



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

The screenshot shows the website's header with a search bar and navigation links. Below the header, three article snippets are displayed, each with a title, a short description, and a date.

Como não eleger um Secretário Geral das Nações Unidas S
 É o caráter singular e inovador da atual eleição para Secretário Geral e as várias anunciadas e percebidas manobras de bastidores que catapultaram a escolha deste ano para o topo da agenda pública
 Escrito por [Mónica Ferro](#), 3 de Outubro de 2016 às 00:00

Que viva a Grécia! DN
 A Grécia parece ter peçonha. Nas reuniões internacionais, Papandreou é objecto de todas as recuadas atenções e de todos os silenciosos desfavores. Os países "periféricos", nos quais se inclui Portugal, nada querem a ter com a Grécia, uma desgraça que dá azar. A simples menção do nome do país faz estremecer de horror os dirigentes da Europa "pobre." A Grécia é-lhes desprezível. Temem o "contágio", e afirmam, com fogaçidade, nada ter a ver com "aquilo". Se a Europa económica e política está a desfazer-se, a Europa moral (o que quer que a expressão signifique) só não cai em estilhaços - porque não existe.
 Escrito por [Baptista Bastos](#), 29 de Junho de 2011 às 00:00

Ortografia é que não P
 Chamem-lhe poligrafia, multigrafia, plurigrafia, arbitriografia, o que quiserem. Ortografia é que não.
 Escrito por [Nuno Pacheco](#), 11 de Março de 2016 às 00:00

Fontes

- Arquivo.pt (API)
- Secção de Opinião (online)

The screenshot shows the GitHub repository page for 'arquivo / pwa-technologies'. The repository has 9 watchers, 7 stars, and 4 forks. The 'Wiki' tab is selected, showing the 'APIs' page. The page was last edited by Daniel Gomes on Mar 1 with 16 revisions. The 'APIs' section lists three active APIs: 'Arquivo.pt API (Full-text & URL search)', 'CDX-server API (URL search)', and 'Memento Timetravel API (URL search)'. The 'Deprecated' section lists two deprecated APIs: 'URL search: OpenSearch-based (deprecated)' and 'Full-text search: OpenSearch-based (deprecated)'. The 'Under development' section lists one API: 'Image Search API (under development)'. A sidebar on the right shows a list of 30 pages, including 'Home', 'APIs', 'Arquivo.pt API v.0.2 (beta version)', 'Compile', 'ConfigureSearch', 'Install', 'L2R4WAIR', 'MainFeatures', 'Memento Time travel API', 'Memento API URL Search in Arquivo.pt', 'Memento API - URL Search in World Web Archives', 'milestones', 'OpenSearch API - Full text Search (deprecated)', 'OpenSearch API - URL Search (deprecated)', and 'OpenSearch based Arquivo.pt API'. A 'Show 15 more pages...' link is at the bottom of the sidebar.

Motor de busca

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

Filtros

- Autor
- Período temporal
- Jornal de publicação

The screenshot shows a search engine interface with a green header. The search term 'eutanásia' is entered in the search box. The results are displayed in a list format, with each result including a title, a brief description, and the author and date of publication. The search results are as follows:

Termo de pesquisa	Nº de Artigos
eutanásia	213
Autor(a): Ex: José Pacheco Pereira	
De: Ex: 1/1/2008	
A: 31/12/2016	
Jornal: Todos	
Pesquisa	

