ARQUIVO DE OPINIÃO

Miguel Won | Prémios Arquivo.pt 2018

SOBRE MIM

Doutorado na área da Física de Partículas (UC)

Investigador Pós-Doc no INESC-ID na área de Processamento de Língua Natural

Focado na análise de textos políticos:

- Debates parlamentares
- Artigos de Opinião

Artigos de Opinião

Publicações de autor, com ou sem periodicidade, que na sua generalidade comentam os mais variados assuntos políticos

São um reflexo da opinião pública mas também um espaço de influência da mesma

Componente essencial do debate público

Memória

Memória da discussão política é essencial ao bom funcionamento das democracias

A memória permite recordar o confronto de ideias, as lógicas argumentativas, bem como os posicionamentos dos diversos atores políticos (muitos cronistas são, foram ou serão eles próprios políticos ativos)

Dada a sua dimensão, é necessária a digitalização deste tipo de corpora

- Motor de busca
- Pesquisas por autor, período temporal ou fonte de publicação

Disponibilidade pública e de fácil utilização (user friendly)

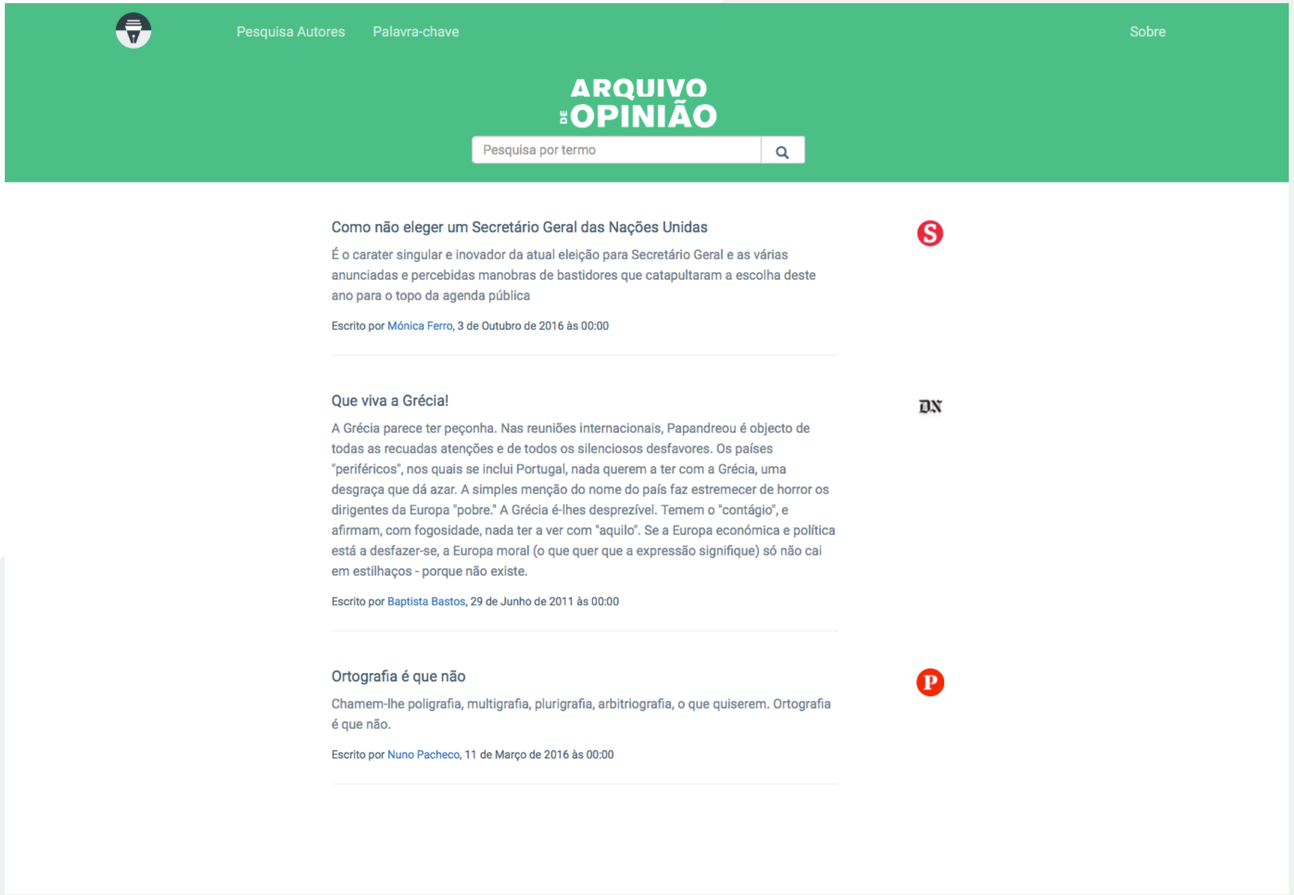
Arquivo de Opinião

Repositório online de artigos de opinião

Indexados mais de 80000 artigos publicados entre 2008 e 2016, nos principais jornais portugueses

