 Botafogo Star conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/knht8Nmp>
 El Mago Branco FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/zslHqFha7>
 El Mago Branco FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/zslHqFha7>
 Los Arranca Toco FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/TbpkuA2X>
 Los Arranca Toco FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/TbpkuA2X>
 Codex Rock conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/6mn87OT>
 D.S.K. F.C conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/XyCnvAxD>
 Conversor FC conquistou o prêmio Fiat Bravo na categoria base, por ter o jogador que mais valorizou na rodada <http://t.co/c7FbE227>
 Fael Nunes Team conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/mv1z5Vt>
 SEP EC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/cpkaiFgg>
 Você conquistou o Prêmio Fiat Bravo da categoria profissional. Por ter o jogador que mais valorizou por 2 rodadas consecutivas.
 SEP EC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/cpkaiFgg>
 Luanique FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/q54F7Q0k>
 Luanique FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/q54F7Q0k>
 Teus Anjinho conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/S5sBGU9e>
 Teus Anjinho conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/S5sBGU9e>
 Sangue no Zojo C.F conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/9mf93AZv>
 Sangue no Zojo C.F conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/9mf93AZv>
 Você conquistou o Prêmio Fiat Cinqucento da categoria profissional. Por ter fazer 500 pontos em 2 turnos.
 Você conquistou o Prêmio Fiat Cinqucento da categoria profissional. Por ter fazer 500 pontos em 2 turnos.
 AgoraOlNuncaFC, conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/NgPMsGH6>
 AgoraOlNuncaFC, conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/NgPMsGH6>
 Galáticos Reis FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos
 Galáticos Reis FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos
 C.A. Linense conquistou o prêmio Fiat Bravo na categoria profissional, por ter o jogador que mais valorizou por 2 rodadas consecutivas
 Mendel Futebol Clube conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/bf1CgwV3>
 Mendel Futebol Clube conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/bf1CgwV3>
 Indiretas FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/d830R3o8>
 Indiretas FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/d830R3o8>
 11 Mitos conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/LbhVxtFH>
 11 Mitos conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/LbhVxtFH>
 Junim S.C conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/OsdthWiz>
 Junim S.C conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/OsdthWiz>
 Sai Capeta FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/oruSICBo>
 Sai Capeta FC conquistou o prêmio Fiat Cinqucento na categoria profissional, por fazer 500 pontos em 2 turnos <http://t.co/oruSICBo>



Informação

-
- ❖ É gerada a partir de uma interpretação sobre os dados, que podem ser:
 - Contextualizados,
 - Categorizados,
 - Calculados ou
 - Condensados.
 - ❖ Advém de fatos sobre uma situação, pessoa ou evento.
 - ❖ Então, **transformar** ou **organizar** os dados em informação com significado permite a execução de algum tipo de **análise**.

Informação

Resumindo...

- ❖ Para transformar dados em informações precisamos que os mesmos sejam:
 - Precisos e oportunos;
 - Específicos e organizados para um propósito;
 - Apresentados dentro de um contexto que lhe dê significado e relevância; e
 - Que podem levar a um aumento na compreensão e diminuição da incerteza.

Informação

❖ A informação é VALIOSA

- Ela pode afetar:
 - o Comportamento,
 - a Decisão ou
 - o Resultado.

❖ Então, a informação, a partir de critérios inicialmente definidos, sob um ponto de vista estratégico e com um caráter informativo obtemos o...



Conhecimento

- ❖ É uma informação **contextual, relevante e açãoável**.

- ❖ É a informação em ação.
- ❖ É uma informação valiosa da mente humana, inclui:
 - **Reflexão**,
 - **Síntese** e
 - **Contexto**.
- ❖ É difícil de estruturar, capturar em computadores, normalmente é tácito e sua transparência é complexa.

Então...

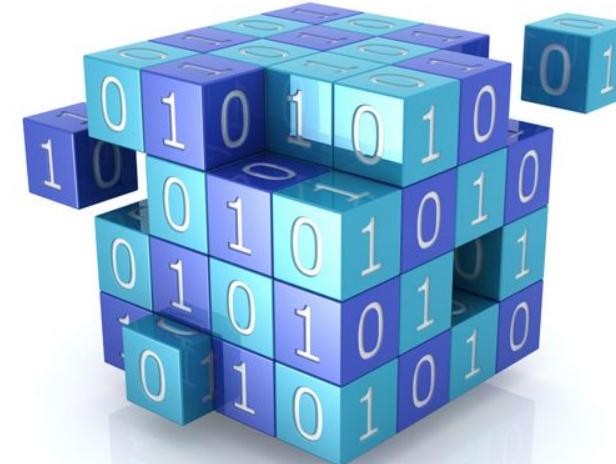
Dados	Informação	Conhecimento
<p>Simples observações sobre o estado do mundo.</p> <ul style="list-style-type: none">• Facilmente estruturado• Facilmente obtido por máquinas• Frequentemente quantificado• Facilmente transferido	<p>Dados dotados de relevância e propósito</p> <ul style="list-style-type: none">• Requer unidade de análise• Exige consenso em relação ao significado• Exige mediação humana	<p>Informação valiosa da mente humana. Inclui reflexão, síntese e contexto</p> <ul style="list-style-type: none">• De difícil estruturação• De difícil captura em máquinas• Frequentemente tácito• De difícil transferência.

Etapas



Dados conforme sua Estrutura

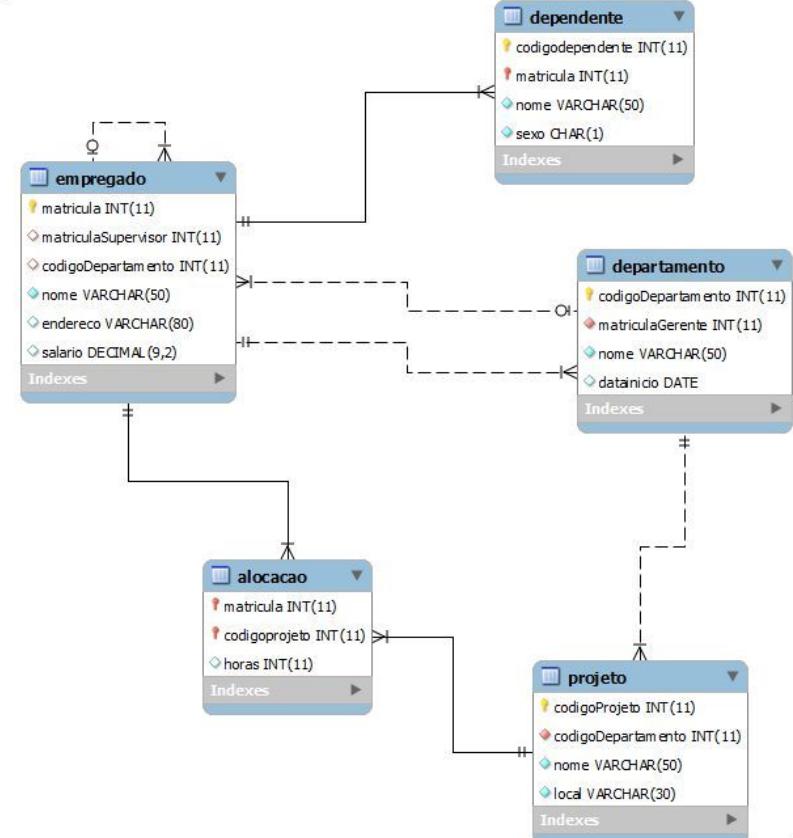
- ❖ Dados estruturados
- ❖ Dados semiestruturados
- ❖ Dados não estruturados



Dados conforme sua Estrutura

❖ Dados estruturados

- Possuem a mesma estrutura de representação rígida e previamente projetada;
- Existe um esquema que estabelece algumas características dos dados que serão armazenados;
- São organizados e gravados em um banco de dados.



Dados conforme sua Estrutura

❖ Dados semiestruturados

- Geralmente não são mantidos em um banco de dados:
 - A maioria dos bancos relacionais admite, por exemplo, XML (eXtensible Markup Language).
- Apresentam organização bastante heterogênea, o que pode dificultar as consultas a esses dados.
- Os dados não são estritamente tipados, mas não são completamente desestruturados.
- Uma análise do dado deve ser feita para que a sua estrutura seja identificada e extraída.

Exemplo de texto e seu correspondente XML

Catálogo de endereços
João Silva
Rua Carijós, 135
Belo Horizonte, MG 30.000
Brasil
31 3335-5556 (preferido)
31 3549-4446
joaosilva@net.com.br
José Almeida
jalmeida@net.com.br

```
<?xml version="1.0"?>
<catálogo de endereços>
<entrada>
  <nome> João Silva </nome>
  <endereço>
    <rua> Carijós, 135</rua>
    <estado> MG </estado>
    <cep> 30.000 </cep>
    <pais> Brasil </pais>
  </endereço>
  <telefone preferido="true">31 3335-4456</telefone>
  <telefone> 31 3594-4446 </telefone>
  <email> joaosilva@net.com.br </email>
</entrada>
<entrada>
  <nome><prim>José</prim>
    <sobren>Almeida</sobren>
  <email> jalmeida@net.com.br </email>
</entrada>
</catálogo de endereços>
```

Dados conforme sua Estrutura

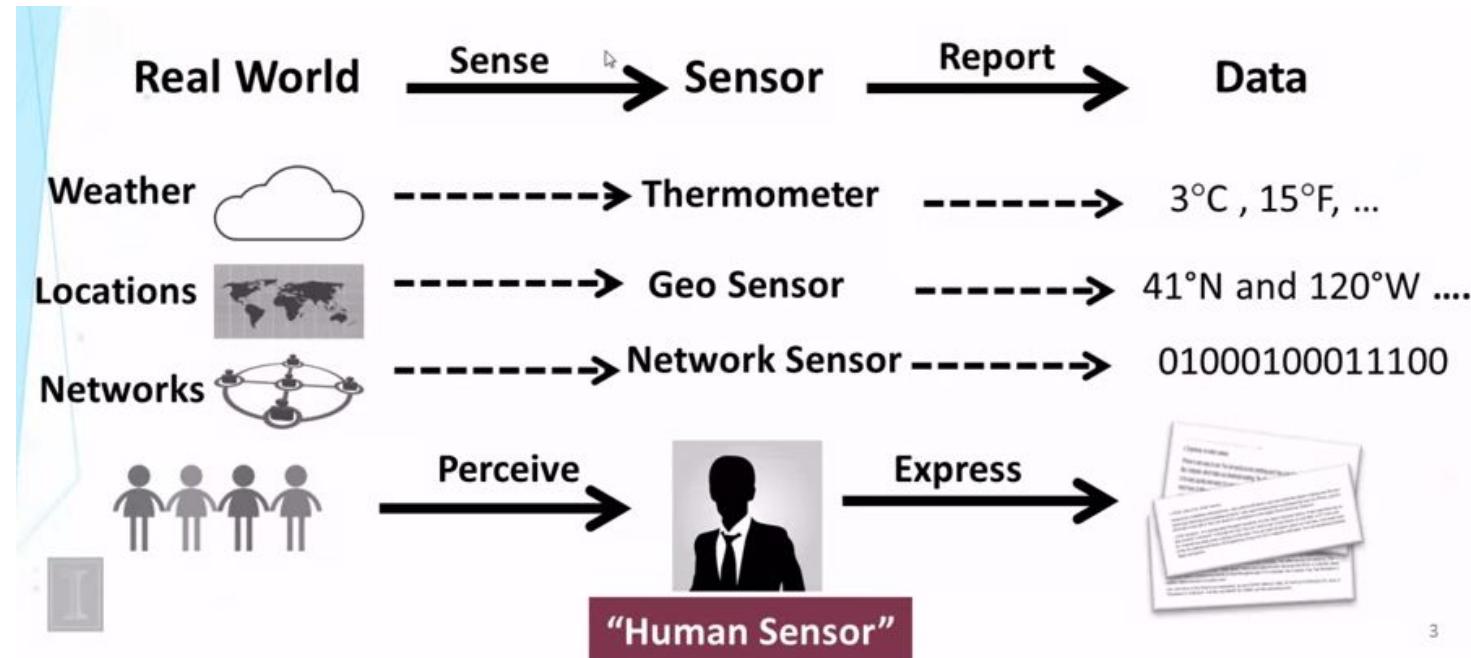
❖ Dados não estruturados

- Aqueles que não possuem uma estrutura definida. Exemplos:
 - Documentos, textos, imagens e vídeos.
- A maior parte dos dados disponíveis na Web são classificados dentro destes formatos.
- Principalmente texto em **linguagem natural**
 - Como lidar com esse tipo de **dado** e obter **informações** de qualidade?