1. Eutanásia	2. Eutanásia de crianças	3. O que é a eutanásia	4. Em defesa da eutanásia, mas...
Mas o que está em causa quando se questiona a descriminalização da eutanásia? ... Eutanásia ('boa morte', segundo o étimo grego) é um conceito que se pode referir a realidades que vão do incitamento ou auxílio ao suicídio até ao homicídio por compaixão (punível com prisão de 1 a 5 anos), passando pelo homicídio a pedido. À luz da Constituição, devemos excluir, à partida, a chamada eutanásia "eugénica".	O parlamento belga, que já legalizara, em 2002, a eutanásia de maiores de 18 anos, eliminou agora essa restrição, passando a admitir a eutanásia de crianças sem limite de idade. ... Além do pedido reiterado da criança, exige-se consentimento parental. A eutanásia (cujo étimo grego significa "boa morte") surge associada a situações de sofrimento insuportável e doença sem esperança e assenta no princípio da autonomia, que reconhece a cada pessoa o direito de dispor da sua vida e morte.	O recente debate sobre a eutanásia é um exemplo disso. ... O debate sobre a eutanásia é obrigatório.	A história da eutanásia – a que o i tem dado o devido destaque e em primeira mão – tem revelado muito desses comportamentos. ... Na edição de hoje damos conta de que a Ordem dos Médicos admite referendar tal hipótese, pois muitos dos profissionais não estarão disponíveis para colaborar com os defensores da eutanásia.
Escrito por Fernanda Palma Dom, 22 Feb 2009	Escrito por Fernanda Palma Dom, 16 Feb 2014	Escrito por José Manuel Diogo Dom, 6 Mar 2016	

Autor

Pesquisa de autores

Exemplo: Tavares

The screenshot shows the 'Autores' search results page on Arquivo.pt. The search bar contains the text 'tavares'. The results are displayed in a grid of eight author cards, each with a name, number of articles, and the publication(s) they wrote for. Some cards also feature a red 'P' icon, indicating a prize.

Nome	Nº de Artigos	Jornais	Prémio
Rui Tavares	829	Público	Sim
Manuel Tavares	200	Jornal de Notícias	Não
Ricardo Tavares	32	Correio da Manhã	Não
Rodrigo Tavares	4	Público	Sim
António Tavares	2	Público	Sim
Carla Tavares	1	Jornal i	Não
Marisa Tavares	1	Público	Sim
Vera Tavares	1	Público	Sim

Autor

Página do autor

Artigos da base de dados

- Filtro por data
- Jornal

Nuvem de palavras-chave

Entidades mencionadas:

- Pessoas
- Locais
- Organizações

The screenshot displays the author profile for Rui Tavares. At the top, there is a navigation bar with 'Pesquisa Autores' and 'Palavra-chave'. The author's name 'Rui Tavares' is prominently displayed, along with statistics: 'Nº de Artigos: 829' and 'Jornais: Público'. Below this, there are search filters for date (De: (ex:1/1/2008) and A: (ex:31/12/2016)) and a 'Filtrar' button. To the right, a word cloud features terms like 'parlamento europeu', 'união europeia', 'governo', and 'pedro passos coelho'. On the far right, there are three sections: 'Pessoas' (listing Pedro Passos Coelho, Cavaco Silva, etc.), 'Locais' (listing Portugal, Europa, etc.), and 'Organizações' (listing Parlamento Europeu, etc.). The main content area shows three article snippets, each with a title, a short text preview, and the author's name and date.

Rui Tavares
Nº de Artigos: 829
Jornais: Público

De: (ex:1/1/2008) A: (ex:31/12/2016)
Todos

1. Razões de esperança P
Aproveitando a minha fama de otimista, há quem me pergunte se podemos esperar que 2017 seja melhor do que 2016. A resposta é não: em 2017 vamos começar a ver...
Escrito por Rui Tavares | Sex, 30 Dez 2016

2. Eu vi 2017, e vai ser só homens P
Digam comigo: queremos mais opiniões de mulheres no debate público português e de jovens e minorias.
Escrito por Rui Tavares | Qua, 28 Dez 2016

3. O ano que vai ter só seis meses P
Contra os desejos pseudo-analíticos de boa parte da opinião convencional, eu creio que 2017 não trará o colapso da UE — e que isso fará uma grande diferença para a década de 2020.
Escrito por Rui Tavares | Ter, 27 Dez 2016

Pessoas
Pedro Passos Coelho (64)
Presidente da República (45)
Cavaco Silva (40)
primeiro-ministro (39)
Passos Coelho (36)