Identificados mais de 3000 autores

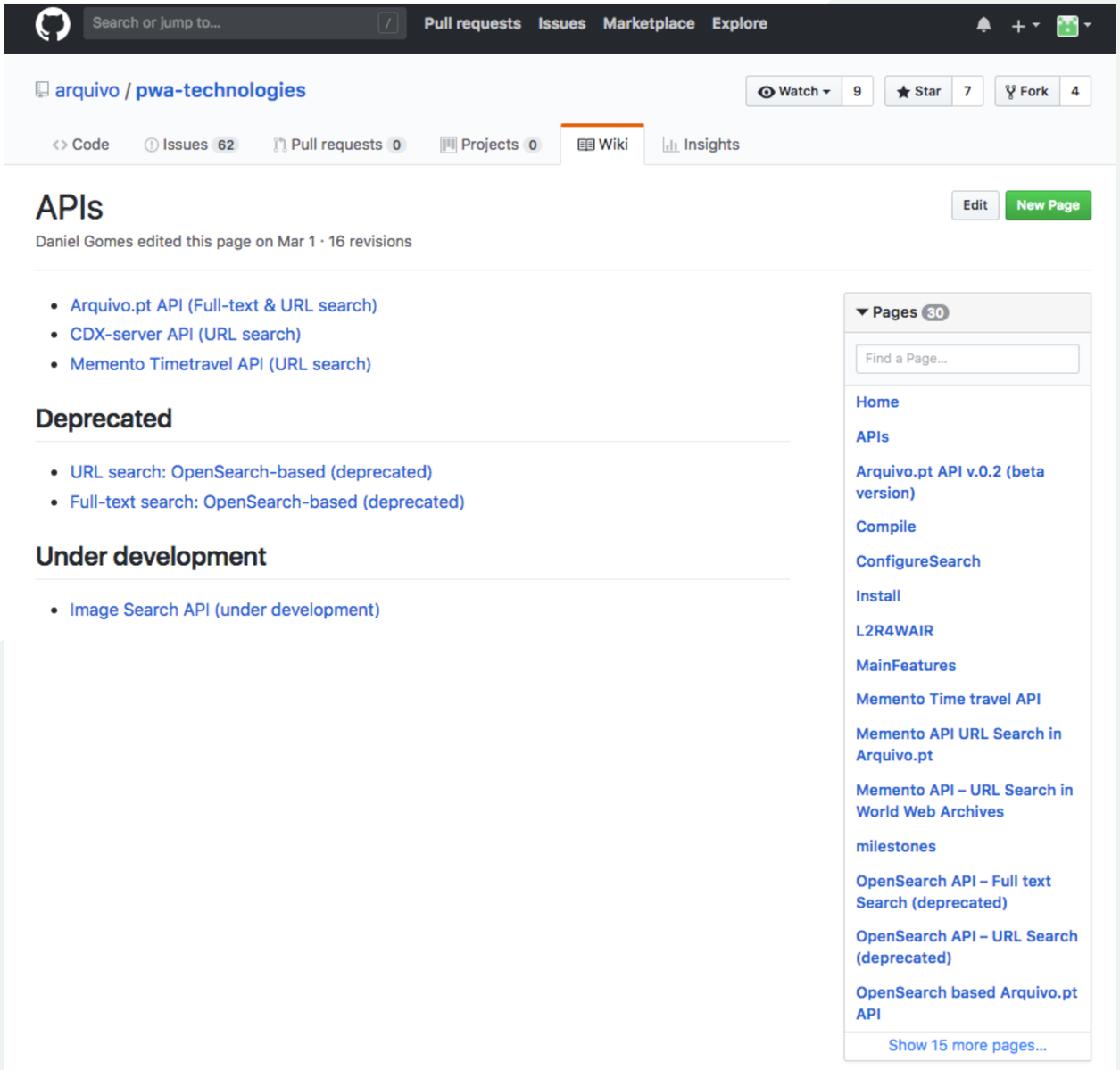
Reconhecidas e indexadas mais de 30000 palavras-chave



