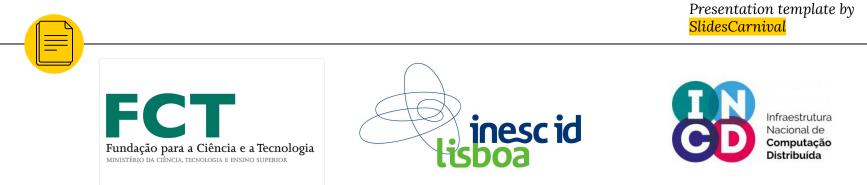
The 'Arquivo de Opinião' archive

Miguel Won Café com o Arquivo.pt





About me

Miguel Won

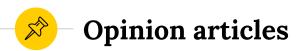
Postdoc researcher at INESC-ID

miguelwon@tecnico.ulisboa.pt





- Political commentary is present in everyday news media:
 - "Experts" in TV broadcasting channels
 - Columnist in newspapers
- This type of opinion plays an important role in the process of the *narrative construction* of the *public realm*:
 - Selection of events
 - Authority position
 - **Deciphers** the political complexities



- In this work we consider only opinion articles from newspapers
- Definition: article usually about the current the state of public affairs, authored by one or multiple authors, that expresses the author's personal opinion
- Two-sided role in respect to public opinion
 - They can be interpreted as a mirror of the public opinion
 - But can also be accused of its main influencer
- Essential component of the public debate



- Memory of political debates allows the recalling of ideas, main debatable issues, the argumentative logics, as well as the political positions of the various political actors (many political commentators are, were or will be themselves active politicians)
- Memory of political discussion is essential to the proper functioning of democracies
- Archives of this type of memory contributes to a healthy public debate
- Such archives should be digital:
 - Search engine
 - Searches by author, time period or media source
 - Public availability and user friendly

Introduction



Arquivo de Opinião

• Digital archive of opinion articles

ARQUIVO ∗OPINIÃO	
Pesquisa por termo Q	
Nova gestão do Novo Banco com tarefa árdua! Recuencia a Junho deste ano, mês em que ocorreu o último aumento do capital do BES. Nessa altura, se considerarmos a última cotação com direitos e o número de acções existentes à data, tinhamos um valor de capitalização bolaista para o BES pouco acima do 3.300 milhões de euros.	٥
Escrito por Carlos Bastardo, 16 de Outubro de 2014 às 00.00	
Duas culturas Decorrenn 50 anos deside o día em que uma figura pesada sublu os degraus da vetusta Casa do Senado, na Universidade de Cambridge, e se dirigiu ao pódio para proferir Escrito por humo Cisto, 17 de Maio de 2009 às 00:00	0
Éxito do turismo nas mãos dos portugueses "Descansa; sem sair de casa". Este bem pode ser o mote que vai caracterizar as férias das famili Exempo preteños Jorge Costa, 9 de Julho de 2012 as 00.00	0



Arquivo construction

Miguel Won Café com o Arquivo.pt



 Arquivo.pt: web archive of .pt domain

• Opinion section (online)



APIs

Daniel Gomes edited this page Mar 1, 2018 · 16 revisions

 Arquivo.pt API (Full-text & URL search) 	▼ Pages (1)
CDX-server API (URL search)	
 Memento Timetravel API (URL search) 	Find a Page
Deprecated	Home
URL search: OpenSearch-based (deprecated)	Arquivo.pt API v.0.2 (beta
 Full-text search: OpenSearch-based (deprecated) 	version)
Under development	Compile ConfigureSearch
 Image Search API (under development) 	Install
	L2R4WAIR
	MainFeatures
	Memento Time travel API



URL identification	Web scraping	Data Cleaning	NLP	
--------------------	--------------	---------------	-----	--

- Search for clues such
 Title as "opiniao"
- web crawling

- Name correction
- Remove html code
- Manual inspection

- Part-of-speech tagging
- NER
- Key-phrases extraction

•

•

۲

Author

Body

Publication date

Miguel Won Café com o Arquivo.pt



Tools:

Web framework:

• Python: nltk, re, scikit-learn, etc.

Django



- Scrapy (web scraping) 🕜 Scrapy
- MongoDB
- B mongoDB

• Lx-Tagger (pos tagging)



• Stanford NER







SP -

Named Entity Recognition (NER)

- Task: given a text as input identify the entities within the text
 - Person names
 - Locations
 - Organizations

NER Examples

Input: Vancouver is a coastal seaport city on the mainland of British Columbia. The city's mayor is Gregor Robertson.

Location

Output: <u>Vancouver</u> is a coastal seaport city on the mainland of <u>British Columbia</u>. The city's mayor is <u>Gregor Robertson</u>.

