

Ciência de Dados e Repositórios: identificando soluções

Jefferson Lima

Tecnologista em Saúde Pública

Plataforma de Ciência de Dados aplicada à Saúde (bigdata.icict.fiocruz.br)



The screenshot shows the login page for the PCD S platform. The header features the ICICT logo and the text 'Instituto de Comunicação e Informação Científica e Tecnológica em Saúde'. The main content area has a dark blue background with a circuit-like pattern. On the left, the 'PCD S' logo is displayed, followed by the text 'Plataforma de Ciência de Dados aplicada à Saúde'. On the right, there is a login form titled 'Acesse a Plataforma' with input fields for 'usuário' and 'senha', a 'Recuperar senha' link, and 'LOGIN' and 'CRIAR CONTA' buttons. A navigation bar at the bottom contains links for 'HOME', 'SOBRE NÓS', 'A PLATAFORMA', 'CONJUNTOS DE DADOS', 'DATA SCIENCE LAB', 'INSTITUIÇÕES PARCEIRAS', and 'CONTATO'.



Plataforma de Ciência de Dados aplicada à Saúde (bigdata.icict.fiocruz.br)



SOBRE NÓS ▾

A PLATAFORMA ▾

CONJUNTOS DE DADOS

DATA SCIENCE LAB ▾

INSTITUIÇÕES PARCEIRAS

CONTATO

Interface Tecnológica

Interface Tecnológica

Você pode acessar os serviços da PCDaS por meio dos ícones abaixo



Análise Visual

Indexação, extração e análise visual de dados



Mineração de Dados e Análise Preditiva

Conexão aos dados da Plataforma via R Studio Server



Data Science Lab

Inovação e Aprendizagem Colaborativa

Plataforma de Ciência de Dados aplicada à Saúde (bigdata.icict.fiocruz.br)

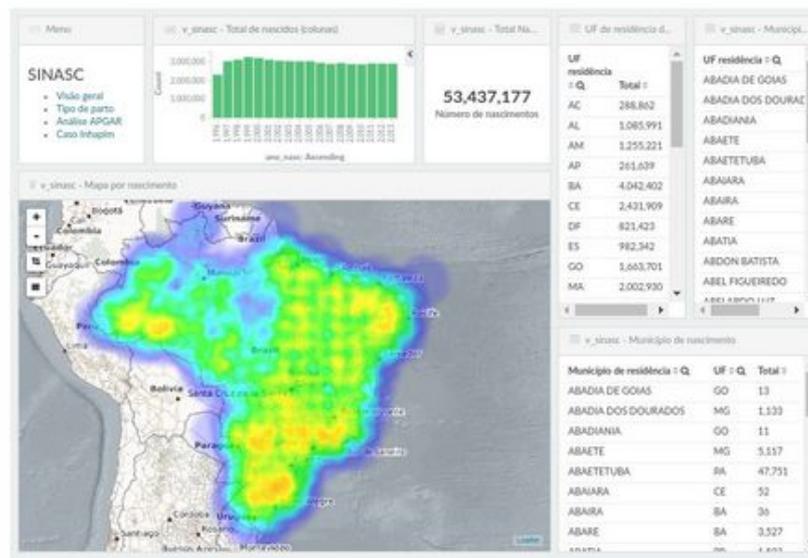


Análise Visual

Indexação, extração e análise visual de grandes quantidades de dados do setor saúde e seus determinantes socioambientais

Acesso qualificado aos dados da Plataforma

- Indexação e disponibilização de grandes bases de dados
- Extração de subconjuntos de dados de interesse para os pesquisadores
- Análise visual de situações de saúde
- Processamento distribuído e escalável



Plataforma de Ciência de Dados aplicada à Saúde (bigdata.icict.fiocruz.br)

