



LUPA DIGITAL

Hugo Veríssimo

ENQUADRAMENTO

┌ O PROBLEMA

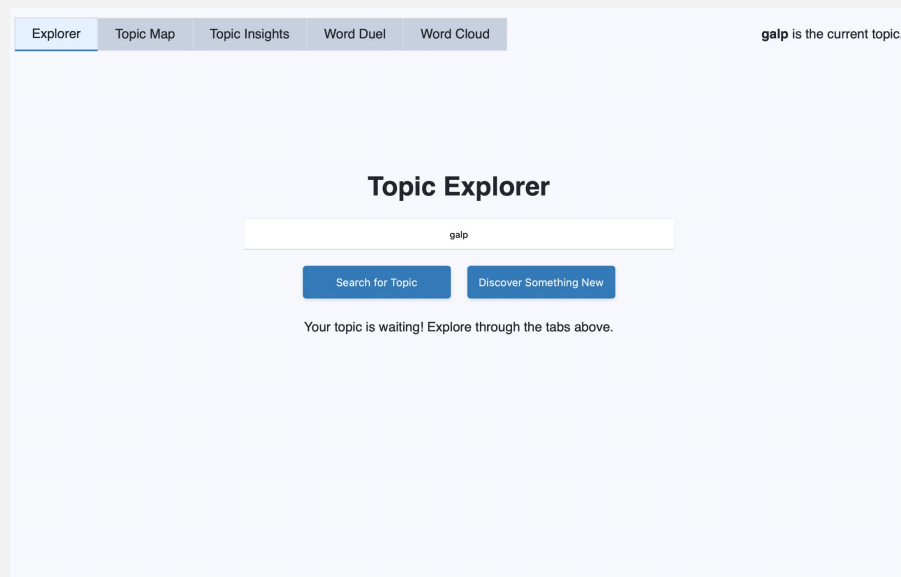
- Explosão de informação digital dificulta a identificação de padrões, fontes e tendências.
- Falta de ferramentas acessíveis para análise escalável e transparente do discurso público.



A PROPOSTA ┐

- Criar uma plataforma de metajornalismo inteligente, interativa e intuitiva.
- Analisar padrões jornalísticos em larga escala.
- Articular ciência de dados, jornalismo e ciência da informação.

MOTIVAÇÃO



5 empresas do PSI-20



1 457 234 tópicos

Perceção do potencial de generalização para qualquer tópico noticioso.