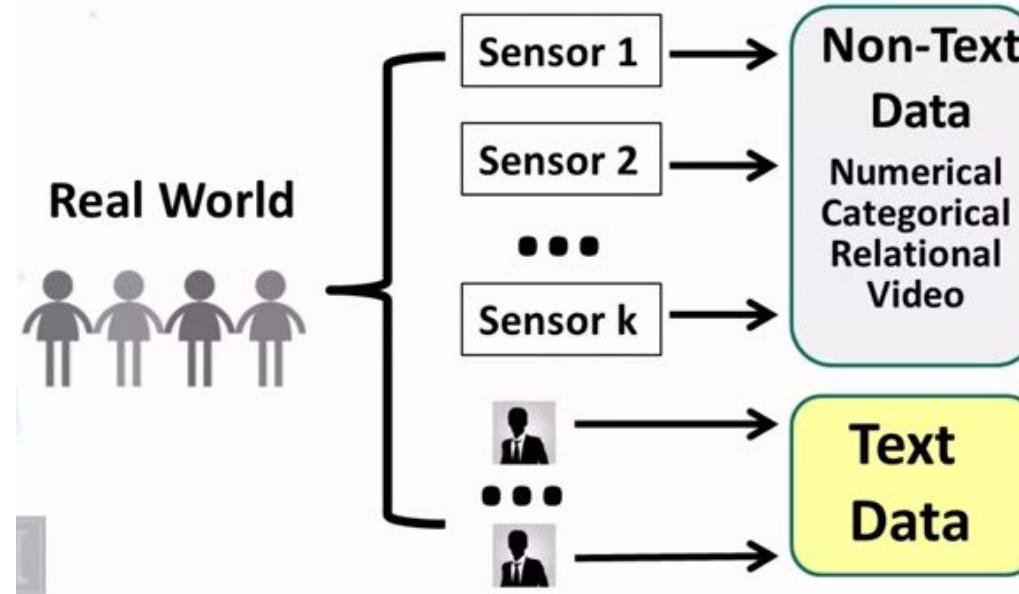
Mineração de Textos

❖ Dados em texto e não textuais



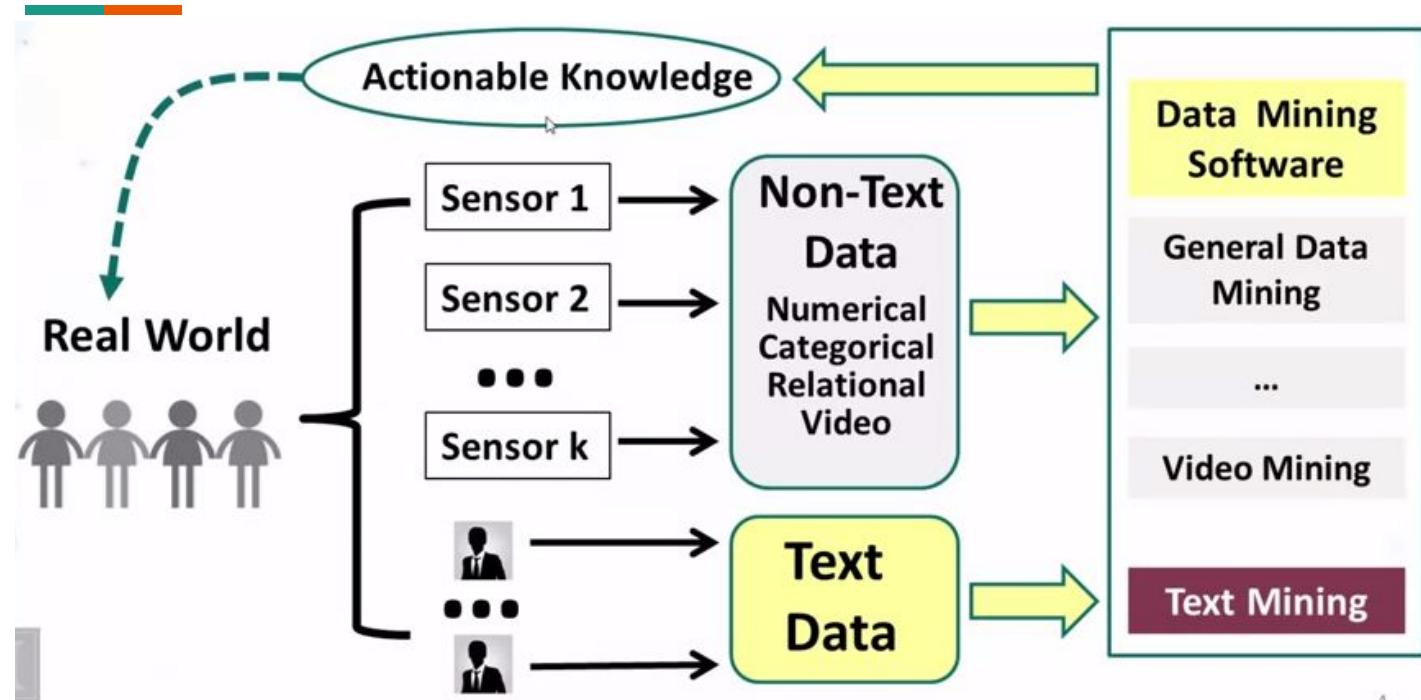
Mineração de Textos

❖ Dados em texto e não textuais



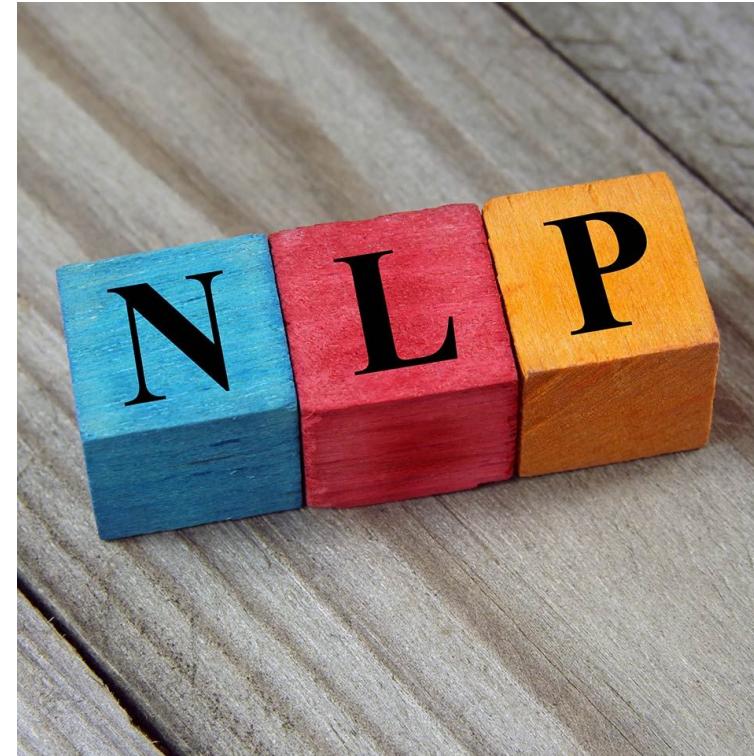
Fonte: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining Paperback by Chengxiang Zhai, Sean Massung

Mineração de Textos



Fonte: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining Paperback by Chengxiang Zhai, Sean Massung

Processamento de Linguagem Natural

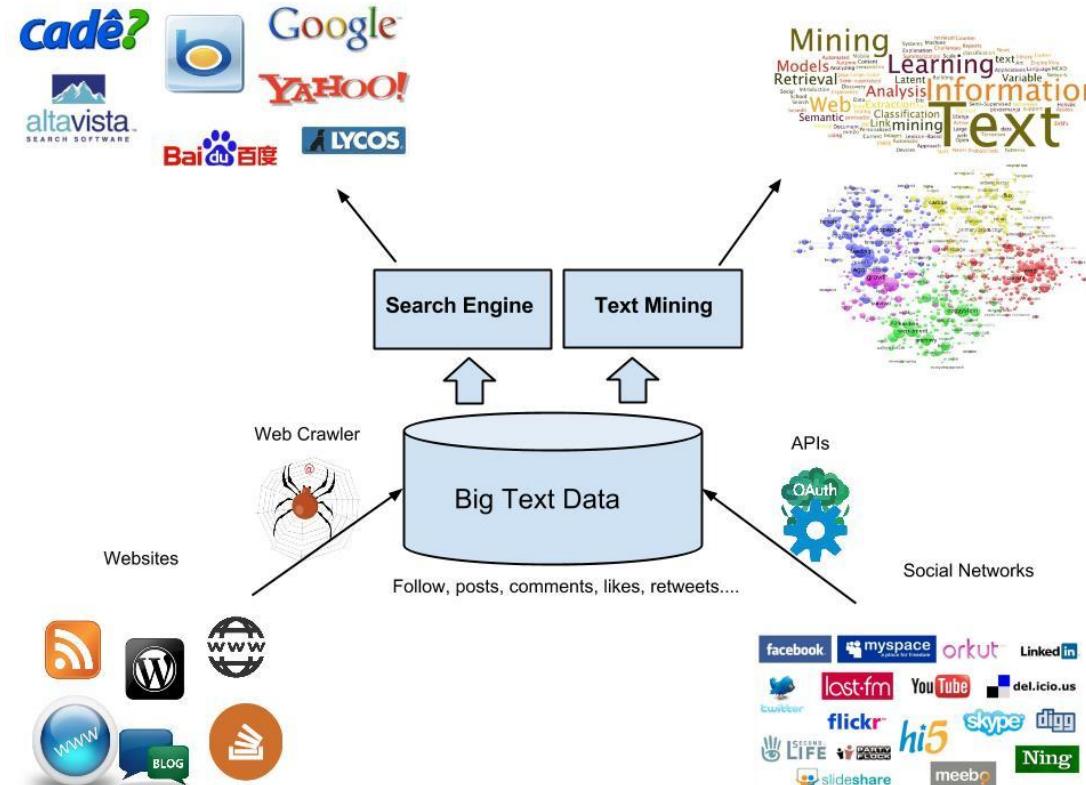


Processamento de Linguagem Natural (NLP)