Projetos de Pesquisa e Desenvolvimento Tecnológico

 <p>Repositório Institucional da Fiocruz</p> <p>Ciência de Dados aplicada ao Arca Ciência de Dados aplicada ao Repositório Institucional da Fundação Oswaldo Cruz - Arca</p>	 <p>Observatório em Ciência, Tecnologia e Inovação em Saúde Visa contribuir para a gestão e formulação de políticas institucionais em ciência, tecnologia e inovação.</p>	 <p>BASIS - Breastfeeding Information System Avaliação de impacto nos indicadores de morbimortalidade neonatal de iniciativas hospitalares pró-aleitamento materno</p>	 <p>Observatório de Epidemiologia Nutricional Desenvolver, testar e avaliar a aplicabilidade de novas recomendações de GPG a partir das curvas do Intergrowth-21st</p>
 <p>CEAG Centro de Estudos Avançados de Governo e Administração Pública</p> <p>Ciência de Dados aplicada à Políticas Públicas Plataforma de Ciência de Dados aplicada à Políticas Públicas</p>	 <p>Fale Conosco da Fiocruz O Fale Conosco é o canal de comunicação da Fiocruz com o cidadão. Seu compromisso é prezar a integridade, transparência e imparcialidade no esclarecimento das dúvidas do cidadão</p>	 <p>Redes de Cuidado Utilização de informações de rotina do SUS para mapear as redes de deslocamento de pacientes com câncer para a realização de tratamento, nos últimos 10 anos</p>	 <p>Extreme Data Lab - DEXL/LNCC Pesquisa e desenvolvimento de algoritmos de machine learning aplicados à saúde</p>

Ciência de Dados



Fonte da imagem: <https://bigdata.icict.fiocruz.br/ciencia-de-dados-aplicada-saude>