Location

Person



x

Named Entity Recognition (cont.)

- Classification task (sequential)
- Many free tools available in the market
 - Stanford NER (CRFs)
 - spaCy (NN)
 - Polyglot (NN)
- We have trained Stanford NER with an annotated corpus for Portuguese (European): CINTIL

Ŕ

- Stanford NER with CINTIL

5-fold Cross-Validation Precision, Recall and F-Measure Results for NER using CINTIL

	True Positive	False Positive	False Negative	Precision	Recall	F-Measure
1	952	221	236	0.81	0.81	0.81
2	846	208	230	0.81	0.79	0.79
3	921	243	261	0.79	0.78	0.79
4	939	209	327	0.82	0.74	0.78
5	892	213	252	0.81	0.78	0.79
Total	4550	1094	1306	0.81	0.78	0.79

NLP

🔊 – Key-phrase extraction

Won, Miguel, Bruno Martins, and Filipa Raimundo. Automatic extraction of relevant keyphrases for the study of issue competition. No. 875. EasyChair, 2019.

- Key-phrase: a word or phrase represents a concept, idea, entity, etc.
 - Refugee Crisis
 - National Health Service
 - António Costa
- Politicians often guide their speeches using key-phrases
- Key-phrase identification can hint us about the topics addressed in a set of speeches



First step: Candidate Selection

- Part-of-Speech tagging followed by a chunk rule:
 - Crise dos Refugiados: NOUN+PREP+NOUN
 - Sistema Nacional de Saúde: NOUN + ADJ + PREP + NOUN
 - António Costa: NOUN + NOUN

Chunking rule (Portuguese): (<NOUN>+ <ADJ>* <PREP>*)? <NOUN>+

🔊 – Second step: rank

- Several methods: TextRank, Phraseness & Informativeness, EmbedRank, etc.
- We can achieve state-of-the-art results with simple heuristic rules:
 - Tf-idf
 - Likelihood metric based in the position
 - Length



Opinion in the Portuguese media



• Frontpage with a search engine

ARQUIVO ©OPINIÃO	
Pesquisa por termo Q	
Nova gestão do Novo Banco com tarefa árdual Recuencia a Juhho deste ano, mês em que ocorreu o último aumento do capital do BES. Nessa altura, se considerarmos a última cotação com direitos e o número de acções existentes datas, inhanos um vaire de capitalização bolisista para o BES pouco acima dos 3.900 milhões de euros. Esente por Carlos Bastardo, 16 de Osubro de 2014 às 00.00	
Duas culturas Decorreram 50 anos desde o dia em que uma figura pesada sublu os degraus da vetusta Casa do Senado, na Universidade de Cambridge, e se dirigiu ao pódrio para proferir Esonte por Nuno Ostab, 17 de Naio de 2009 às 00:00	
Exite do turísmo nas mãos dos portugueses "Descunsar, sem sair de casa". Este bem pode ser o mote que vai caracterizar as férias das famili Escrite por Antóleo Jorge Costo, 9 de Julho de 2012 às 0.00	



Search engine (mongo)

- Search for text of phrase
- Filters:
 - author
 - time interval
 - source

Pesquisa Auto	res Palavra-chave	Sobre
Termo de pesquisa	Nº de Artigos: 213	
eutanásia	1. Eutanásia	000000
Autor(a):	Mas o que está em causa quando se questiona a descriminalização da eutanásia?	
Ex: José Pacheco Pereira	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	
De:	(punível com prisão de 1 a 5 anos), passando pelo homicídio a pedido.À luz da	
Ex:1/1/2008	Constituição, devemos excluir, à partida, a chamada eutanásia "eugénica".	
A:	Escrito por Fernanda Palma Dom, 22 Fev 2009	
31/12/2016		
Jornal:	2. Eutanásia de crianças	00000
Todos 💠	O parlamento belga, que já legalizara, em 2002, a eutanásia de maiores de 18 anos,	
Pesquisa	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 insuportávele doença sem esperança e assenta no princípio da autonomia, que reconhece a cada pessoa o direito de dispor da sua vida e morte. Escrito por Fernanda Palma Dom, 16 Fev 2014	
	3. O que é a eutanásia O recente debate sobre a eutanásia é um exemplo disso O debate sobre a eutanásia é obrigatório.	Remainer Remainer
	Escrito por José Manuel Diogo Dom, 6 Mar 2016	



• Search for author

• -3500 available authors

	Pesquise por nome	Q	
Direção Nº de Artigos: 2980 Jornals: Público Diário de Notícias Express Jornal i	So Sábado	Miguel Esteves Cardoso № de Artigos: 2675 Jornais: Público	C
Fernando Sobral Nº de Artigos: 2358 Jornais: Jornal de Negócios Correio da Mar	unhă	Camilo Lourenço Nº de Artigos: 1724 Jornais: Jornal de Negócios	۵
Ferreira Fernandes Nº de Artigos: 1571 Jornais: Diário de Notícias	DN	Helena Garrido Nº de Artigos: 1432 Jornais: Jornal de Negócios Sábado Correio d	a Manhã

Miguel Won Café com o Arquivo.pt



Author (cont.)