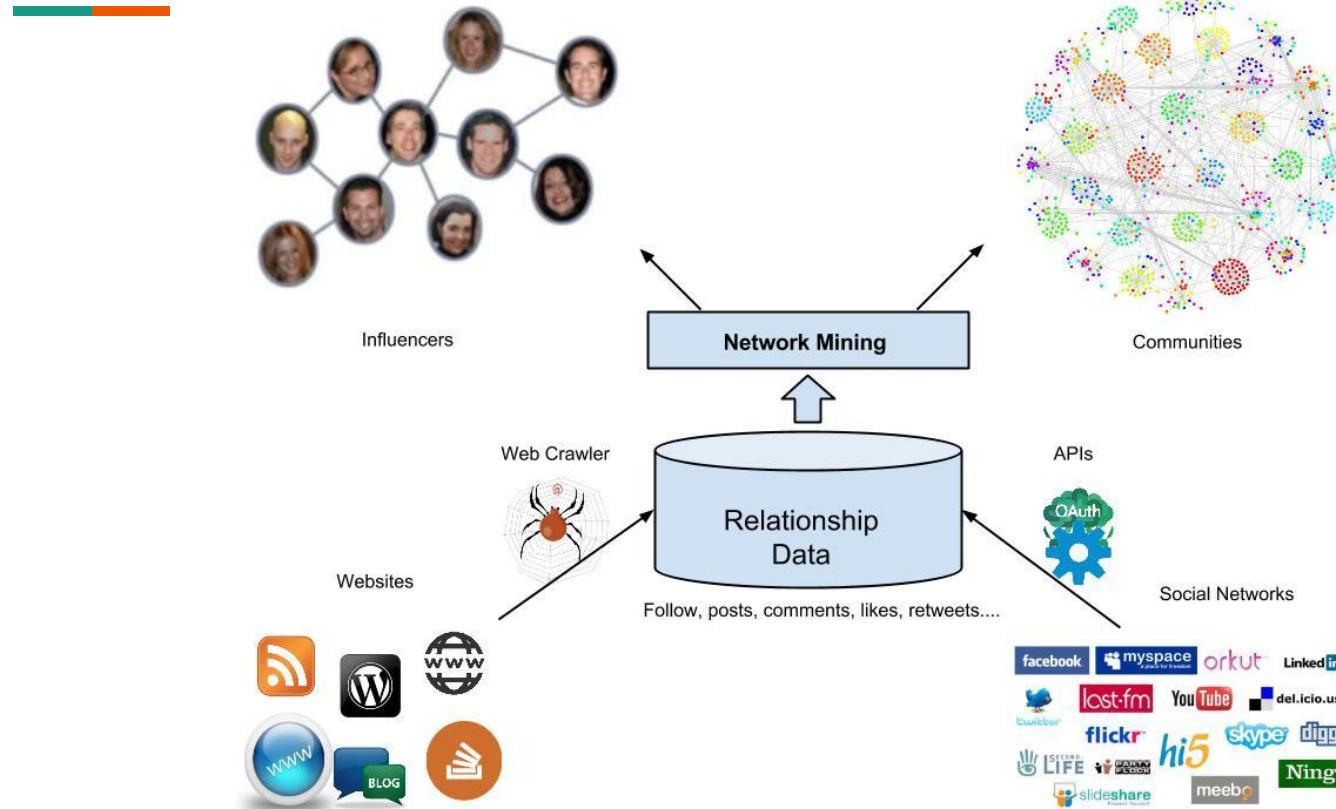
❖ NLP - Natural Language Processing

- NLP é a base para mineração de texto
- Computadores estão muito longe de serem capazes de entender linguagem natural
 - É preciso escalar computacionalmente e em cobertura, o que dificulta o uso de métodos muito profundos e limitados a certos domínios
- Na prática são utilizados métodos (rasos) estatísticos de NLP como base, enquanto humanos fornecem ajuda quando necessário.

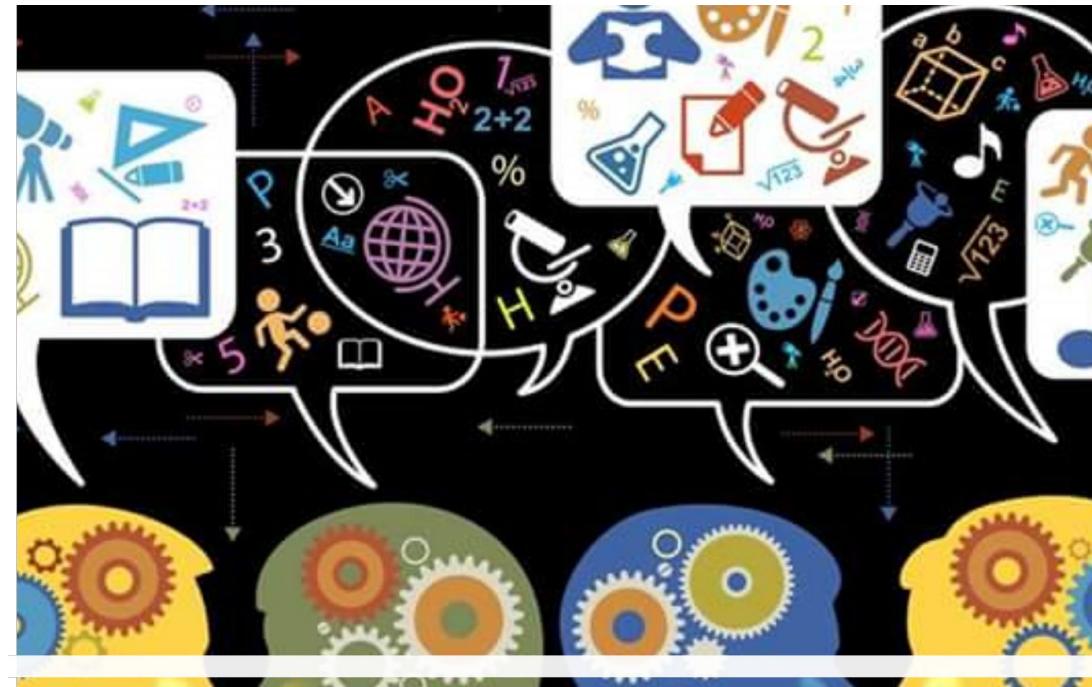
Mineração da Web e Redes Sociais



Network Mining



Conhecimento, Colaboração e Marketing



A Sabedoria das Multidões

❖ Inteligência Coletiva

- Inteligência compartilhada ou de grupo formada através da colaboração, esforço coletivo, competição entre indivíduos e aparecem como consenso em tomadas de decisão.
- “Ninguém sabe tudo, porém todos sabem alguma coisa” -
Paulo Freire
- O termo aparece na **sociobiologia**, na **ciência política** e aplicações de **crowdsourcing**.



A Sabedoria das Multidões

- Inteligência Coletiva: Crowdsourcing
- Crowd (multidão) e Outsourcing (terceirização)
- Conceito de interação social, baseado na construção coletiva de soluções com benefícios a todos.
- Processo de comunicação que pode ser encontrado em comunidades *open source*
- Criação de softwares complexos
- Conhecimento colaborativo (wiki's)
- Ferramentas de gerenciamento de conhecimento em organizações de grande porte



A Força das Multidões

- **Crowdfunding**
- A prática de financiar um projeto ou empreendimento através de contribuições monetárias de um grande número de pessoas, tipicamente através da internet.
- “Um grande número de pessoas unindo seu poder econômico para sustentar uma companhia, organização ou projeto em que acreditam”

(2017) - \$34.4 billion...

(2025) - potential could be between \$90 billion and \$96 billion.."[Fonte](#)



A Força *destrutiva* das Multidões

❖ Como NÃO utilizar um Crowdfunding: Case Zebeléo

≡ EXAME.COM

Lula Trump

Crowdfunding da hamburgueria de Leo e Bel Pesce é cancelado

À primeira vista, a ideia de abrir o Zebeléo até parecia boa. Mas, por uma má escolha de forma de arrecadação, o projeto se queimou com muitos usuários.

Por [Mariana Fonseca](#)
© 5 set 2016, 17h02

f
t
d
e
...



[Fonte](#)

A Força *destrutiva* das Multidões

- **Como NÃO utilizar um Crowdfunding: Case Zebeleu**
- “Muitos apontaram que as contribuições eram caras diante das recompensas recebidas”
- “Não oferecia nenhum tipo de participação no negócio em troca das contribuições”
- “Para os críticos, os três sócios poderiam muito bem conseguir um investidor para o seu negócio, ou colocar a mão no bolso pela hamburgueria, em vez de pedir dinheiro para as pessoas comuns.”

A Força *destrutiva* das Multidões

➤ A gafe chamou a atenção do país e..... O caso foi muito além....

Bel Pesce: empreendedora malcompreendida ou farsa midiática?

POR EQUIPE TECMUNDO - EM POLÉMICA - 02 SET 2016 - 18H07



Fonte

"Bel Pesce afirmava ter cinco diplomas no MIT e ter trabalhado em grandes empresas de tecnologia nos Estados Unidos"

A Força *destrutiva* das Multidões

- Vários outros casos...

The Dog Haüs: como destruir a reputação de um restaurante nas redes sociais

Depois de "mal entendido", Shemuel Shoel, dono da lanchonete The Dog Haüs, volta a xingar cliente. Reputação foi rebaixada

BRUNO FERRARI

[Fonte](#)

Das 4,8 estrelas, seu perfil no Facebook passou a ter apenas 1,5.

jornalista Leka Peres em sua visita à lanchonete The Dog Haüs, no bairro do Itaim, em São Paulo. Disse que aprovou o hotdog, mas desaprovou a decoração considerada por ela machista.

"Caramba Qt gente infeliz nesse mundo, isso é decoracao bando de babaca , aqui repaeotos a todos , ficou ofendido??? Come HotDog em outro pico".

Inteligência Competitiva

- Processo sistemático e **ético** de coleta, tratamento e análise da informação sobre atividades dos concorrentes, tecnologias e mercado, visando subsidiar a tomada de decisão e atingir as metas estratégicas da empresa.
- Coletar informações do ambiente externo para entender as forças e fraquezas dos competidores;
- Avaliar sua própria competitividade;
- **Prever*** as intenções dos competidores, as expectativas dos clientes e do mercado como um todo.



*No sentido de ser proativo e antecipar da melhor maneira possível algumas tendências com base na experiência. Não existe previsão de fato, a menos que o passado se repita em alguma forma de padrão.

Redes e Mídias Sociais

- Uma **rede social** é uma **estrutura social** composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que partilham valores e objetivos comuns.
- As **mídias sociais** são **sistemas online** projetados para permitir a interação social a partir do compartilhamento e da criação colaborativa de informação nos mais diversos formatos.
 - *"Um grupo de aplicações para Internet construídas com base nos fundamentos ideológicos e tecnológicos da Web 2.0, e que permitem a criação e troca de conteúdo gerado pelo utilizador."* - Andreas Kaplan e Michael Haenlein



Mundo Social

Internacional

A Islândia prepara nova constituição. Via Facebook

por Paula Thomaz — publicado 07/07/2011 19h06, última modificação 17/07/2011 11h17

Redes sociais são amplamente utilizadas para discussão popular sobre o tema; para especialista brasileiro, tamanho diminuto do país facilita. Por Paula Thomaz

 Recomendar 92

 +1 1

 Share

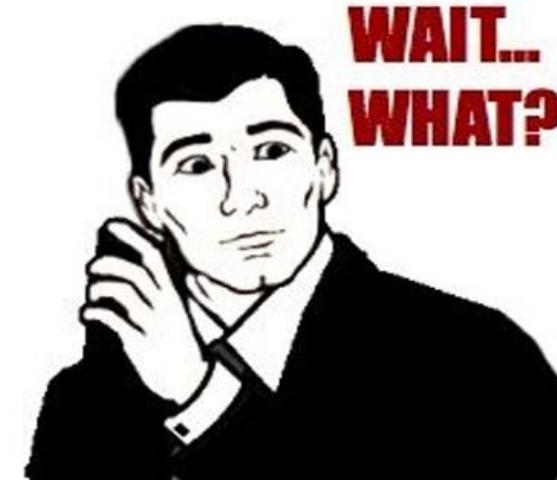
 Tweetar 8

Três anos depois de passar por uma crise financeira que derrubou o primeiro-ministro Geir Haarde, quebrou vários dos bancos nacionais e causou uma queda drástica do valor da moeda local, a Islândia aos poucos tenta se reerguer. E o passo mais incisivo para a volta por cima tem sido a discussão para a criação de uma nova Constituição que dê novos ares à base legal do estado islandês. E com uma grande novidade: esta nova constituição está sendo discutida via



O Facebook entra no mercado com alto valor. Mas corre o risco de não corresponder às expectativas.
Foto: iStockphoto