Fontes

- Arquivo.pt (API)

- Secção de Opinião (online)

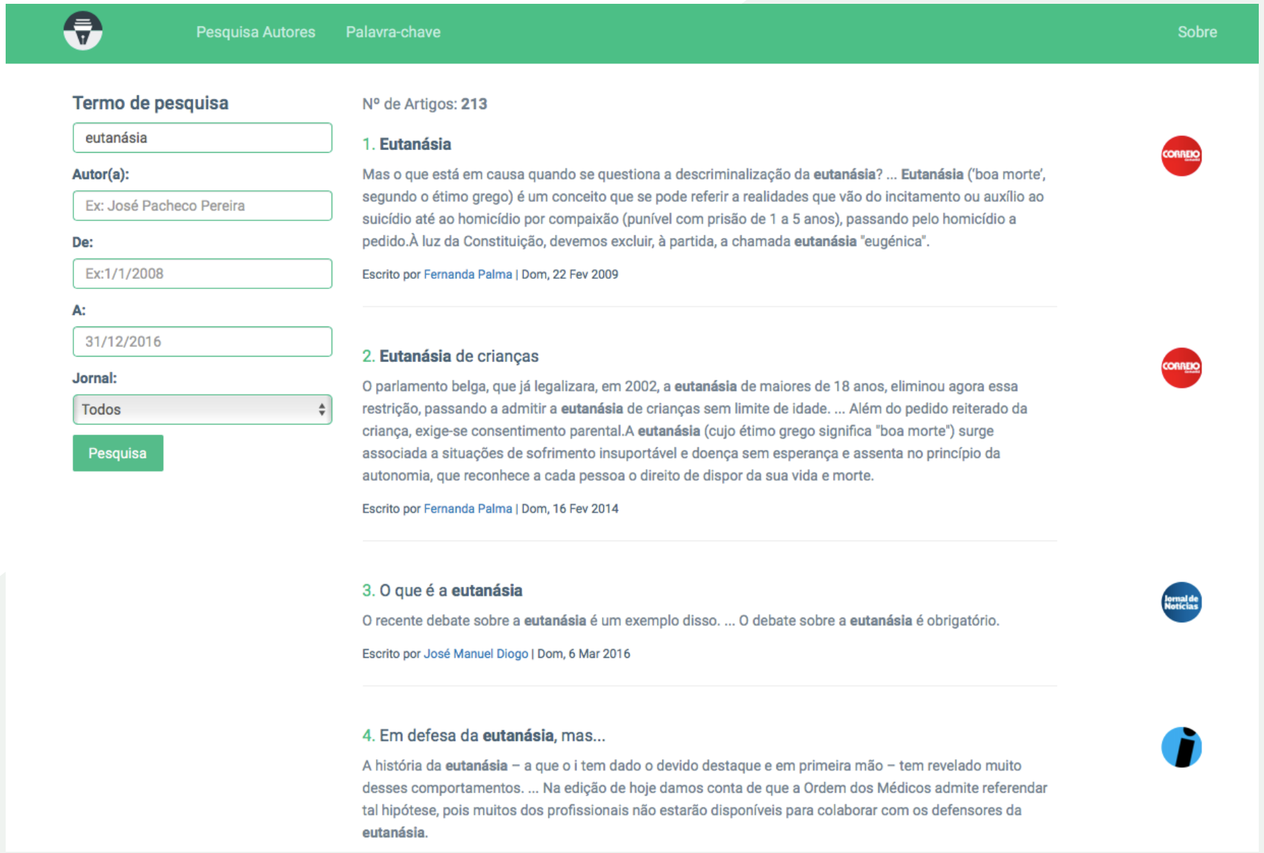


Motor de busca

Disponível um motor de busca que permite procuras por palavra (raiz) ou expressão exacta ("")