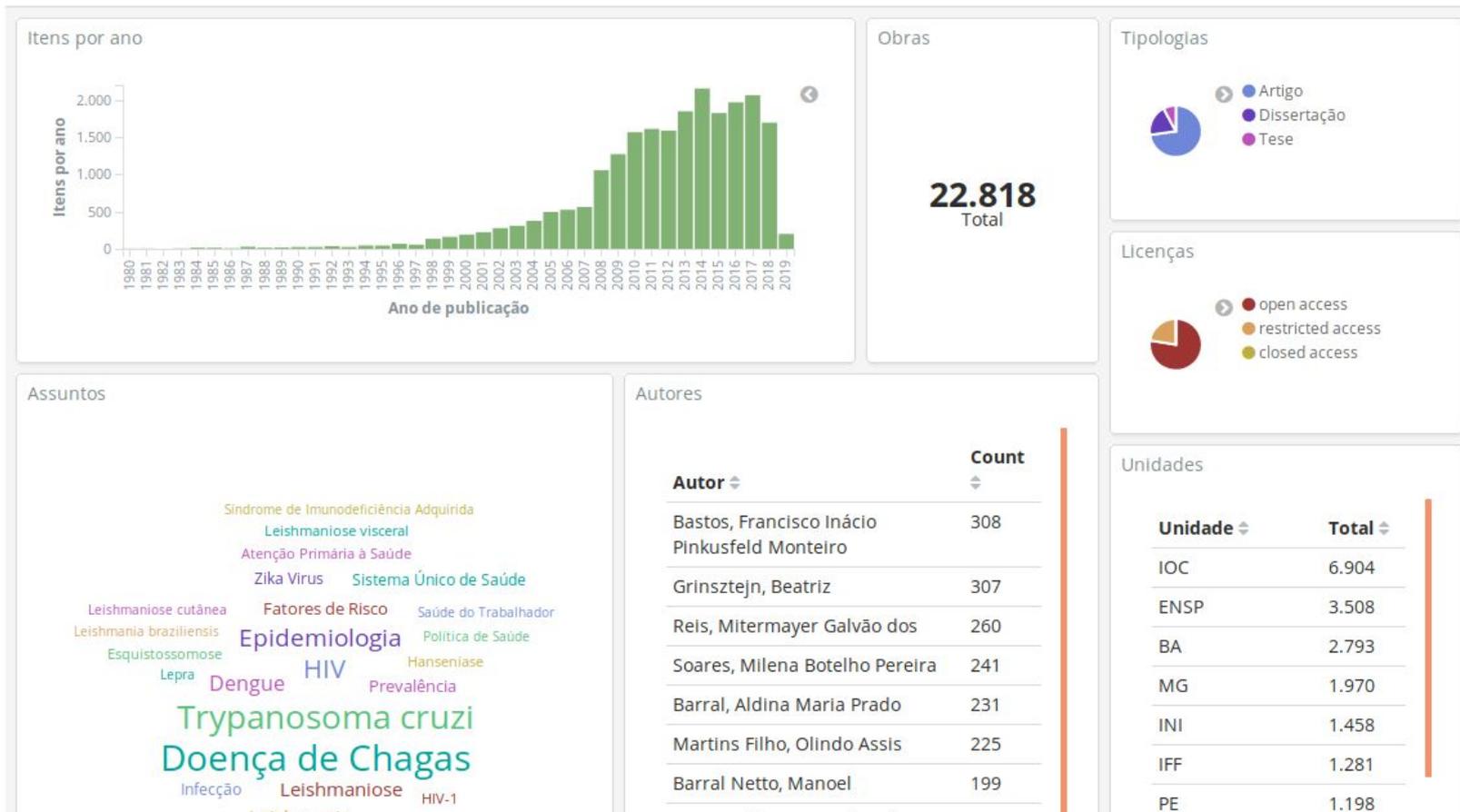
Ciência de Dados

- Inteligência artificial;
- Machine Learn;
- Big Data;
- Visualização da Informação;
- Mineração de Dados;
- ...

Parceria PCDaS e ARCA

- Visualização da Informação;
- Curadoria dos dados;
- Mineração de Textos.

Parceria PCDaS e ARCA - Visualização da Informação



Parceria PCDaS e ARCA - Curadoria dos dados

Idioma ↕	Obras ↕
eng	12.410
por	10.378
spa	140
Português (Br)	11
fra	7
Português	1

É importante garantir a legibilidade por máquina!

Parceria PCDaS e ARCA - Mineração de Textos (alguns exemplos)

- Identificação de Clusters de documentos;
- Identificação de conceitos correlacionados;
- Extração de Entidades Nomeadas;
- Sistemas de recomendação;
- Identificação automática de descritores.