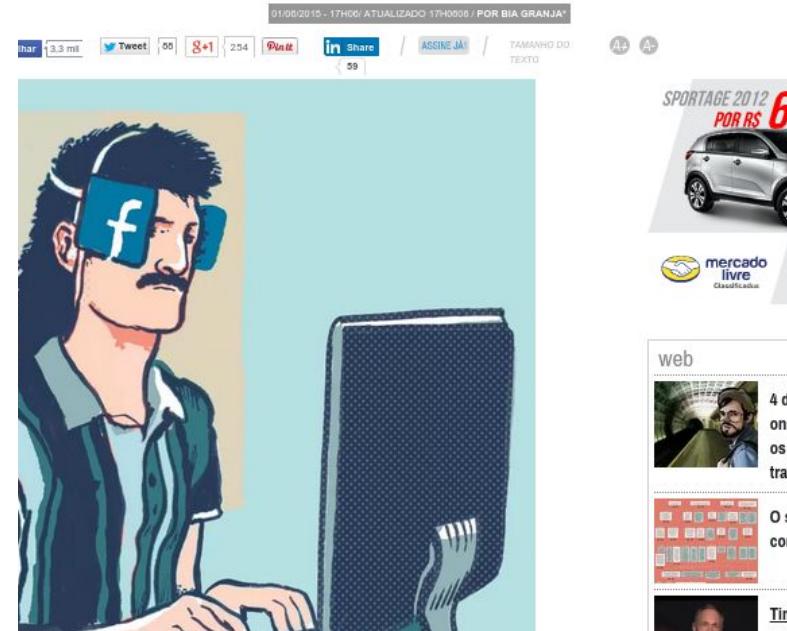

Proposta de Emenda Constitucional → **Constituição Federal de 1988** → **Norma Internacional** → **Islândia: uma experiência constitucional para o Brasil?**



Mundo Social

Bem-vindos à Zuckernet: os efeitos de conhecer o mundo através de uma única rede social

Não experimentar a verdadeira web, pode ser problemático



The collage illustrates the concept of Zuckernet, showing how people can become so engrossed in their online social networks that they miss out on experiencing the real world. It features:

- A cartoon illustration of a man with a mustache wearing Facebook sunglasses, sitting at a computer.
- A screenshot of a news article from UOL.com.br dated 01/06/2015, with social sharing icons for LinkedIn, Twitter, Google+, and Pinterest.
- An advertisement for a Kia Sportage 2012 car, priced at R\$ 63,000.
- A snippet of a web search results page showing a thumbnail of a person and some text.

Mundo Social

FOLHA DE S.PAULO

[MENU](#) [ASSINE](#) [ENTRAR](#) [BUSCAR](#)

[poder](#) > governo bolsonaro lava jato entrevista da 2^a

AGÊNCIA LUPA PIAUÍ

Na onda de Bolsonaro, Twitter vira canal de anúncios e poemas

Rede social adotada por Donald Trump atrai nomes do poder como FHC e Mourão



Tuites publicados nos últimos meses pelo ex-presidente do STF Ayres Britto, pelo ex-presidente Fernando Henrique e pelo vice-presidente eleito, Hamilton Mourão

Reprodução

[Fonte](#)

Mundo Social



Tweets 40.5K Following 45 Followers 58.1M Likes 7 Moments 6

Donald J. Trump 

@realDonaldTrump

45th President of the United States of America 

Washington, DC

Instagram.com/realDonaldTrump

Joined March 2009

[Tweet to Donald J. Trump](#)

3 Followers you know



2,977 Photos and videos



Tweets **Tweets & replies** **Media**

Donald J. Trump  @realDonaldTrump · 58m
2019 National Prayer Breakfast



President Trump Delivers Remarks at the 2019 National Prayer Breakfast...
The White House @WhiteHouse

4.3K 4.4K 17K

Donald J. Trump  @realDonaldTrump · 4h
PRESIDENTIAL HARASSMENT! It should never be allowed to happen again!

38K 17K 82K

Donald J. Trump  @realDonaldTrump · 4h
Democrats at the top are killing the Great State of Virginia. If the three failing pols were Republicans, far stronger action would be taken. Virginia will come back HOME Republican in 2020!

8.9K 12K 51K

Who to follow · Refresh · View all

President Trump  @PO...
[Follow](#)

Barack Obama  @Bara...
[Follow](#)

The White House  @W...
[Follow](#)

[Find people you know](#)

Trends for you · Change

#sddaaepocaque
Qual é a época que mais deixou saudades?

#EUQueriaQueAsPessoas
11 Tweets com sugestões para um mundo melhor

#QuintaDetremuraSDV
16.5K Tweets

#chuvraRJ
Número de mortes provocadas por temporal Rio só pra 6

#fradiobrasil
1,511 Tweets

#LulaLivreJá
4,582 Tweets

#OABInimigaDoBrasil
3,032 Tweets



Jair M. Bolsonaro 

@jairbolsonaro

Capitão do Exército Brasileiro, eleito 38º Presidente da República Federativa do Brasil. 

Brasília, Brazil
[bolsonaro.com.br](#)
Joined March 2010

[Tweet to Jair M. Bolsonaro](#)

7 Followers you know



1,992 Photos and videos



1:05 190K views

Secretary Pompeo Meets with Brazilian Foreign Minister Araújo
Secretary of State Michael R. Pompeo meets with Brazilian Foreign Minister Ernesto Henrique Fraga Araújo at the State Department on February 5, 2019. - U.S. Department of State

Who to follow · Refresh · View all

Eduardo Bolsonaro  @...
[Follow](#)

Flávio Bolsonaro  @Fl...
[Follow](#)

Danilo Gentili  @Danilo...
[Follow](#)

[Find people you know](#)

Tweets **Tweets & replies** **Media**

Jair M. Bolsonaro  @jairbolsonaro · 5h
Começamos mais uma quinta-feira combatendo o bom combate. Temos uma missão e vamos cumprí-la. Precisamos estar unidos para transformar o Brasil em um local mais seguro para os cidadãos de bem! Não perderemos esta oportunidade única! Contem conosco! Nenhum assassino irá nos parar!

[Translate Tweet](#)

4.0K 5.6K 43K

Jair M. Bolsonaro Retweeted
Department of State  @StateDept · Feb 5
Today, @SecPompeo welcomed Brazil's Foreign Minister @ernestofarajau to the State Department. @Itamaraty_EN @EmbaixadaEUA @BrazilinUSA



Who to follow · Refresh · View all

Eduardo Bolsonaro  @...
[Follow](#)

Flávio Bolsonaro  @Fl...
[Follow](#)

Danilo Gentili  @Danilo...
[Follow](#)

[Find people you know](#)

Trends for you · Change

#sddaaepocaque
Qual é a época que mais deixou saudades?

#EUQueriaQueAsPessoas
11 Tweets com sugestões para um mundo melhor

#QuintaDetremuraSDV
16.5K Tweets

#chuvraRJ
Número de mortes provocadas por temporal no Rio só pra 6

#fradiobrasil
1,545 Tweets

#LulaLivreJá
4,582 Tweets

#BBdebate

#OABInimigaDoBrasil

Relacionamentos Criativos

 MauroJunior Junior ▶ Bradesco 
24 de outubro de 2011 · 

Banco Bradesco querido
 Quisto por mim e os meus
 Tens sua morada paulista
 Bem na Cidade de Deus
 Vejam que bela homenagem
 O próprio Deus concebeu
 Para a sua cidade
 O vosso Banco escolheu
 Eu até que me poria
 Em alta colina à bradar
 Peito banhado em verdade
 Bradesco em primeiro lugar
 Mas venho por outro motivo
 O que findou meu sorris
 Para por fim ao martírio
 Um favor vou lhes pedir
 Plena falta de cuidado
 Digna de um jabuti
 Fazendo compras no mercado
 O meu cartão eu perdi
 Antes que eu passe fome
 Faço a solicitação
 Ao meu Banco preferido
PRECISO DE OUTRO CARTÃO!
 @bradesco

Banco Bradesco querido
Quisto por mim e os meus
Tens sua morada paulista
Bem na Cidade de Deus



Bradesco Mauro querido cliente
 Pra você ter outro cartão
 à sua agência deve ir pessoalmente

Mas não será por motivos fúteis
 Você irá cadastrar uma nova senha
 E seu cartão chegará em até 7 dias úteis

Agradecemos a sua compreensão
 E sempre que precisar
 Pode contar com a nossa colaboração

;-)
 Yesterday at 09:17 · Like ·  1,556 people

Relacionamentos Criativos

POR LUIZA BELLONI VERONESI - EM NEGÓCIOS / COMO-VENDER-MAIS - 015 MAI, 2014 15H55

Netshoes e Centauro travam divertida disputa por cliente no Twitter; confira

Cliente intermedia disputa de empresas e ganha desconto de 25%

 em ambas. Discutam entre vocês qual é a melhor pra eu comprar.
[Detalhes](#)  [Responder](#)  [Retweetar](#)  [Curtir](#)  [Mais](#)

 **Pedro Tessarolo** @pedrotri · 9 de mai
@[siganetshoes](#) e @centauroesporte preciso saber disso antes das 17 horas se não eu vou ter que ir para a academia descalço :(

[Detalhes](#)  [Responder](#)  [Retweetar](#)  [Curtir](#)  [Mais](#)

 **Centauro** @centauroesporte · 9 de mai
Quem vai sair na frente é você agora, @pedrotri! 10% de desconto pra você. Fechado? bit.ly/1m7iqKy

[Detalhes](#)  [Responder](#)  [Retweetar](#)  [Curtir](#)  [Mais](#)

MAIS LIDAS



IMPRESSIONA ANALISTAS

Amazon Brasil cresce 167% em tráfego após expansão



...



NÃO É SÓ A VALE

Tragédia de Brumadinho paralisa venda de ativo de R\$ 1 bilhão da Usiminas



...

MAKE MONEY

Relacionamentos Criativos

- Pedro Tessarolo** @pedrotrl - 9 de mai
 @siganetshoes e @centauroesporte quero um tênis e vi que está o mesmo preço em ambas. Discutam entre vocês qual é a melhor pra eu comprar.
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Pedro Tessarolo** @pedrotrl - 9 de mai
 @siganetshoes e @centauroesporte preciso saber disso antes das 17 horas se não eu vou ter que ir para a academia descalço :(
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Centauro** @centauroesporte - 9 de mai
 Quem vai sair na frente é você agora, @pedrotrl! 10% de desconto pra você. Fechado? bit.ly/1m7iqky
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Pedro Tessarolo** @pedrotrl - 9 de mai
 @centauroesporte nocaute na @siganetshoes é isso mesmo? O cartão de crédito já está na mão.
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Netshoes** @siganetshoes - 9 de mai
 @pedrotrl @centauroesporte a luta só acaba quando o último beija a lona. 20% de desconto e temos um acordo?
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Pedro Tessarolo** @pedrotrl - 9 de mai
 @siganetshoes POR MIM SIM! 20% DE DESCONTO TÁ LIBERADO PRA TER COPA. SERÁ QUE AGORA A @centauroesporte CONSEGUE REVIDAR?
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Centauro** @centauroesporte - 9 de mai
 @pedrotrl @siganetshoes O round ainda não acabou! Fechado em 25% de desconto?
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Pedro Tessarolo** @pedrotrl - 9 de mai
 @centauroesporte @siganetshoes Isso está igual Anderson Silva e Weidman, e pelo jeito foi a Netshoes que quebrou a perna!
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)
- Netshoes** @siganetshoes - 9 de mai
 @pedrotrl @centauroesporte a gente cobre os 25% de desconto. Em quem você confia mais para entregar seu produto?
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)

Netshoes e Centauro "brigam" por consumidor no Twitter

Os perfis das duas empresas no Twitter começaram a dialogar quando um consumidor lançou um desafio para ambas

Pedro Tessarolo @pedrotrl - 9 de mai
 @siganetshoes acordo fechado.
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)

 **Netshoes** @siganetshoes
[Seguir](#)

@pedrotrl boa, Pedro! Vamos te mandar uma DM para fechar o pedido, blz?

[Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)

16:34 - 9 de mai de 2014

Responder a @siganetshoes @pedrotrl

Pedro Tessarolo @pedrotrl - 9 de mai
 @siganetshoes manda a maquininha do cartão por DM que já vou passar!
[Detalhes](#) [Responder](#) [Retweetar](#) [Curtir](#) [Mais](#)

Marketing Social e Monitoramento

- ❖ **Por que as empresas devem aderir às mídias sociais?**
- Cada vez mais pessoas estão passando a usar regularmente as redes sociais
- Se as pessoas buscam a empresa dentro das mídias sociais, é possível que queiram manter e ampliar o relacionamento com elas.



Marketing Social e Monitoramento

❖ E o que acontece se as empresas não aderirem?

- As empresas, de uma forma ou de outra, estão sendo comentadas nas mídias sociais.
- A polaridade das discussões pode ser positiva ou negativa, e não há controle sobre essas interações, apenas influência.
- Não participar das conversas pode significar que a empresa endossa o que é dito, seja positiva ou negativamente.

Marketing Social e Monitoramento

- ❖ Lembra do caso?

O silêncio é a pior estratégia para Bel Pesce

[Fonte](#)

Após a polêmica hamburgueria "Zebeléo", blogueiro publicou um texto com acusações sérias à empreendedora e palestrante. Bel deveria falar

“...a mesma dinâmica das redes sociais que a colocou no olimpo do empreendedorismo do Brasil pode tirá-la de lá...”

Marketing Social e Monitoramento

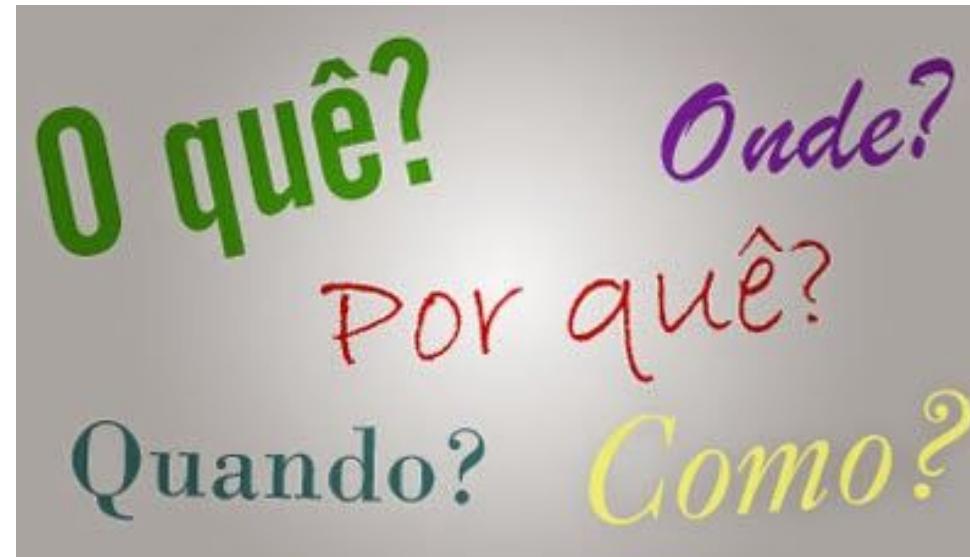
❖ Como as empresas podem usar as mídias sociais?

- Lançamento e divulgação de produtos
- Análise do comportamento dos consumidores, tendências e oportunidades
- Monitoramento da marca para gestão estratégica
- Comunicação, suporte e relacionamento com o cliente
- Gasta-se muito em pesquisas de mercado e sistemas de qualidade.
 - Mas e a voz que vem da Web?

Marketing Social e Monitoramento

❖ Diagnóstico

- Saber o quê, quem, quando, onde e como estão falando da empresa



Tesla

[STORIES](#)[NEWS](#)[STARTUPS](#)[OFFICES](#)[LEARN](#)[OFFICE HUMOUR](#)

Elon Musk Receives Product Suggestion On Twitter, Tesla Implements It 6 Days Later

Posted on January 10, 2017 by [OfficeChai Team](#)

[Fonte](#)



Loic Le Meur

@loic

Follow

Now Musk has 6.42 million Twitter followers,

@elonmusk the San Mateo supercharger is always full with idiots who leave their tesla for hours even if already charged.

1:00 AM · 11 Dec 2016 · San Mateo, CA



37



322



Elon Musk

@elonmusk

Follow

@loic You're right, this is becoming an issue. Supercharger spots are meant for charging, not parking. Will take action.

1:20 AM · 11 Dec 2016



259



2,241

thousands of such tweets per day. But he immediately

Tesla

[STORIES](#)[NEWS](#)[STARTUPS](#)[OFFICES](#)[LEARN](#)[OFFICE HUMOUR](#)

Elon Musk Receives Product Suggestion On Twitter, Tesla Implements It 6 Days Later

Posted on January 10, 2017 by [OfficeChai Team](#)

[Fonte](#)



Loic Le Meur

@loic

Follow

Now Musk has 6.42 million Twitter followers,

@elonmusk the San Mateo supercharger is always full with idiots who leave their tesla for hours even if already charged.

1:00 AM - 11 Dec 2016 · San Mateo, CA



37

322



Elon Musk

@elonmusk

Follow

@loic You're right, this is becoming an issue. Supercharger spots are meant for charging, not parking. Will take action.

1:20 AM - 11 Dec 2016

259 2,241

thousands of such tweets per day. But he immediately

Tesla

[STORIES](#)[NEWS](#)[STARTUPS](#)[OFFICES](#)[LEARN](#)[OFFICE HUMOUR](#)

Elon Musk Receives Product Suggestion On Twitter, Tesla Implements It 6 Days Later

Posted on January 10, 2017 by [OfficeChai Team](#)

[Fonte](#)



Loic Le Meur

@loic

Follow

@elonmusk the San Mateo supercharger is always full with idiots who leave their tesla for hours even if already charged.

1:00 AM - 11 Dec 2016 · San Mateo, CA



37



322



Elon Musk

@elonmusk

Follow

@loic You're right, this is becoming an issue. Supercharger spots are meant for charging, not parking. Will take action.

1:20 AM - 11 Dec 2016



259



2,241

thousands of such tweets per day. But he immediately

Tesla

[STORIES](#)[NEWS](#)[STARTUPS](#)[OFFICES](#)[LEARN](#)[OFFICE HUMOUR](#)

Elon Musk Receives Product Suggestion On Twitter, Tesla Implements It 6 Days Later

Posted on January 10, 2017 by [OfficeChai Team](#)

[Fonte](#)



Loic Le Meur 
@loic

 [Follow](#)

Now Musk has 6.42 million Twitter followers,

@elonmusk the San Mateo supercharger is always full with idiots who leave their tesla for hours even if already charged.

1:00 AM - 11 Dec 2016 · San Mateo, CA



 37

 322



Elon Musk 
@elonmusk

 [Follow](#)

@loic You're right, this is becoming an issue. Supercharger spots are meant for charging, not parking. Will take action.

1:20 AM - 11 Dec 2016

  259  2,241

thousands of such tweets per day. But he immediately

Marketing Social e Monitoramento

❖ Planejamento

- O Refletir sobre a maneira mais adequada de entrar nas redes e se preparar
- Estabelecer objetivos e metas
- Alocar pessoas e recursos
- Preparar as pessoas e as regras a serem respeitadas (políticas de uso).

Gafes

 **LG_France**
@LG_France

Fonte

Nos smartphones ne se plient pas, ils sont naturellement incurvés ;)
[#bendgate](#) pic.twitter.com/ZoHdXyUWkU



September 25, 2014 at 4:16 AM
via Twitter for iPhone

☆ 61 188

◀ ▶ ★ ☰ ⚙

September 25, 2014 at 4:16 AM
via Twitter for iPhone



Introdução

❖ Revisão

- A Web como um imenso repositório de informações
- Necessário processar grandes volumes de texto em linguagem natural
- Um ambiente de colaboração em massa como a Web pode gerar conhecimento, ferramentas e até financiamentos de projetos
- Empresas devem acompanhar as interações em busca de uma vantagem competitiva e melhor relacionamento com clientes

Recuperação da Informação



Recuperação da Informação

- Conceitos básicos de RI
- Mais ênfase na recuperação de informação na Web

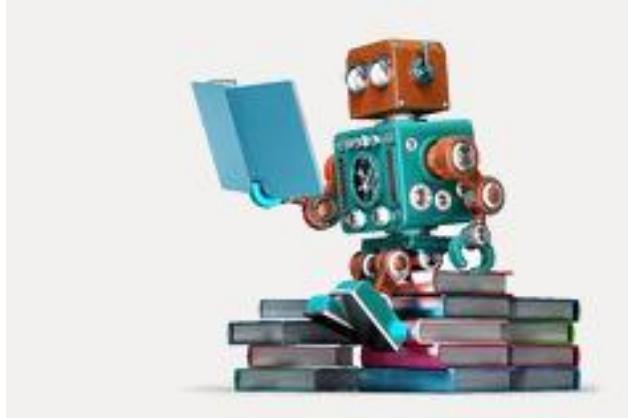


Fonte:<http://irlab.csie.ntu.edu.tw/lab/>

Recuperação da Informação

❖ O que é RI ?

- Recuperação da informação é uma subárea da ciência da computação que trata da recuperação **automática da informação**
- Normalmente contida em **documentos**.

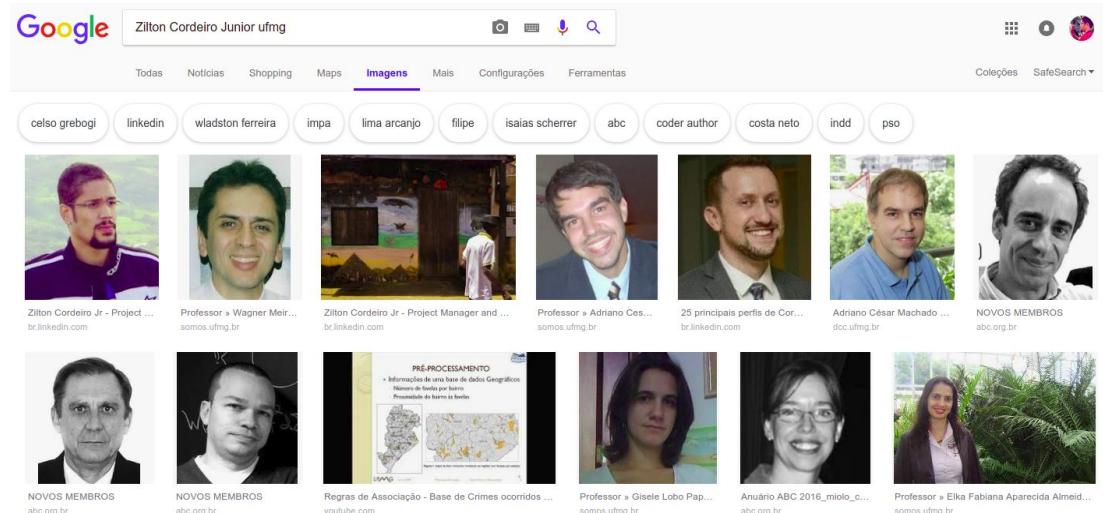


Recuperação da Informação

◆ Documentos

- Fazem o papel de **registros de dados**
- Podem conter qualquer tipo de mídia (texto, imagem, som)
- Normalmente são compostos de textos em linguagem natural

- Ou de informação textual
associada a outros tipos
de dados.



Ex: Google Imagens

Recuperação da Informação

❖ Recuperar Dados vs Recuperar Informação

Recuperação de Dados:

- Tarefas precisas
- Sistemas não visam incorporar o significado do que está sendo buscado
- Respostas devem ser corretas

Recuperação da Informação

❖ Recuperar Dados vs Recuperar Informação

Recuperação de Informação:

- Tarefas imprecisas
- Sistemas tentam modelar o significado do que está sendo buscado
- Objetivo é trazer as melhores respostas
(normalmente não há o conceito de resposta correta)

Recuperação da Informação

❖ Exemplos

Recuperação de **Dados**:

- Uma busca por documentos que **contém** a palavra Belo Horizonte

Recuperação de **Informação**:

- Uma busca por **bons documentos** que falam sobre a cidade de Belo Horizonte

Histórico

Surgimento na década de 60 [Leia +]

- Principal objetivo era automatizar o acesso a informação em bibliotecas.