Filtros

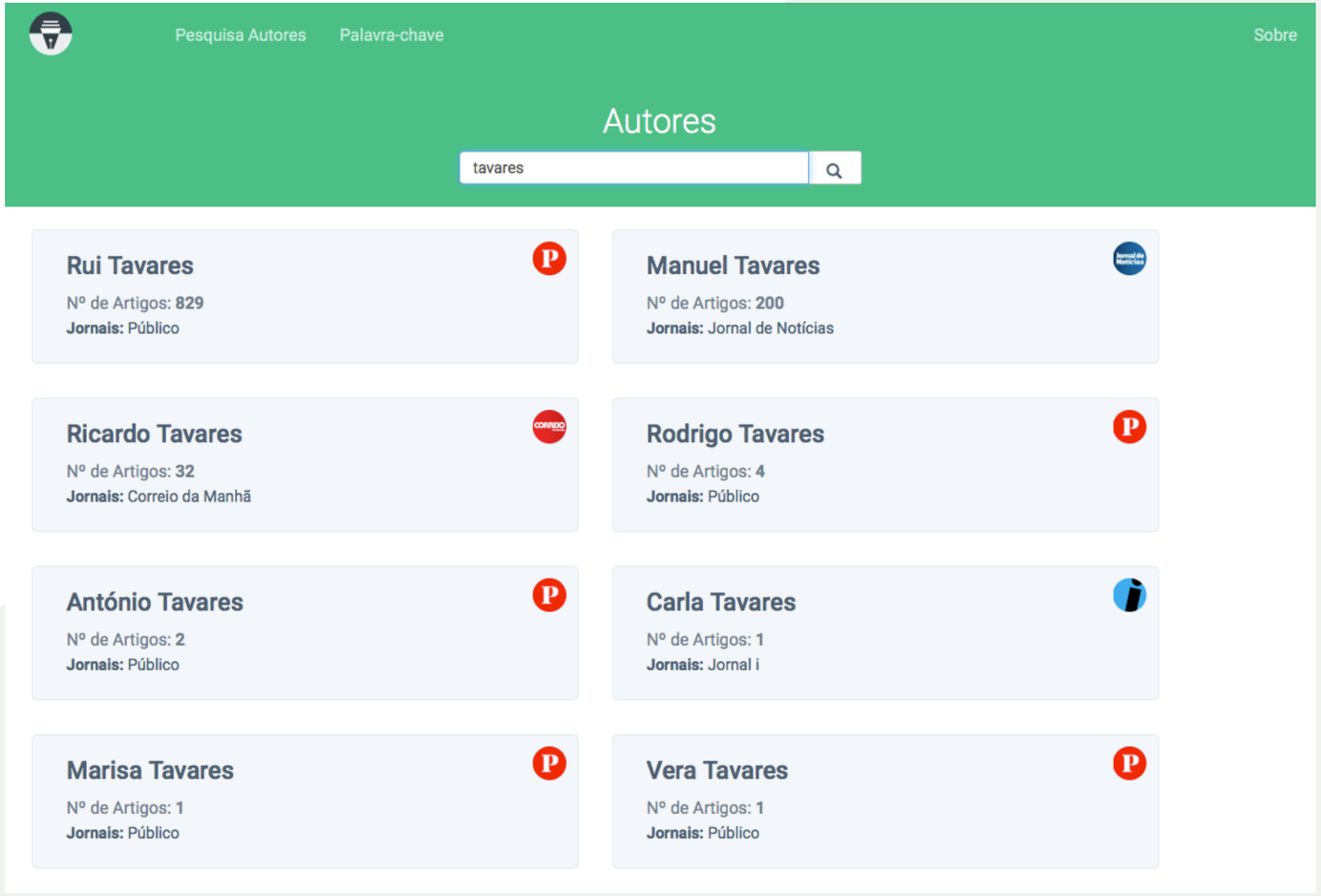
- Autor
- Período temporal
- Jornal de publicação