Mineração de Textos: Identificação de Clusters de documentos

O GLOBO MENU

SOCIEDADE ▾

26/09/2017

TECNOLOGIA

FGV usa inteligência artificial para agilizar trabalho de historiadores

Algoritmo faz em poucos dias tarefa que levaria décadas

POR SÉRGIO MATSUURA

26/09/2017 16:08 / atualizado 26/09/2017 16:16



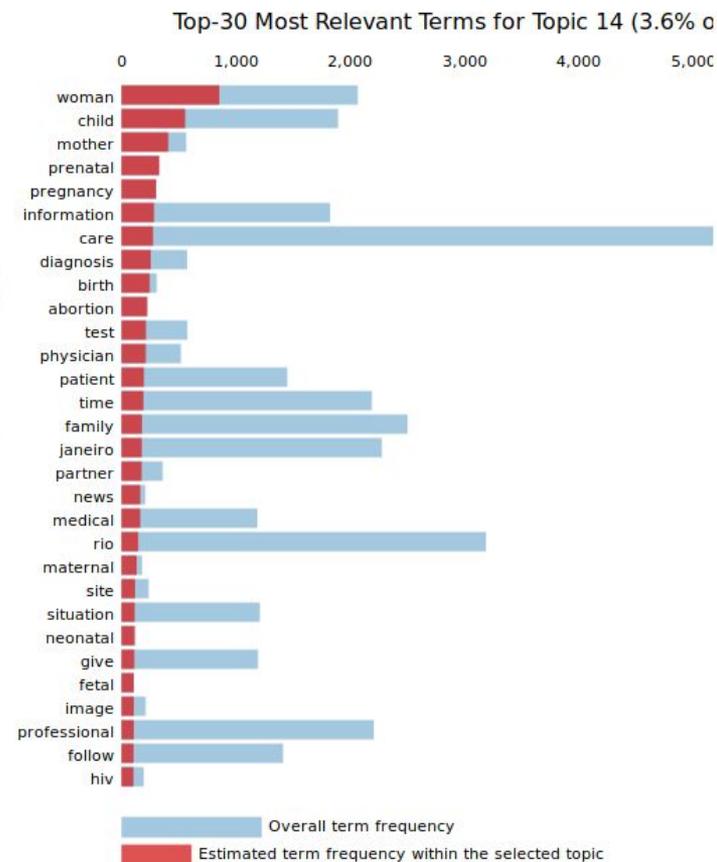
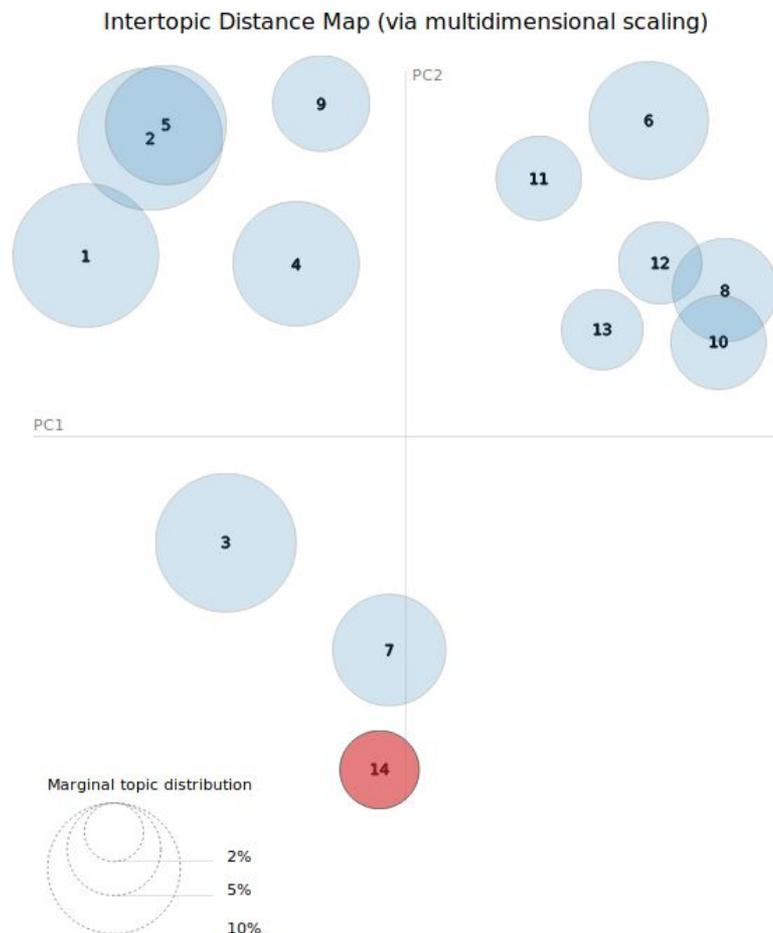
Ferramenta criada pela Universidade Columbia, em parceria com a FGV, transforma documentos em dados - HISTORY-LAB

— Eu fiquei surpreendido. A máquina oferece grupos de documentos com uma coesão temática que só um perito poderia identificar — disse Moreli. — E seria uma tarefa que um historiador levaria décadas.

Alexandre Moreli (CPDOC/FGV)

<https://oglobo.globo.com/economia/fgv-usa-inteligencia-artificial-para-agilizar-trabalho-de-historiadores-21872633>