ARQUITETURA E METODOLOGIA

Recolha de ligações

Extração de texto

Mineração de dados

Lupa Digital

 **arquivo.pt**
(CDX-server API)

20 fontes
de informação



3 056 418 URLs,
desde 1998 até à atualidade

ARQUITETURA E METODOLOGIA

Recolha de ligações

Extração de texto

Mineração de dados

Lupa Digital

3 056 418 URLs

 **arquivo.pt**
(páginas arquivadas)

 **Spark**
APACHE

BeautifulSoup

 **learn**
scikit

349 519 textos,
únicos e válidos

ARQUITETURA E METODOLOGIA

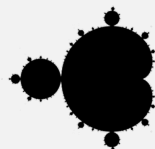
Recolha de ligações

Extração de texto

Mineração de dados

Lupa Digital

349 519 textos



TextBlob

spaCy

349 519 registos

Data de publicação

Fonte de informação

Perceção do sentimento

Tópicos mencionados

ARQUITETURA E METODOLOGIA

Recolha de ligações

Extração de texto

Mineração de dados

Lupa Digital

349 519 registos



LupaDigital

LUPA DIGITAL

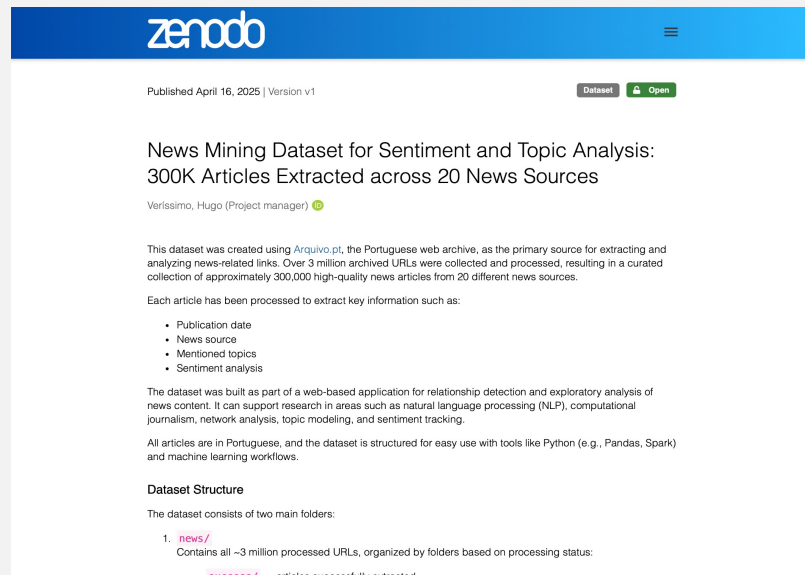
Recolha de ligações

Extração de texto

Mineração de dados

Lupa Digital

Conjunto de dados



The screenshot shows the Zenodo dataset page for the "News Mining Dataset for Sentiment and Topic Analysis: 300K Articles Extracted across 20 News Sources". The page includes the Zenodo logo, publication date (April 16, 2025), version (v1), and a brief description of the dataset. It also lists the project manager, Verissimo, Hugo, and provides a list of key information extracted from each article, such as publication date, news source, mentioned topics, and sentiment analysis. The dataset structure is also detailed, showing two main folders: "news/" and "surveys/".

Published April 16, 2025 | Version v1

Dataset Open

News Mining Dataset for Sentiment and Topic Analysis: 300K Articles Extracted across 20 News Sources

Verissimo, Hugo (Project manager)

This dataset was created using *Arquivo.pt*, the Portuguese web archive, as the primary source for extracting and analyzing news-related links. Over 3 million archived URLs were collected and processed, resulting in a curated collection of approximately 300,000 high-quality news articles from 20 different news sources.

Each article has been processed to extract key information such as:

- Publication date
- News source
- Mentioned topics
- Sentiment analysis

The dataset was built as part of a web-based application for relationship detection and exploratory analysis of news content. It can support research in areas such as natural language processing (NLP), computational journalism, network analysis, topic modeling, and sentiment tracking.

All articles are in Portuguese, and the dataset is structured for easy use with tools like Python (e.g., Pandas, Spark) and machine learning workflows.

Dataset Structure

The dataset consists of two main folders:

1. **news/**
Contains all ~3 million processed URLs, organized by folders based on processing status:

doi.org/10.5281/zenodo.15231163

Aplicação web



The screenshot shows the Lupa Digital web application interface. It features a dark blue header with the Lupa Digital logo and navigation links (Início, Sobre, Grafo). The main content area has a large heading "A Lupa que Revela Conexões" and a subheading "A Lupa Digital permite explorar e analisar notícias arquivadas ao longo da era digital, desde 1998 até à atualidade. Descubra conexões e padrões entre mais de 1 500 000 tópicos, entidades e eventos." Below this is a search bar with the placeholder text "Procure sobre um tópico...". At the bottom, it mentions "Baseado em dados do arquivo.pt" with the Arquivo.pt logo.

LupaDigital

Início Sobre Grafo

A Lupa que Revela Conexões

A Lupa Digital permite explorar e analisar notícias arquivadas ao longo da era digital, desde 1998 até à atualidade. Descubra conexões e padrões entre mais de 1 500 000 tópicos, entidades e eventos.

Procure sobre um tópico...