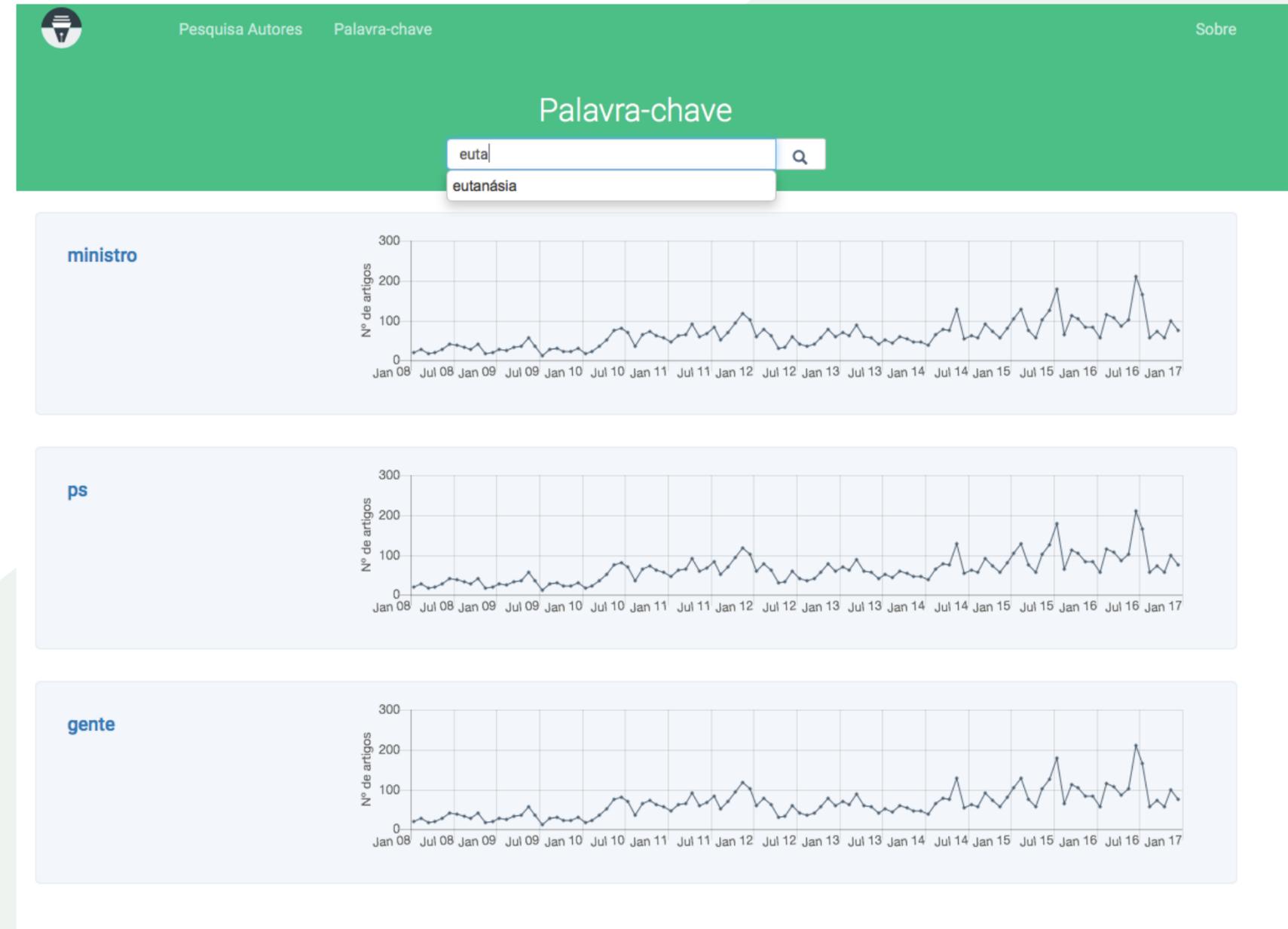
Locais
Portugal (327)
Europa (251)
EUA (116)
Grécia (80)
Espanha (72)

Organizações
Parlamento Europeu (253)
União Europeia (184)
Portugal (153)
BE (141)
União (124)

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

Pesquisa Autores Palavra-chave
Sobre

Termo de pesquisa

Author:

From date:

To date:

Jornal:

[Pesquisa](#)

Palavra-Chave: eutanásia

Artigos

Artigos por jornal

Jornal	Nº de artigos
Correio da Manhã	~20
Diário de Notícias	~25
Expresso	~10
Jornal de Notícias	~25
Jornal I	~15
Público	~110
Jornal de Negócios	~15
Sábado	~15