Mineração de Textos: Identificação de Clusters de documentos

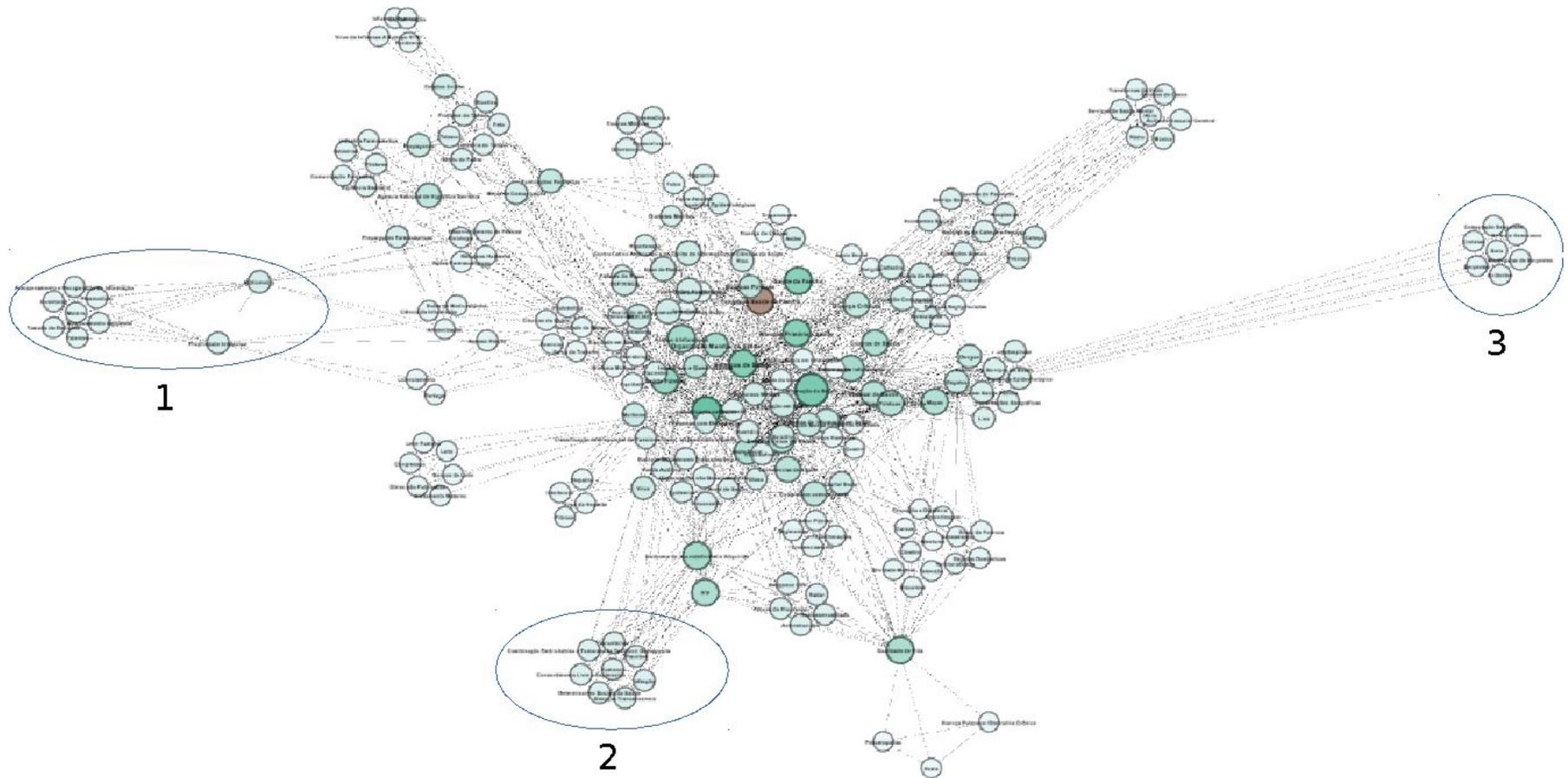


1. $saliency(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for
2. $relevance(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Six