Autor

Pesquisa de autores

Exemplo: Tavares



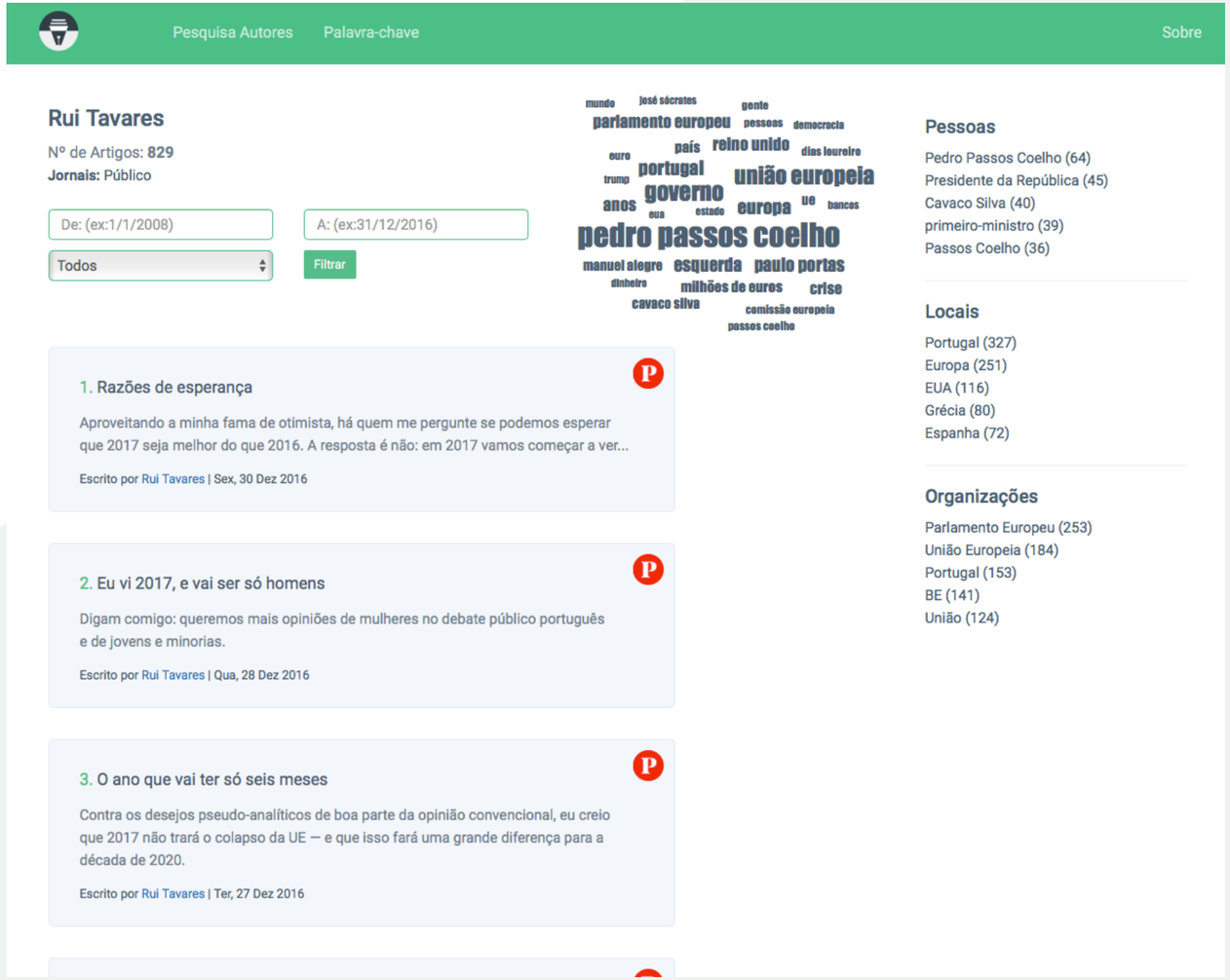
Autor

Página do autor

Artigos da base de dados