Principais focos até final dos anos 80:

- Catálogos de bibliotecas, jornais, revistas, enciclopédias eletrônicas e bases de dados de empresas.

Foco mais recente na Web

- Abundância de informação não estruturada
- Publicação sem controle central e diversidade
- Dificuldade na busca de informação específica

Exemplos de Problemas

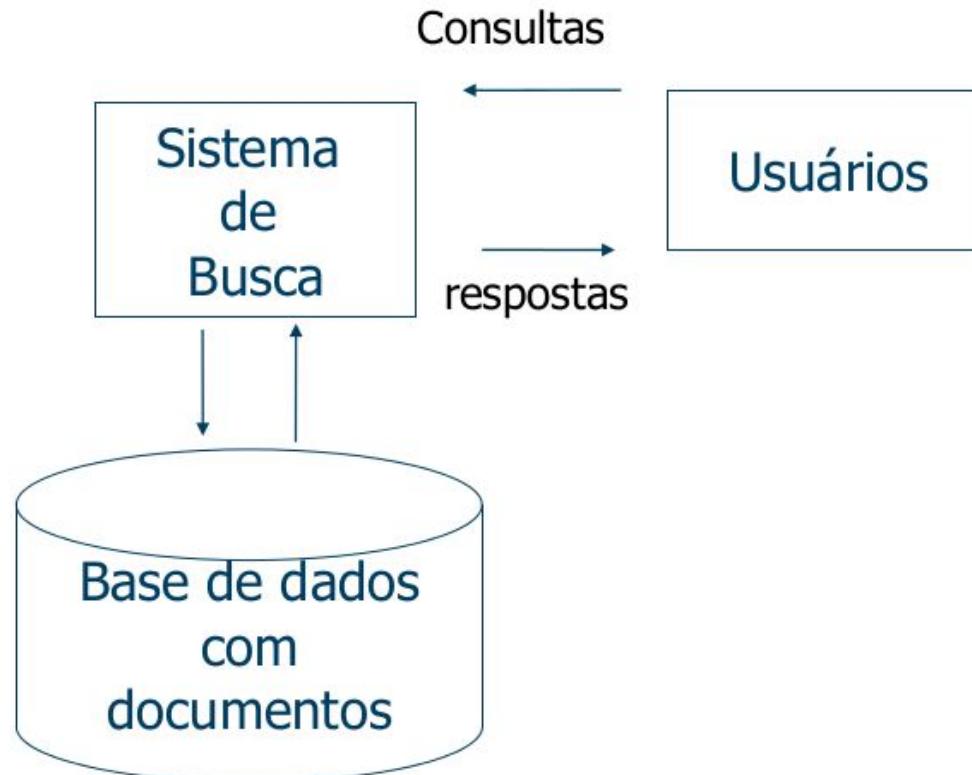
- Busca
- Filtragem e Recomendação (entrega de informação)
- Classificação

Exemplos de Problemas

❖ Busca

- Tipo mais comum
- Usuários apresentam uma consulta e há uma base de dados contendo documentos onde o sistema busca respostas
- **Base de dados:** documentos
- **Entrada:** consultas dos usuários
- **Objetivo:** retornar os documentos que melhor atendem às consultas

Sistemas de Busca



Ranking de Similaridade

Google Microphone Search

Todas Maps Notícias Imagens Shopping Mais Configurações Ferramentas

Aproximadamente 10.800.000 resultados (0,54 segundos)

Pós Graduação PUC-RS Online | Cursos EAD em todas as áreas
[Anúncio](#) online.pucrs.br/puc ▾ (51) 3937-5012
Cursos de Pós-graduação PUC-RS. Conheça os Professores. Saiba mais. Presencial ou Online. Conheça nossas Pós - Conheça nossas Extensões

Processo Seletivo Simplificado PUC Minas
<https://www.pucminas.br/> ▾
Candidato se inscreve aqui. 2. A PUC Minas entrará em contato, em até 48 horas, com o resultado e informações sobre a matrícula. Seleção por redação. 1.

Resultados de pucminas.br Search

Aluno
Guia do Aluno · Guia Prático do Aluno · Sistema de Gestão ...

Vestibular
Vestibular · Busca · PUC > Formas de Ingresso > Vestibular ...

Virtual
Graduação - Disciplinas a distância - Licenciatura - ...

Pós-graduação
O IEC PUC Minas oferece cursos de pós-graduação Formação e ...

Graduação
Graduação. Localização. Todos os campi/ ... Graduação. Arcos ...

Biblioteca
Atribuições. Coordenar o processo de seleção e aquisição de ...

 Map data ©2019 Google

Pontifícia Universidade Católica de Minas Gerais
Instituição de ensino superior em Belo Horizonte, Minas Gerais 

A Pontifícia Universidade Católica de Minas Gerais é uma instituição de ensino superior, privada e católica brasileira situada em Belo Horizonte, capital do estado de Minas Gerais. [Wikipédia](#)

Chanceler: Walmer Oliveira de Azevedo
Número de alunos: 63.528
Sede: Belo Horizonte, Minas Gerais
Fundação: 12 de dezembro de 1958
Subsidiárias: PUC Minas, Campus de Poços de Caldas, MAIS

Ex-alunos notáveis Ver mais 1

Aécio Neves, Cármen Lúcia, Fernando Pimentel, Chico Pinheiro, Clésio Andrade

Pesquisas relacionadas Ver mais 5


PUC-SP - Campus Monte Al... São Paulo
Universid... Federal de Minas Ge... Belo Horizonte
Pontifícia Universid... Católica... Rio de Janeiro
Universid... FUMEC Belo Horizonte
PUC Minas Poços de Caldas

Máquinas de busca para a Web

Cadê?



DuckDuckGo

Google
YAHOO!

Ask **bing™**
Aol Search.

Bai du 百度

Privacidade



- <https://www.theguardian.com/technology/2014/apr/04/duckduckgo-gabriel-weinberg-safe-searches>
- <https://duckduckgo.com/privacy>

Máquinas de busca para a Web

❖ Principais Componentes

- Coletor
- Indexador
- Processador de consultas: Ranking e Eficiência

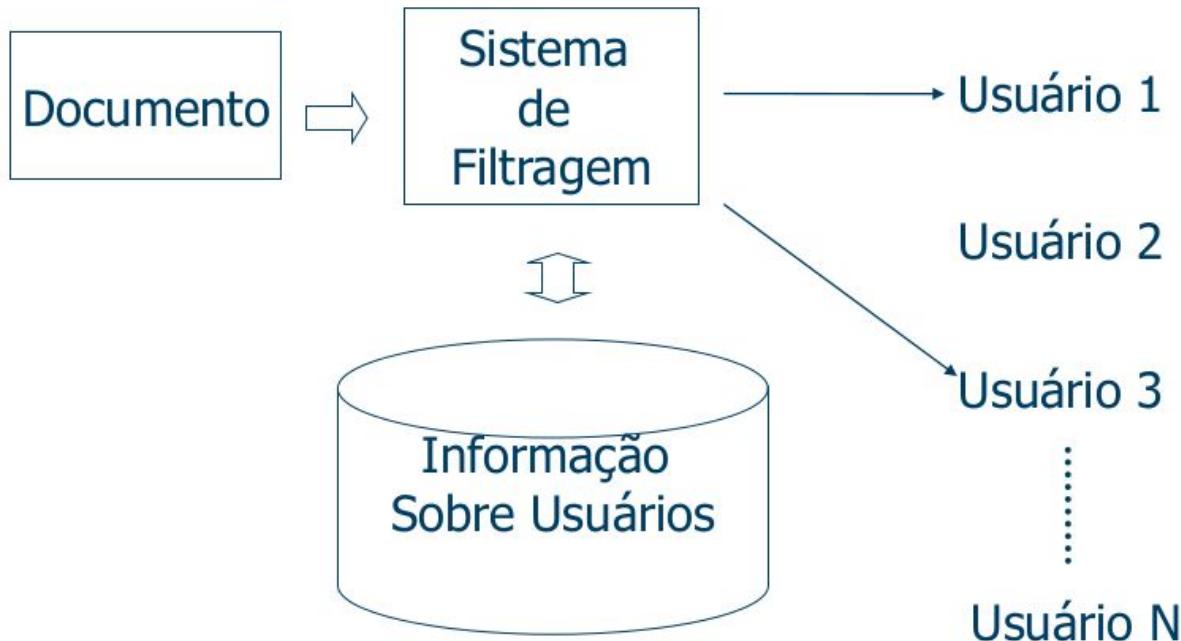
Sistemas de Busca

- 
- Bibliotecas digitais.
 - Enciclopédias, jornais, revistas e livros disponíveis em meios eletrônicos
 - Serviços de busca para a Web
 - Busca em SGBDs com ranking
 - Busca interna em sites institucionais

Filtragem e Recomendação

-
- Inverso do problema de busca
 - Base de dados: os interesses de cada usuário
 - Entrada: documentos
 - Objetivo: identificar os usuários que se interessam pelos documentos

Filtragem e Recomendação



Sistemas de Filtragem

- 
- Recorte de jornais (clipping)
 - Controle de correspondência
 - Filtragem de informação em redes ponto a ponto

Clipping



JAC T40 é o novo crossover compacto da marca chinesa que chega em junho

JAC T40 é o novo crossover compacto da marca chinesa que chega em junho ... 1 Minutos de Leitura A JAC Motors anunciou a chegada do crossover compacto T40 ao mercado nacional, o que deve acontecer em ... [Leia mais](#)



Os esportivos australianos mais legais da história, parte 1

Os esportivos australianos mais legais da história, parte 1 Dalmo Hernandes 3 fevereiro, 2017 O Quando a Holden encerrar suas atividades como fabricante de veículos, não será o fim apenas para os carros... [Leia mais](#)



Barrichello e Kubica em Le Mans, McLaren rompe tradição, polêmica em Daytona e mais!

Barrichello e Kubica em Le Mans, McLaren rompe tradição, polêmica em Daytona e mais! Projeto Motor 3 fevereiro, 2017 O Seja bem-vindo a mais uma edição do Racing News, com a seleção das principais notícias... [Leia mais](#)



Conheça os 10 carros mais vendidos no mundo em 2016

Conheça os 10 carros mais vendidos no mundo em 2016 ... 2 Minutos de Leitura O mercado mundial viu o Toyota Corolla emplacar nada menos que 1,3 milhão de unidades em 2016. Embora seja um número bem ex... [Leia mais](#)



Novo Cruze Hatch será lançado na China com câmbio de dupla embreagem



Dieselgate virou um bom negócio para muitos americanos – Compra de TDIs usados gera excelente lucro



Dodge Challenger Hellcat tem preço a partir de R\$ 700 mil



Ranking: empresas com melhor atendimento ao consumidor

Twitter A revista EXAME divulgou seu ranking EXAME/IBRIC de melhores

Recomendação

Customers who viewed this item also viewed these products



Dualit Food XL1500
Processor

\$560

Add to cart



Kenwood kMix Manual
Espresso Machine

★★★★☆

\$250

Select options



Weber One Touch Gold
Premium Charcoal
Grill-57cm

\$225

Add to cart



NoMU Salt Pepper and
Spice Grinders

\$3

View options

Recomendação

The screenshot shows the Netflix homepage. At the top left, the title "NETFLIX ORIGINAL HOUSE of CARDS" is displayed, followed by a five-star rating, the year 2015, the age rating 16+, and information about 3 Seasons and 51 episodes. Below this, a promotional image of Kevin Spacey as Frank Underwood sitting at a desk in an office is shown. To the right of the image is a "SEARCH & MENU" button. On the left side of the main content area, there's a section titled "Popular on Netflix" featuring thumbnails for "HOUSE of CARDS", "FULLER HOUSE", "SUITS", "NARCOS", and "BETTER CALL SAUL". Below this, another section titled "Top Picks for Takafumi" features thumbnails for "INCEPTION", "IN TIME", "THE GREAT GATSBY", and a "NETFLIX" promotional image.