Mineração de Textos: Identificação de Clusters de documentos

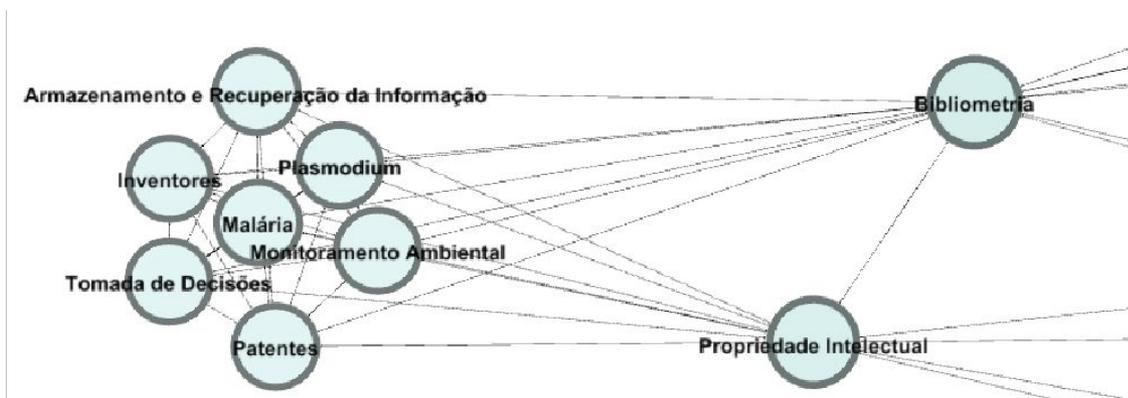


Mineração de Textos: Identificação de conceitos correlacionados

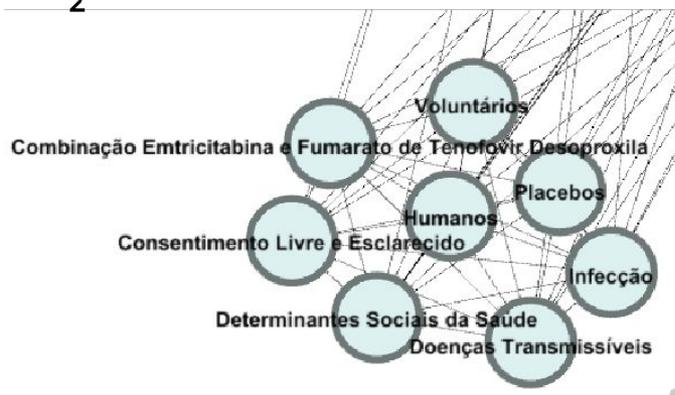


Mineração de Textos: Identificação de conceitos correlacionados

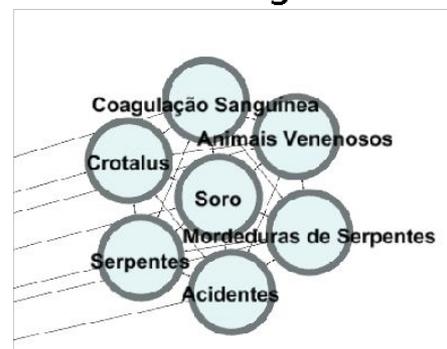
1



2



3



Mineração de Textos: Extração de Entidades Nomeadas

This note is an analysis of the communication produced by the global mining company Vale and circulated in the Paraopeba Valley's region, which includes the municipalities of Congonhas, Belo Vale and Brumadinho. The material analyzed was distributed in 2018 after the disaster caused by Vale in the municipality of Mariana (MG). In the study, it was hypothesized the construction of the public image of Vale in front of the Mariana and the recent disaster in Brumadinho (MG) was based on the euphemism semantics, a speech modality which favours the image idealized by the company about itself.

<https://dandelion.eu/semantic-text/entity-extraction-demo/>

Mineração de Textos: Extração de Entidades Nomeadas

This note is an **analysis** of the **communication** produced by the global **mining** company Vale and circulated in the **Paraopeba** Valley's region, which includes the municipalities of **Congonhas**, **Belo Vale** and **Brumadinho**. The material analyzed was distributed in 2018 after the disaster caused by Vale in the **municipality** of Mariana (MG). In the study, it was hypothesized the construction of the **public image** of Vale in front of the Mariana and the recent disaster in **Brumadinho** (MG) was based on the **euphemism semantics**, a **speech modality** which favours the image idealized by the company about itself.

0 persons

0 works

0 organisations

3 places

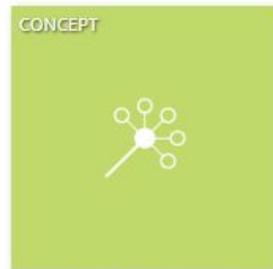
0 events

10 concepts



Analysis

TOP ENTITY



Communication

TOP ENTITY



Mining



Paraopeba



PLACE



PLACE



PLACE

TOP I



CONCEPT

TOP I

Mineração de Textos: Sistemas de recomendação

Netflix oferece US\$1 milhão para quem criar sistema de recomendação mais eficiente

21/09/2009 - POR | [NOTÍCIA](#)



A companhia de locação de filmes online Netflix [está reeditando](#) uma experiência bem-sucedida: um concurso em que premia com US\$1 milhão quem criar um sistema de recomendação mais eficiente. A empresa anunciou hoje o vencedor da primeira edição.