Relacionadas:

despenalização

testamento vital

aborto

cuidados paliativos

consentimento

legalização

interrupção voluntária

ivg

gravidez

homossexualidade

Números I

Artigos: 85 530

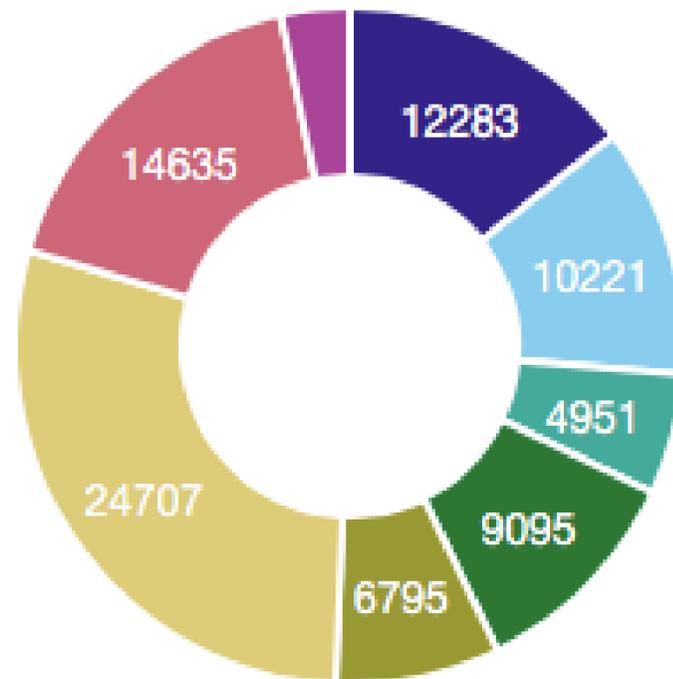
Autores distintos: 3571

Jornais:

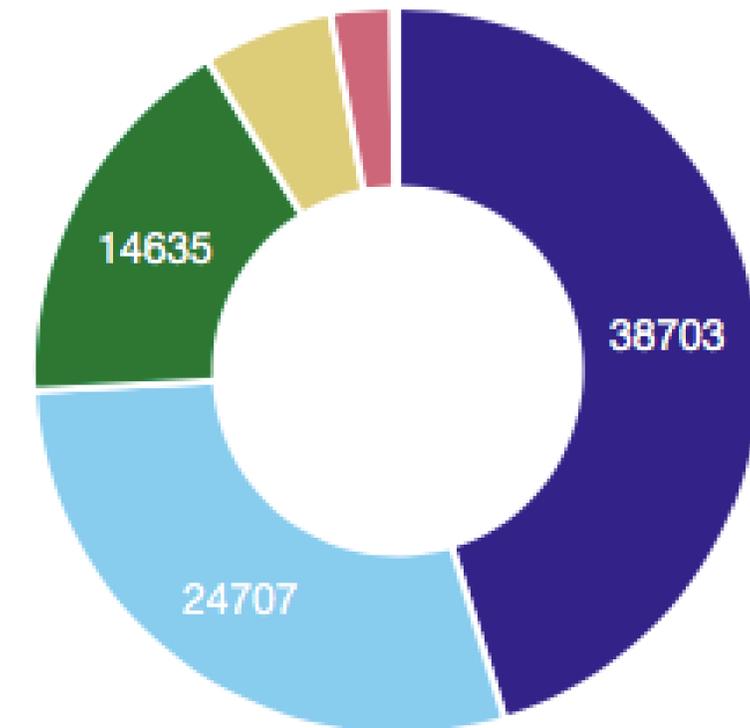
- Correio da Manhã
- Diário de Notícias
- Expresso
- Jornal de Notícias
- Jornal i
- Público
- Jornal de Negócios
- Sábado

Números II

Artigos por jornal



Artigos por fonte



Versão 2.0

Incluir mais jornais

Alargar o período temporal

Adicionar informação das redes sociais:

- 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)
 - Stanford NER (reconhecimento das entidades)
 - Django (front-end) 