Baseado em dados do **arquivo.pt**

lupa-digital.pt

CONJUNTO DE DADOS

Viabiliza:

- Integração de novas fontes e atualizações com notícias recentes.
- Estudos de viés, polarização e evolução narrativa com base em NLP.
- Análise comparativa entre fontes e tendências editoriais.
- Investigação do impacto de eventos históricos nas narrativas.

timestamp	source	archive	id	probability	keywords	sentiment	significant_keywords
202201	SAP0	https://arquivo.p...	2587516	0.503894	{ME0 Energia -> 1...	0.5882089495658874	[desporto, austri...
201708	RTP	https://arquivo.p...	668092	0.5042335	{casa -> 1, frota...	0.7012464761734009	[apontar, crescim...
201412	Jornal de Negócios	https://arquivo.p...	258624	0.52641004	{inglãs -> 1, cu...	0.6120871663093567	[governo, novo, b...
201602	RTP	https://arquivo.p...	386591	0.5151844	{casa -> 1, segun...	0.6212668776512146	[concurso, chavel]
201005	Público	https://arquivo.p...	140235	0.60642016	{UGT -> 11, Dague...	-0.28381410241127014	[ugt, movimento, ...]
201906	SAP0	https://arquivo.p...	1527555	0.531442	{STE -> 3, govern...	0.7262091994285583	[alteracao, model...
202103	Expresso	https://arquivo.p...	2374638	0.53970325	{governo -> 4, si...	-0.01	[elogio, antonio, ...]
201902	Dinheiro Vivo	https://arquivo.p...	1387482	0.58162355	{casa -> 4, Turis...	0.5689757227897644	[portugal, fonte, ...]
201908	RTP	https://arquivo.p...	1612263	0.51769644	{casa -> 2, gover...	0.4765915036201477	[sao vicente, mad...
201702	O Mirante	https://arquivo.p...	557856	0.53151506	{casa -> 1, refer...	0.8240159511566162	[alcanena]
202009	Jornal de Negócios	https://arquivo.p...	2207746	0.54178274	{saber -> 1, smar...	0.01	[reducao, contrac...
201906	RTP	https://arquivo.p...	1542776	0.5529164	{Kuciak -> 1, Mar...	0.5608742594718933	[atletico madrid, ...]
201802	Dinheiro Vivo	https://arquivo.p...	868760	0.55604887	{casa -> 4, ganhe...	0.435367739200592	[sinal, trabalho, ...]
201912	Jornal de Negócios	https://arquivo.p...	1770781	0.526249	{Conferência Negó...	0.5451344609260559	[der, aprovar, ne...
201707	Correio da Manhã	https://arquivo.p...	654373	0.5489004	{participe -> 1, ...}	-0.5889040231704712	[manha, represent...
201812	SAP0	https://arquivo.p...	1323728	0.60221225	{Sérvulo -> 1, mo...	0.6396513342857361	[dinamarca, preco...
202002	Público	https://arquivo.p...	1953421	0.64698744	{irregularidade -...	0.5414410591125488	[rui pinto, portu...
202303	Público	https://arquivo.p...	3034857	0.5487394	{Pet Fun -> 2, fr...	0.7260831356048584	[casa, comportame...
201702	RTP	https://arquivo.p...	539436	0.58189535	{governo -> 3, ap...	0.6064995288848877	[crescimento, fin...
202104	SIC Notícias	https://arquivo.p...	2402778	0.50842965	{crise -> 2, fóru...	-0.7334808111190796	[vitima, area, ef...

only showing top 20 rows

APLICAÇÃO WEB



Lupa Digital

[Início](#) [Sobre](#) [Grafo](#) [🔍](#)

A Lupa que Revela Conexões

A Lupa Digital permite explorar e analisar notícias arquivadas ao longo da era digital, desde 1998 até à atualidade. Descubra conexões e padrões entre mais de 1 500 000 tópicos, entidades e eventos.