<http://colunas.revistaepocanegocios.globo.com/tecneira/2009/09/21/netflix-oferece-us1-milhao-para-quem-criar-sistema-de-recomendacao-mais-eficiente/>

Mineração de Textos: Identificação automática de descritores

Dissertação:

Análise lexicográfica da produção acadêmica da Fiocruz: uma proposta de metodologia

<https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/17458/JeffersonLima-Disserta%C3%A7%C3%A3o.pdf>

Mineração de Textos: Identificação automática de descritores

Motivação:

1. O crescimento exponencial de textos em formato digital;
2. É razoável afirmar que a classificação de conteúdos não é uma ciência exata;
3. Mesmo nos casos em que ainda seja possível a classificação manual dos conteúdos, há um caráter dinâmico ligado aos descritores que não costuma ser capturado;
4. É uma tarefa difícil prever as necessidades futuras dos usuários.

Mineração de Textos: Identificação automática de descritores

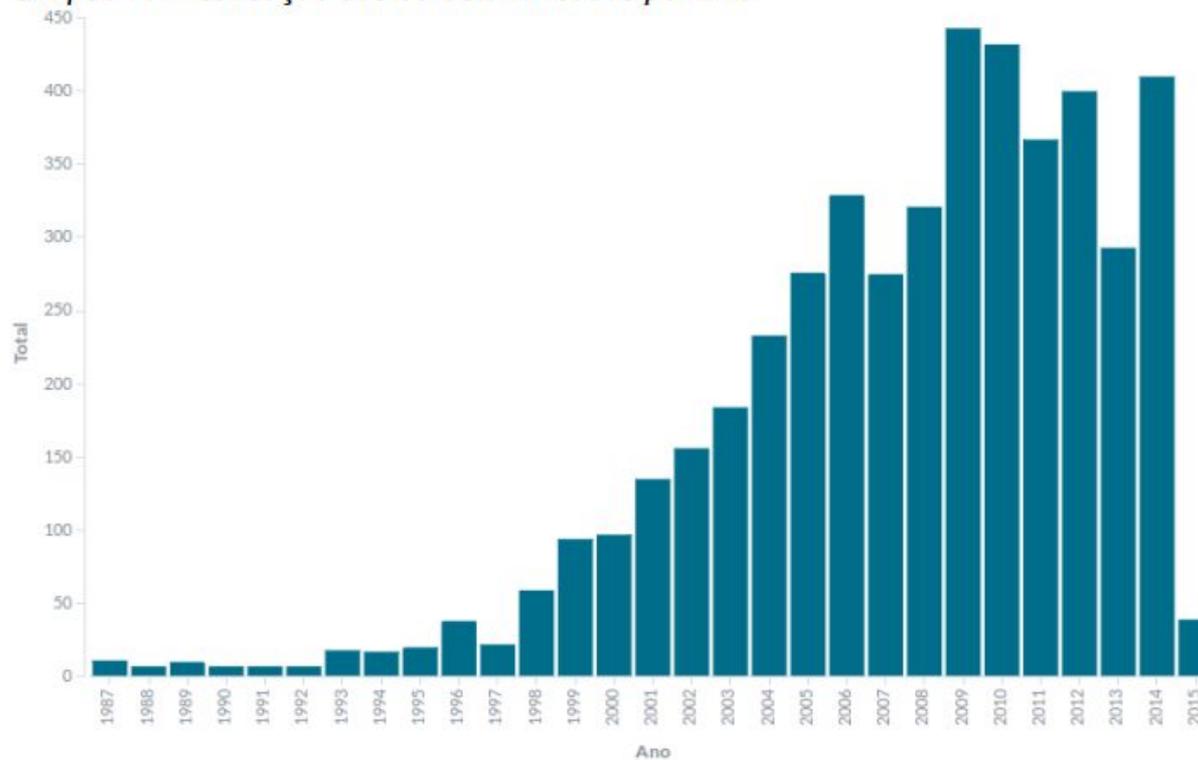
Etapas:

1. Obtenção do Corpus;
2. Pré-processamento dos documentos;
3. Captura de dados do vocabulário Descritores em Ciências da Saúde (DeCS);
4. Cruzamento entre n-grams e o DeCS para a identificação de descritores para os documentos.