- Filtro por data
- Jornal

Nuvem de palavras-chave

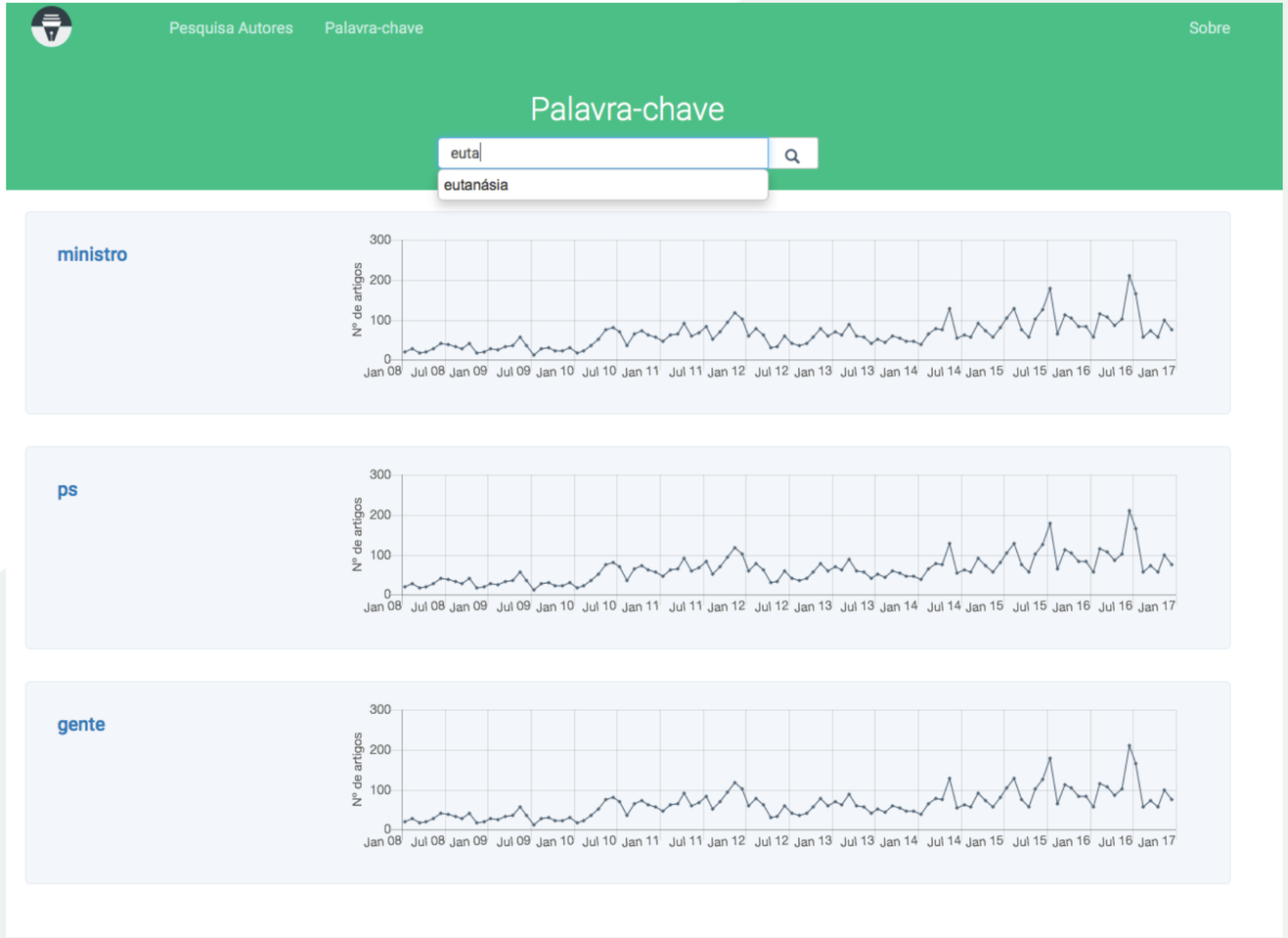
Entidades mencionadas:
- Pessoas
- Locais
- Organizações