NETFLIX ORIGINAL
HOUSE of CARDS

★★★★★ 2015 16+ 3 Seasons 51

Watch the Series

Is it true that absolute power corrupts absolutely? Congressman Frank Underwood absolutely intends to find out.

SEARCH & MENU

Popular on Netflix

HOUSE of CARDS

FULLER HOUSE

SUITS

NARCOS

BETTER CALL SAUL

Top Picks for Takafumi

INCEPTION

IN TIME

THE GREAT GATSBY

NETFLIX

Recomendação

SECTIONS HOME SEARCH

The New York Times

MEDIA

“o algoritmo deles mostrou que os filmes do **Kevin Spacey** eram muito vistos, assim como os **dirigidos por David Fincher**, e que uma série britânica dos anos 90 sobre os **bastidores sujos do Parlamento** tinha uma interessante legião de seguidores...”

[Fonte](#)

Giving Viewers What They Want



David Carr

THE MEDIA EQUATION

FEB. 24, 2013



[Fonte](#)

Recomendação



Seria Stranger Things uma obra de arte
do algoritmo da Netflix?

[Fonte](#)

Outros problemas relacionados

- Classificação de documentos
- Remoção de informação ruidosa (duplicatas, informação inútil)
- Coleta de dados na web
- Geração de resumos/extracão informação em texto
- Problemas de agrupamento (clustering)

Classificação

Notícias

- Últimas notícias
- Notícias próximas de...
- Sugerido para você

Mundo

Trump garante que bloqueio judicial a seu voto migratório será "cancelado"

Terra Brasil - há 50 minutos

O presidente dos Estados Unidos, Donald Trump, tachou neste sábado como "ridícula" a decisão de um juiz americano de suspender o voto migratório temporário que ele tinha imposto a sete países de maioria muçulmana e refugiados, e garantiu que ...

Pai do suspeito do ataque no Louvre diz que filho é inocente

Estado de Minas - há 1 hora

O pai do suspeito de ter atacado na sexta-feira um grupo de militares no Museu do Louvre, em Paris, assegurou neste sábado que seu filho é inocente e que não apresentava sinais de radicalização. Reda El Hamahmy, um general de polícia reformado, ...

Continente desaparecido há 200 milhões de anos é encontrado debaixo do Oceano Índico

Globo.com - 3 de fev de 2017

[G](#) [+](#) [M](#) [T](#) [f](#)

Duplicatas



Estado de Minas

Cobertura em
tempo real

[Novo presidente do Senado é acusado de receber propina para liberar MP para Odebrecht](#)

Estado de Minas - 2 de fev de 2017



Eleito nesta quarta-feira, dia 1º, com 61 votos presidente do Senado e do Congresso Nacional, o senador Eunício Oliveira (PMDB-CE) era conhecido entre os executivos da Odebrecht como o "Indio", apelido utilizado pelo chamado departamento de ...

Entenda como fica composição da Câmara e situação de Temer perante Legislativo [Blasting News](#)
[Que Congresso é esse?](#) [Istoe](#)

Detalhada: Eunício é novo presidente do Senado; veja quem vai disputar comando da Câmara [Terra Brasil](#)



[Portal de Notícias...](#)



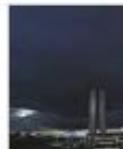
[Portal de Notícias...](#)



[Portal de Notícias...](#)



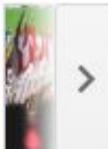
[Blasting Ne...](#)



[Istoe](#)



[O Nortão J...](#)



[Port...](#)



Dúvidas?

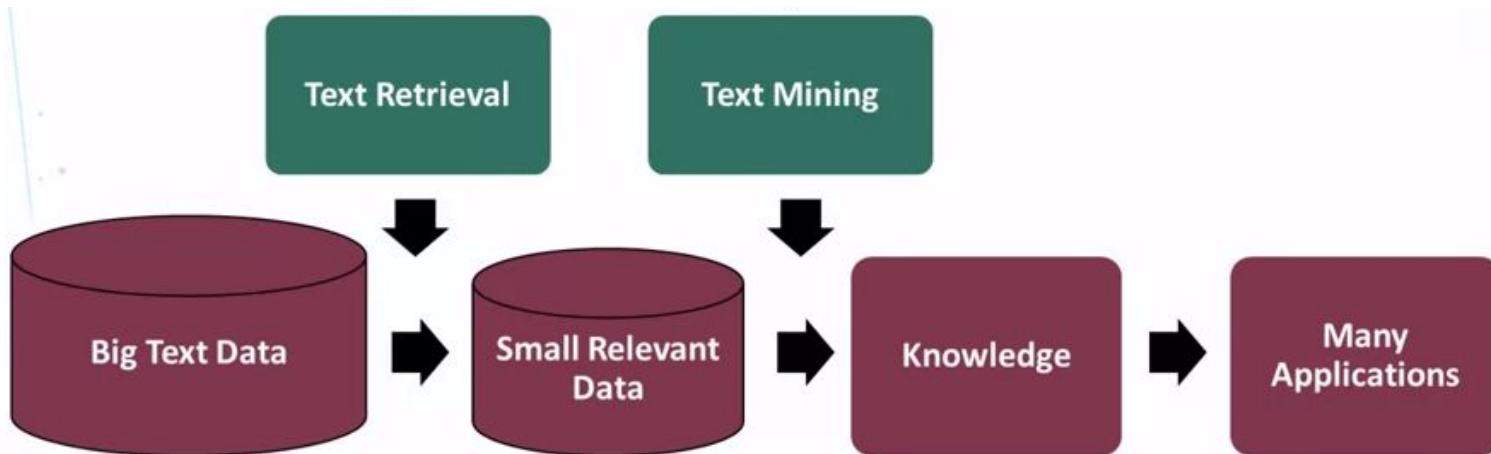
- O que é um Sistema de RI?
- Você poderia citar exemplos de problemas relacionados à área?

Dúvidas?

- O que é um Sistema de RI?
 - Trata da recuperação automática da informação contidas em grandes bases de dados
- Você poderia citar exemplos de problemas e soluções relacionadas?
 - Busca na web: Google, bing
 - Busca em sites institucionais e bibliotecas digitais: Lattes, wikis
 - Ecommerce e recomendação: Amazon, Submarino, Netflix...

Recuperação e Mineração de Textos

❖ Recuperação e Mineração



Inspiração para Projeto Final

Testes com a Língua Portuguesa





**No próximo slide:
Do que se trata o conjunto de documentos
recuperados da Web?**

Do que se trata?

sexos
teria discordar sexo
beljando vao dúvida apelidada direitos perfume
camp ala enfim cerveja revolta vcs concordar mim
chile alguém falava marca graça gerando prova tipo coerência malcontnt
colocar anticapitalista famosa mulher deu publicitária saco empresa mostrando publicidade boicote homofóbica
fizeram objeto ninguém casais causa pra continua conar preocupado atingiu
henrique desgosto reclamou polêmica pastormalafaia vídeo povo amigos gerou
colonogarner caralho juliano jedd verão vou polêmicas alguma gera
pediu vontade fuder

mãe amor proxima fala merda produto fosse
namorados assista galera post opinião unimed
dela... machista socorro mesmo olha manifestou sociedade filme
dê-lhe malafala deslike lucrou respeite entrei
engajadinha rentável mostra diferentes dela

fragrâncias sofrido estao comprar promove gay encheu marciomr10 itaipava mkreuz
gosto garoto telefone parar fazendo encherem fique incrível lançou comprei
governo intense tio pessoas falar inclusive incapacidade começo
instrumental lorigomessilva tirar lulu novas agr produtos barbarafpavan gays cade calmo impressionada encaretamos chega

manelra parecido ama nadyagospel acesse viu diferenças desboicote boicotar
mau pois criaram música de... brasileira ameaças
nego progressista deviam trivago vários
reclamando smileforbleber santos

boticário propaganda campanha



**O que mais chamou a atenção?
Qual o clima da discussão?**

Fluxo para Prática - Enriquecimento

- *Data Analysis - Boticario*
- Enriquecer palavras com tags semânticas
- Detectar entidades
- Configurar dicionários de entidades
- Análise de sentimento

O que mais chamou a atenção?

❖ Filtragem de Adjetivos e Verbos nos Tweets

governo
graça! 11 falava
[homossexual Unimed mulheres sutiã dê-lhe
começou acho gerando pra olha sociedade...
cerca polêmica encherem mim
inclusive encare tamos garoto presentes
gostei interfere falar casais tizeram povo comprar
instrumental marca fique ninguém [REDACTED]
mau nego anticapitalista
mostrando

O que mais chamou a atenção?

❖ Filtragem de Adjetivos e Verbos nos Tweets

governo
graça! 11 falaya
homossexual Unimed mulher sutíldê-lhe
começou acho gerando pra olha sociedade...
cerca polêmica encherem mim
inclusive encaretamos garoto presentes
gostei interfere falar casais fizeram povo comprar
instrumental marca fique ninguém [REDACTED]
mau nego anticapitalista
mostrando

O que mais chamou a atenção?



02/06/2015 18h05 - Atualizado em 03/06/2015 18h05

Propaganda de O Boticário com gays gera polêmica e chega ao Conar

Propaganda gerou reações homofóbicas e ameaças de boicote à marca. Em queixa ao Conar, consumidores consideraram comercial desrespeitoso.

Do G1, em São Paulo



[Vídeo](#)

Quais figuras são importantes na discussão?



Figuras importantes na discussão

- ❖ Detecção de Entidades: Pessoas, instituições, lugares...
-

U BUI I IUAKIU
Boticario O
Boticário
Itaipava O Boticário
Lulu Santos

Por quê? Existem mais referências?

❖ Lulu Santos?

➤ Busca + Mineração

The screenshot shows a Google search results page. The search query 'boticario lulu santos' is entered in the search bar. Below the search bar, there are tabs for Web, Images, Videos, News, Shopping, More, and Search tools. The 'Web' tab is selected. A message indicates 'About 235,000 results (0.35 seconds)'. The first result is a sponsored link from Boticario's website: 'Boticario - boticario.com.br' (Ad). It includes a link to 'www.boticario.com.br/' and promotional text: 'Compre na Loja Online O Boticário em até 5x s/Juros e Frete Grátis!* Entrega em todo o Brasil · até 5x sem Juros'. To the right of this result are two other links: 'Promoções O Boticário' and 'R\$110 = Frete Grátis' under 'Make B Lumina' and 'Perfumaria Masculina 20%'. The second result is a news article from rd1.ig.com.br titled '"1 a 0 na caretice", dispara Lulu Santos sobre comercial de ...' with a link to 'rd1.ig.com.br/1-a-0-na-caretice-dispara-lulu-santos-so...'. The third result is another news article from www.revistaforum.com.br titled 'Lulu Santos elogia propaganda com casais gays: "1 x 0 na ca...' with a link to 'www.revistaforum.com.br/.../lulu-santos-elogia-propa...'. Both news articles mention the date as Jun 8, 2015.

Por quê? Existem mais referências?

- ❖ Itaipava? Faz sentido?

The screenshot shows the header of the Brasil Econômico website. The logo 'Brasil Econômico' is at the top left, with 'Brasil' in yellow and 'Econômico' in black. To the right is a search bar with the word 'BUSCAR' and the text 'enhanced by Google'. Below the header is a navigation bar with links: CARREIRAS, EMPREENDEDORISMO, EMPRESAS, FINANÇAS PESSOAL, IMPOSTO DE RENDA, TECNOLOGIA, PRA FRENTES SEMPRE, and MAIS SITES.

Itaipava terá de mudar campanha por ser 'excessivamente provocativa'

Por Luis Philipe Souza - iG São Paulo | 18/06/2015 14:47 - Atualizada às 18/06/2015 18:00



Tamanho do texto + -

Duas peças de ponto de venda da campanha "Itaipava 100%"
foram consideradas 'excessivamente provocativas' pelo órgão

Right now.