* <https://www.loc.gov/about/general-information/> e <https://www.loc.gov/about/fascinating-facts/>

Mineração de Textos: Identificação automática de descritores

Gráfico 1: Distribuição das obras analisadas por ano



<https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/17458/JeffersonLima-Disserta%C3%A7%C3%A3o.pdf>

Mineração de Textos: Identificação automática de descritores

Título		Vigilância em saúde dos trabalhadores: potencialidades da matriz FPEEEA
Descritores	Cadastrados manualmente	atenção primária à saúde, saúde do trabalhador, mineração, matrix, brasil, <i>mining</i> , <i>worker's health surveillance</i> , -indicadores, matriz, <i>worker's health</i> , vigilância em saúde do trabalhador
	Identificados automaticamente	atenção primária à saúde, saúde do trabalhador, mineração, vigilância em saúde pública, trabalhadores, silicose, técnicas de planejamento, vigilância em saúde do trabalhador
	Em comum	vigilância em saúde do trabalhador, saúde do trabalhador, mineração, atenção primária à saúde

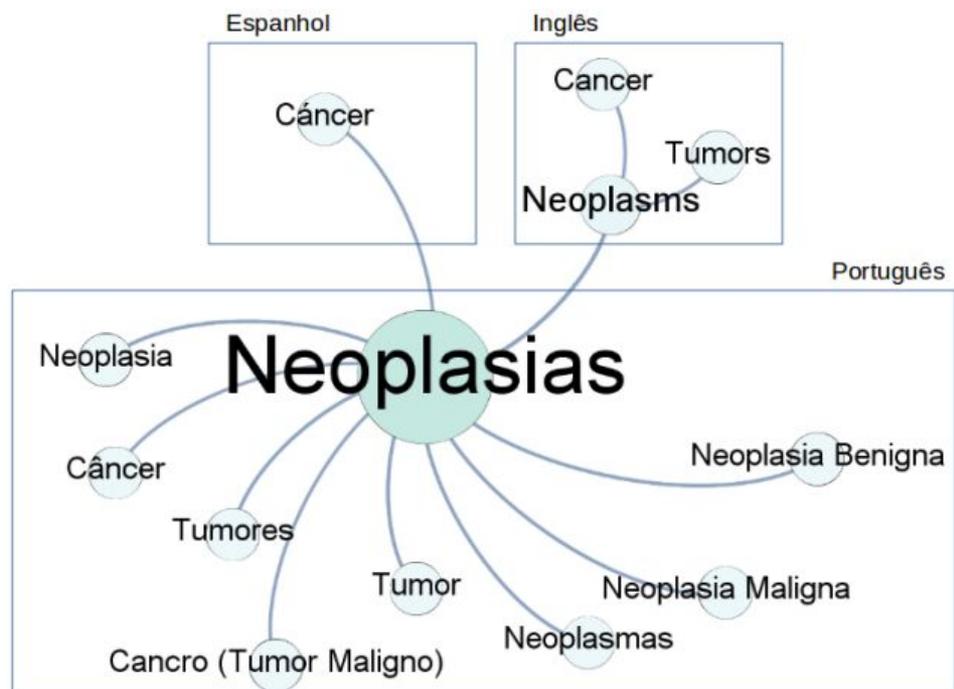
<https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/17458/JeffersonLima-Disserta%C3%A7%C3%A3o.pdf>

Mineração de Textos: Identificação automática de descritores

Título	A construção da intersetorialidade no Programa Bolsa Família em Manguinhos, no Rio de Janeiro
Descritores cadastrados manualmente	Ação intersetorial, Programas e Políticas de Nutrição e Alimentação, Programa Saúde da Família, Descentralização, Política Social, Pobreza, Estudos de Casos, Brasil, Public policy, Social policy, Intersectoral action
Descritores identificados automaticamente	Ação Intersectorial, Conhecimentos, Atitudes e Prática em Saúde, Políticas Públicas, Fome, Segurança Alimentar e Nutricional, Assistência Social, Governo Federal, Imunoglobulina D, Cadastro
Em comum	-

<https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/17458/JeffersonLima-Disserta%C3%A7%C3%A3o.pdf>

Mineração de Textos: Identificação automática de descritores (aumento da revocação)



<https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/17458/JeffersonLima-Disserta%C3%A7%C3%A3o.pdf>

Se a internet deixar...



Nossa contribuição:



Contato:

Jefferson Lima
Tecnologista em Saúde Pública

jefferson.lima@icict.fiocruz.br



Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

www.facebook.com/fiocruz.icict

[twitter.com/@Icict_fiocruz](https://twitter.com/Icict_fiocruz)

www.youtube.com/videosaudefio

www.icict.fiocruz.br