Procure sobre um tópico... 🔍

Baseado em dados do  **arquivo.pt**



Banco de Portugal

Resultados de Procura

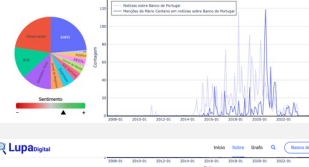
Foram encontrados 2267 resultados!



Banco de Portugal & Mário Centeno

Has 2267 notícias sobre Banco de Portugal, foram encontradas 177 menções em relação a Mário Centeno. Visualize os relacionamentos de notícias e tópicos associados ao tópico selecionado.

- ✓ Fontes de Informação
- ✓ Percepção de Sentimentos
- ✓ Evolução da Relação
- ✓ Notícias Associadas



Visualize as notícias associadas ao tópico selecionado e explore a rede de tópicos e entidades relacionadas.



IMPACTO SOCIAL

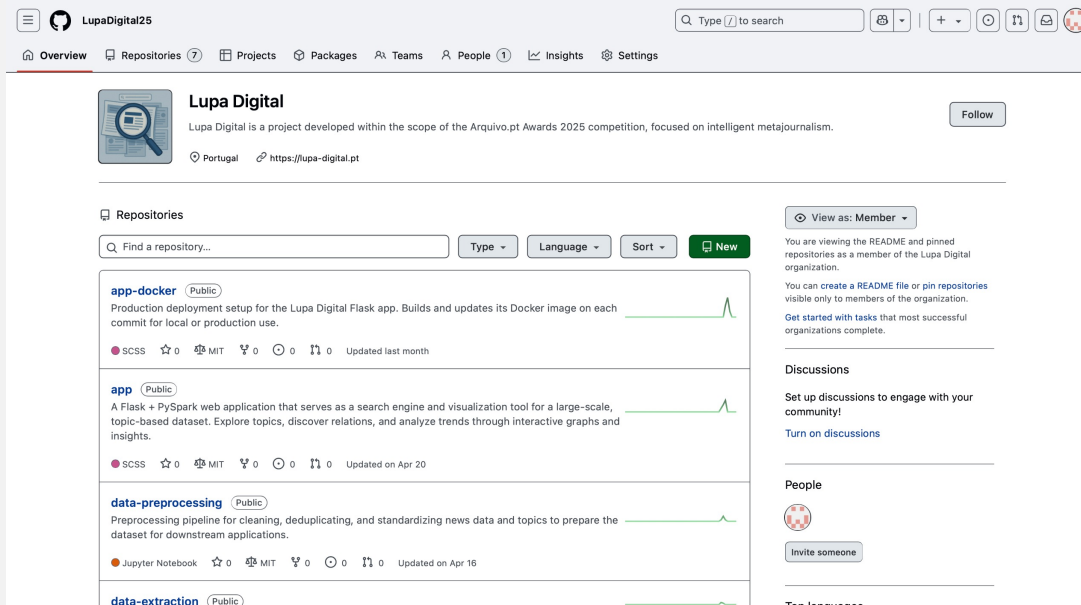


- Acesso facilitado e visual à evolução histórica de temas, entidades e discursos na imprensa portuguesa.
- Promoção da literacia mediática e da análise crítica do discurso jornalístico.
- Contribuição para a sustentabilidade democrática através da transparência informacional.
- Exploração de padrões e relações através do grafo interativo.
- Utilidade transversal: cidadãos, estudantes, jornalistas, etc.

- Demonstração prática de uma abordagem de metajornalismo inteligente escalável.
- Dados estruturados para estudos históricos, narrativos, de viés e enquadramento mediático.
- Estrutura modular e de código aberto permite:

Extensão contínua do conjunto de dados;

Integração com outros sistemas de informação.



IMPACTO CIENTÍFICO





LUPA DIGITAL

lupa-digital.pt

Agradecimentos

Prof. Alina Trifan

Cristiano Nicolau

Helena Veríssimo

João Cardoso

Lara Pereira