Detectamos algum “movimento”?

Detecção de Entidades - Português

❖ Amplificadores



Sentimento - Português

- ❖ Sentimento -> É preciso refinar e se adaptar ao contexto



É possível usar tagcloud para **refinar** classe negativa e positiva de termos ao usar em análises seguintes

Busca por Entidades

- ❖ Exemplo de procura pelas principais companhias de saúde mais citadas na rede em determinado período
-



Análise de Sentimento

- ❖ Exemplo de sentimento das palavras-chave encontradas no texto



Trechos de Trabalhos

The screenshot shows a news website with a dark header bar. The top navigation includes links for 'PODER', 'BRASIL', 'MUNDO', 'ECONOMIA', 'DESENHOS 247', '2016', 'MÍDIA', 'SAÚDE 247', 'GÁSIS', and various regional sections like 'Alagoas 247', 'Amapá 247', 'Bolívia 247', 'Brasília 247', 'Ceará 247', 'Goiás 247', 'Maranhão 247', 'Mato Grosso 247', 'Paraná 247', 'Pernambuco 247', 'Ribeira 247', 'Rio Grande do Sul 247', 'SP 247', 'Sergipe 247', and 'Tocantins 247'. Below the header is a main menu with categories: Home, Brasil, Política, Finanças, Empresas, Agronegócios, Internacional, and Opinião. Sub-menus include Executive, Congresso, Estados e Municípios, Partidos, and Judiciário.

MOVIMENTOS DE HIP HOP SE UNEM CONTRA O GOLPE

Fiesp oferece filé mignon a manifestantes pró-impeachment na Paulista

governo sitiado

Citação de Dilma a impeachment de Lugo abre crise com Paraguai

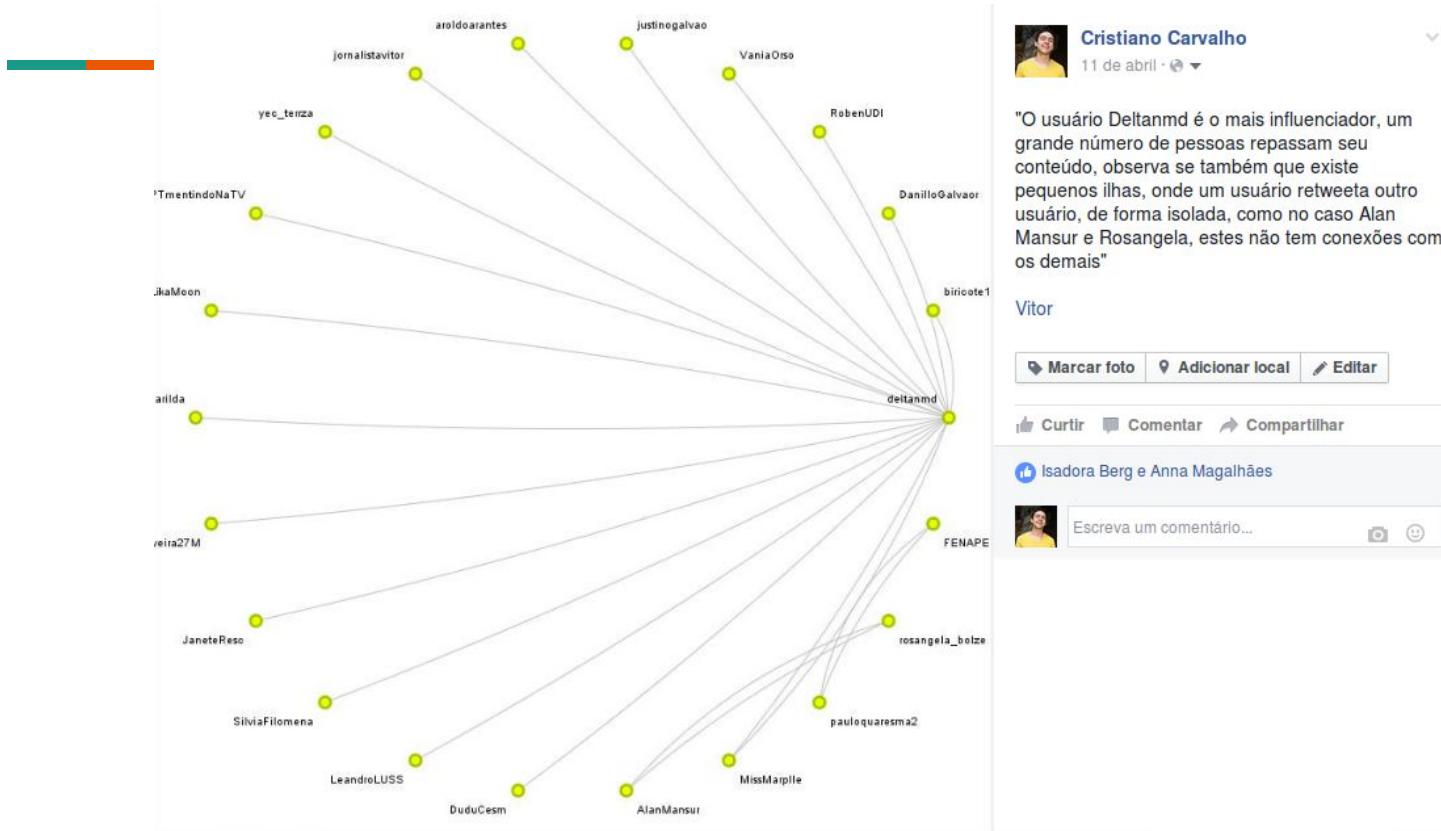
Filé mignon contra pão com mortadela

A central image is a political cartoon by William Guerreiro. It depicts a yellow rubber duck standing next to a small pile of food. A speech bubble from the duck says 'AMÉN'. The cartoon is signed 'WILLIAM GUERREIRO' at the bottom right.

“Na busca por #impeachmentjá encontramos a palavra “Mercosul” destoante das principais.

Na busca por #naovaitergolpe, a análise chamou a atenção por conter as palavras: hip hop, filé mignon e stones.”

Trechos de Trabalhos



Trechos de Trabalhos

Table "default" - Rows: 308 Spec - Columns: 3 Properties Flow Variables			
Row ID	GameName	PositiveNegative	Count*(Result)
Row2	7 Days to Die	Recommended	5
Row3	8BitBoy™	Recommended	5
Row4	A Lenda do Herói	Recommended	5
Row11	Amazing Frog?	Recommended	5
Row14	Angry Video Game Nerd II: ASSimilation	Recommended	5
Row25	Axiom Verge	Recommended	5
Row30	Ballistic Overkill	Recommended	5
Row31	Balrum	Recommended	5
Row34	Battle Brothers	Recommended	5
Row35	BattleBlock Theater®	Recommended	5
Row36	Besiege	Recommended	5

Jogos mais recomendados pelos usuários na categoria "Indie" considerando comentários na página.

Pedro

Marcar foto Adicionar local Editar

Curtir Comentar Compartilhar

Anna Magalhães



Escreva um comentário...



Trechos de Trabalhos

definitivamente amado imprensa crise esquecer preciso
odeio surpreender escuro paz gosto luz lindo chama amo
odiosa intimidade tipo pro respeito amante inferno
bonito crush vontade viver loucura simpatia
gloss problema travesti tenso sorrindo
sabedoria errado melhor pecado perda
infelizmente universal pena culpa saudade chorar superar



Cristiano Carvalho

11 de abril ·

""....Palavras como "pecado", "loucura" e "crush". Porém ao pesquisarmos o contexto em que essas palavras ocorrem, descobrimos que seu sentido não é negativo. Por exemplo, a palavra "pecado" aparece em tweets declarando que o usuário quer cometer "pecado" com o padre. Ainda, que sentem um "crush" pelo Padre. E que este vai a loucura no snap, com posts humorísticos....""

Isadora

Marcar foto Adicionar local Editar

Curtir Comentar Compartilhar

3

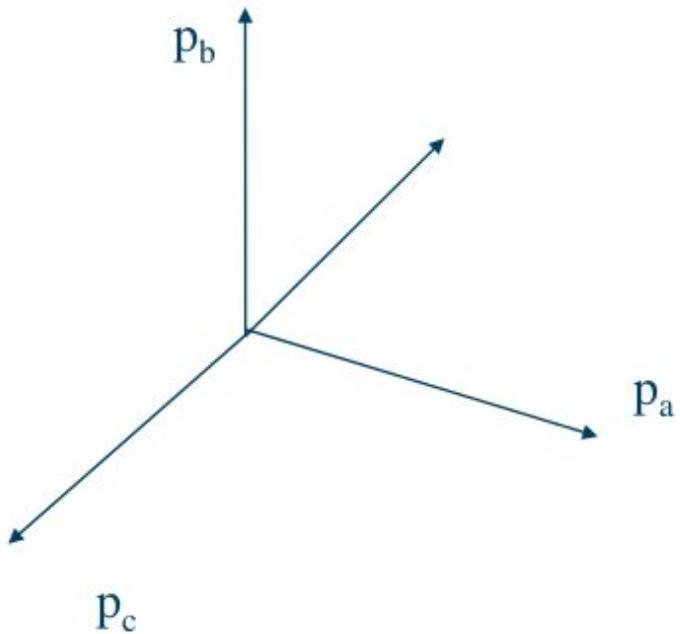
A disciplina - RI

❖ Plano de Ensino

- **Unidade 01:** Conceitos de inteligência competitiva e coletiva, crowdsourcing e redes sociais. Recuperação da informação e Máquinas de busca. Desafios da Mineração na web e nas redes. Exemplos de Projetos da disciplina.



Próxima aula...



Esse assunto era o que você imaginava?



A central image of a hand holding a small, wrapped gift box with a ribbon, positioned over a collage of international "thank you" greetings.

The text in the image includes:

- GRAZIE** (Italian)
- Mamana** (Portuguese)
- Chokrane** (Korean)
- DANK JE** (Dutch)
- ASANTE** (Swahili)
- Kiitos** (Finnish)
- Obrigado** (Portuguese)
- MATONDO** (Swahili)
- SPASIBO** (Russian)
- Kiitos** (Finnish)
- Obrigado** (Portuguese)
- DANK JE** (Dutch)
- Kia Ora** (Maori)
- SPASIBO** (Russian)
- DANK JE** (Dutch)
- ASANTE** (Swahili)
- Mamana** (Portuguese)
- Grazie** (Italian)
- Chokrane** (Korean)
- Kiitos** (Finnish)
- Nirringrazzjak** (Swahili)
- SPASIBO** (Russian)
- Mochchakkeram** (Khmer)
- ASANTE** (Swahili)
- Obrigado** (Portuguese)
- SPASIBO** (Russian)
- Kiitos** (Finnish)
- Matondo** (Swahili)
- Matur Nuwun** (Burmese)
- Chokrane** (Korean)
- Raibh Maith Agat** (Irish)
- Arigato** (Japanese)
- Multumesc** (Romanian)
- Mochchakkeram** (Khmer)
- MAAKE** (Swahili)
- SPASIBO** (Russian)
- Kiitos** (Finnish)
- DANK JE** (Dutch)
- Salamat** (Filipino)
- Matur Nuwun** (Burmese)
- Chokrane** (Korean)
- Raibh Maith Agat** (Irish)
- Kiitos** (Finnish)
- Spasibo** (Russian)
- ASANTE** (Swahili)
- Merci** (French)
- DANK JE** (Dutch)
- SPASIBO** (Russian)
- Mochchakkeram** (Khmer)
- Chokrane** (Korean)
- Grazie** (Italian)
- Raibh Maith Agat** (Irish)
- Terma Kasih** (Indonesian)
- Dankon** (Lao)
- Maake** (Swahili)
- Ua Tsang Rau Koj** (Lao)
- Obrigado** (Portuguese)
- DANK JE** (Dutch)
- Mochchakkeram** (Khmer)
- Raibh Maith Agat** (Irish)
- Obrigado** (Portuguese)