Palavra-chave

Pesquisa de palavras-chave com sugestão

Exemplo: eutanásia

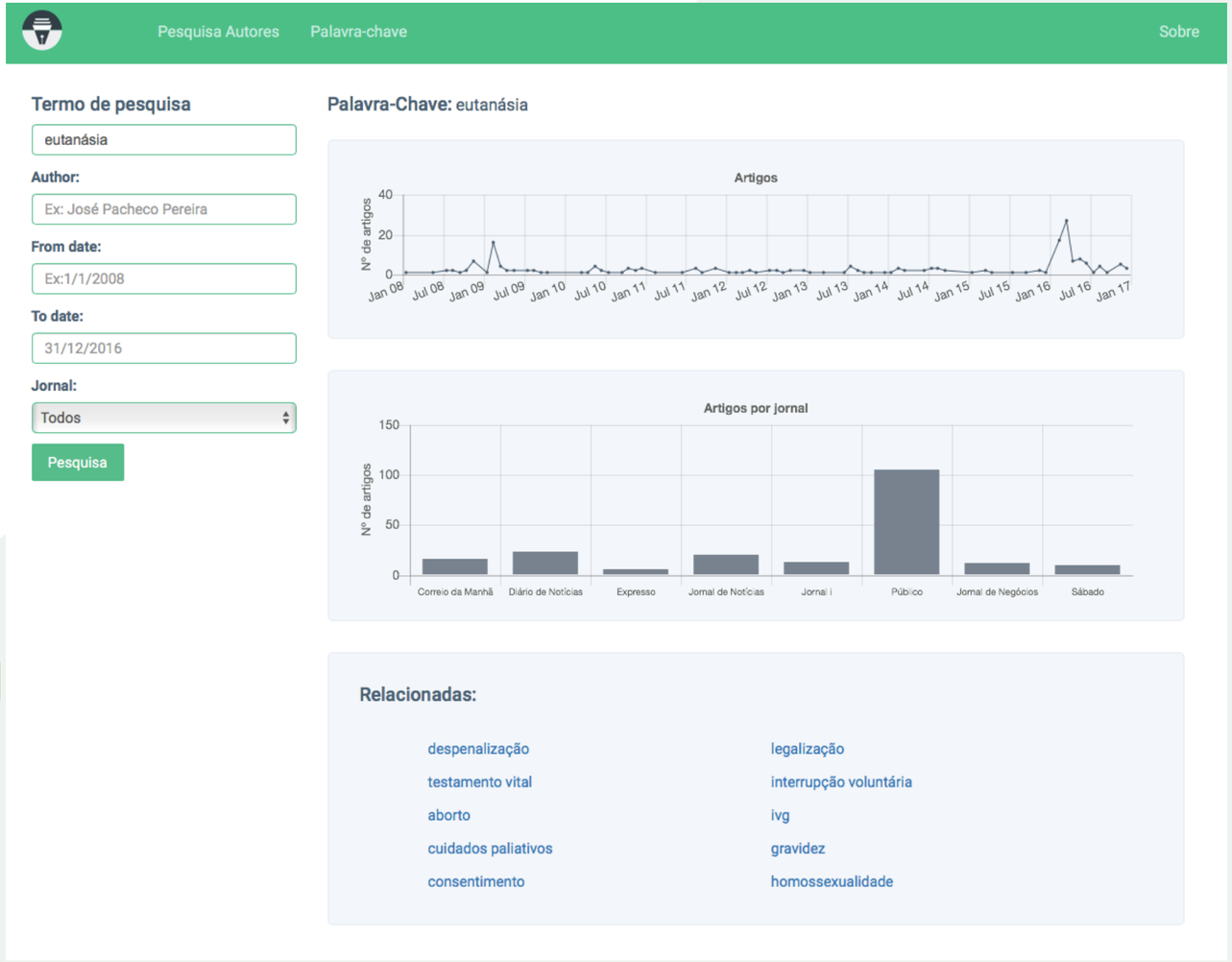


Palavra-chave

Resultados por:

- Período temporal
- Jornal de publicação
- Palavras-chave relacionadas

Também existe possibilidade de filtragem por autor, período temporal e jornal de publicação



Números I

Artigos: 85 530

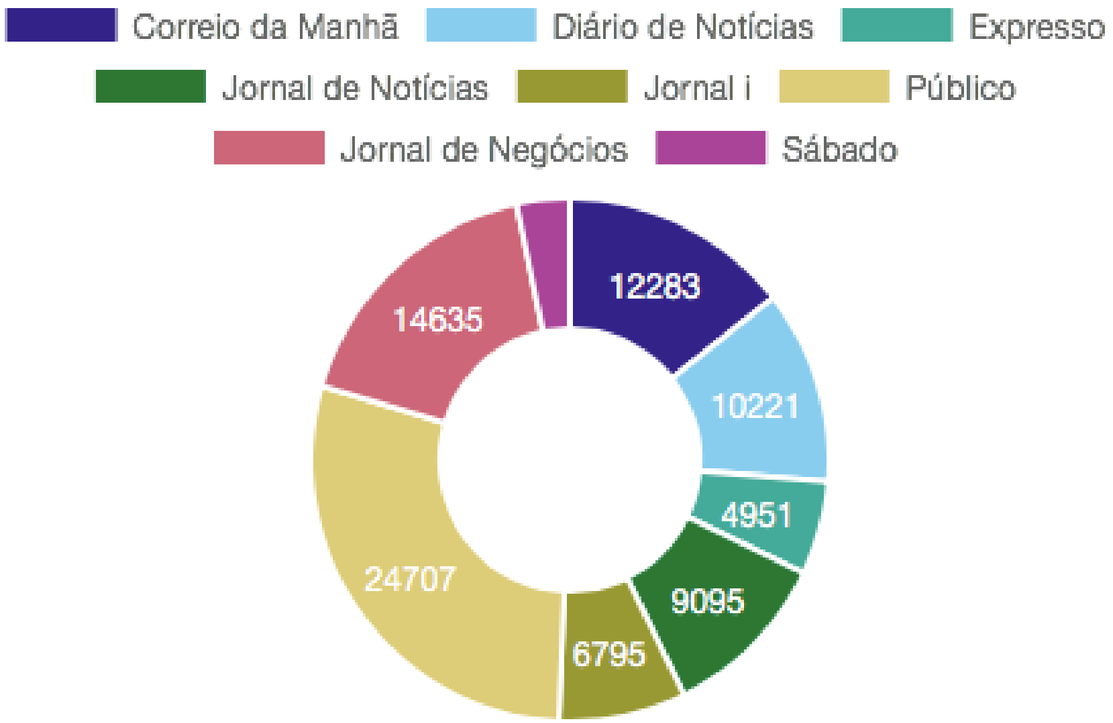
Autores distintos: 3571

Jornais:

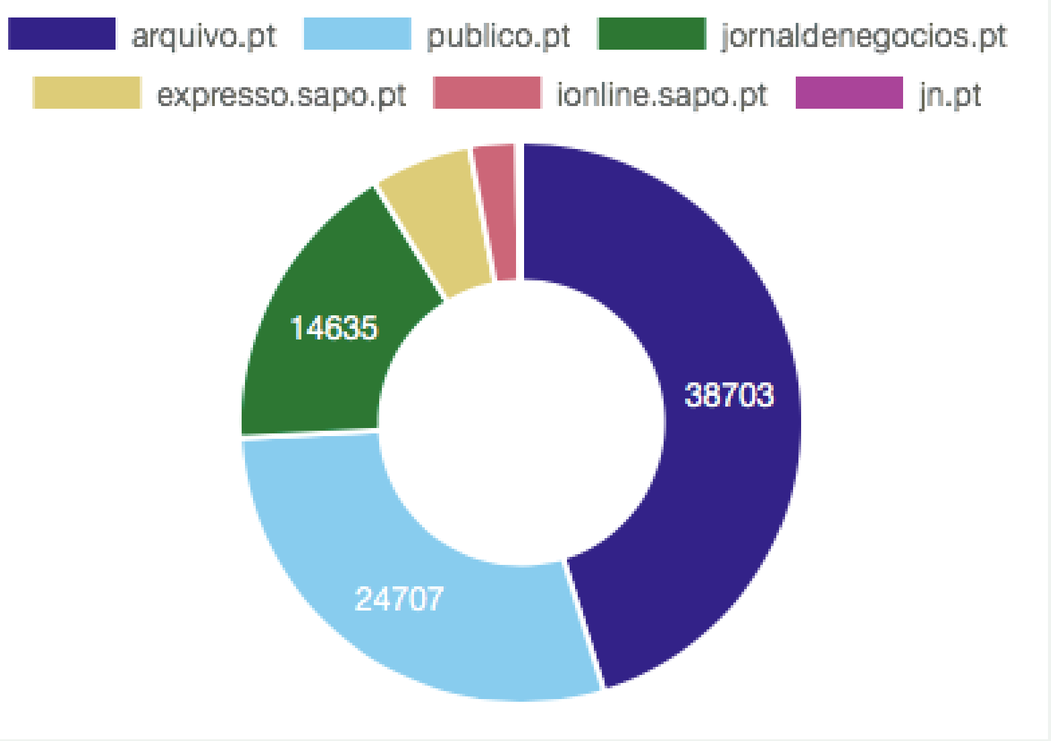
- Correio da Manhã
- Diário de Notícias
- Expresso
- Jornal de Notícias
- Jornal i
- Público
- Jornal de Negócios
- Sábado

Números II

Artigos por jornal



Artigos por fonte



Versão 2.0

Incluir mais jornais

Alargar o período temporal

Adicionar informação das redes socias:

- indexar as páginas públicas dos autores
- incluir métricas de partilhas e interações dos artigos de opinião
- monitorização em tempo real

www.arquivodeopiniao.pt

Obrigado!

Processamento de Língua Natural (NLP)

Campo da Ciência da Computação que tem como objectivo estudar a interacção entre computadores e línguas humanas (naturais)

Divisão em tarefas computacionais:

- etiquetagem morfo-sintática
- reconhecimento de entidades
- classificação de textos
- tradução automática

...

Base de trabalho incide tradicionalmente sobre corpora (oral ou escrito)

Pipeline

1. Identificação dos artigos de opinião (web scraping)

2. Identificação dos elementos dos artigos: título, autor, corpo, data de publicação, etc

3. Limpeza dos texto (em particular os nomes)

4. Processamento: etiquetagem morfo-sintática, extração entidades

- Python: nltk, re, pymongo, ...
- Scrapy (web scraping) Scrapy
- LX-Tagger (etiquetagem morfo-sintática)
- Standford NER (reconhecimento das entidades)
- Django (front-end) django