- Each author has its page
- Key-phrases cloud

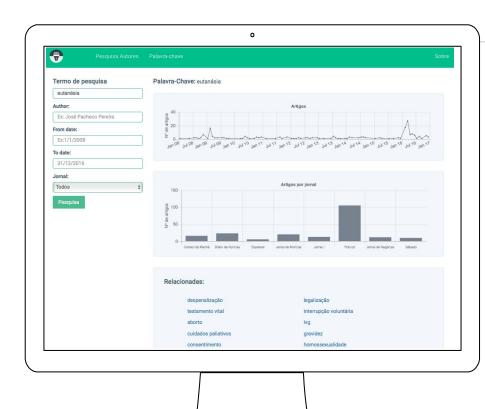
- Mentioned entities:
 - Persons names
 - Locations
 - Organizations

Rui Tavares i ^o de Artigos: 829 omais: Público De: (ex:1/1/2008) A: (ex:31/12/2016) Todos Filtrar	tenercità tenerata enventa unata enventa enventa enventa unata esta enventa	Pessoas Pedro Passos Coelho (64) Presidente da República (45 Cavaco Silva (40) primeiro-ministro (39) Passos Coelho (36)
 Razões de esperança Aproveltando a minha fama de otimista, há quem me pergunte se podemo: que 2017 seja melhor do que 2016. A resposta é não: em 2017 vamos com Econto por fuit Tavares [Sex, 30 bez 2016 		Portugal (327) Europa (251) EUA (116) Grécia (80) Espanha (72)
 Eu vi 2017, e vai ser só homens Digam comigo: queremos mais opiniões de mulheres no debate público po e de jovens e minorias. 	P	Organizações Parlamento Europeu (253) União Europeia (184) Portugal (153) BE (141) União (124)



 Search indexed key-phrases (with autocomplete)

- Outputs
 - No. articles by time and source
 - Related (word embeddings)



Miguel Won Café com o Arquivo.pt



3571 Authors

30 000 key-phrases

9 Years of publications (2008–2016)







- Add additional sources: Observador, Jornal Económico, ECO, Visão
- 2016-Present
- Social Media:
 - Twitter
 - Facebook
- Real time monitoring
- Add political position
- New Frontend
- Contas de utilizador?



A	1. 0. :	
Pesquisar Arquive	o de Opinião	Sign up
DN Eco Expresso IOnline Jornal Económico J	ornal de Negócios JN (Observador Público Sábado Visão
João Monteiro Pinto	Laurinda Alves	\circ
Covid-19, gestão de risco e o	Quantos alur	nos vão ficar para
"novo normal"	trás?	iob vao noar para
Nov 11	Nov 11	
This a asda da sda da	No atual contexto di	igital as dificuldades
	multiplicam-se	
®(22) ƒ (1000) 以(10) Fonte	e @(22) f (1000) £7(10)	Fonte
Recentes Populares Top		Sobre
Anne-Marie	DN	
Tester in the second se		O Arquivo de Opinião é um arquivo digital da opinião publicada nos principais canais
Juntos venceremos o combate contra o Co	ovia 19	da imprensa portuguesa.
January 1, 2014		
Neste momento, a rápida propagação global do coronavírus implica que	e ninguém está a salvo desta	Tendências
pandemia até que todos estejamos em segurança.		
¹	Fonte	António Costa Covid-19
and and the American Starts		Marcelo Rebelo de Sousa
Carlos Pimenta	E	Crise
Cântico por um futuro melhor	_	Europa Alemanhã
		Trump
January 1, 2014		Pandemia
O Observatório de Economia e Gestão de Fraude (OBEGEF) promove n		
reflexão sobre as temáticas da fraude, da corrupção, da economia não-		Autores + partilhados
integridade e da transparência, contribuindo deste modo para a formaç mais esclarecida e mais participativa Carlos Pimenta	ao de uma opinião pública	João Miguel Tavares
mais coolarcolaa o mais paraopativa Ganos Fillionta		Laurinda Alves

Miguel Won Café com o Arquivo.pt

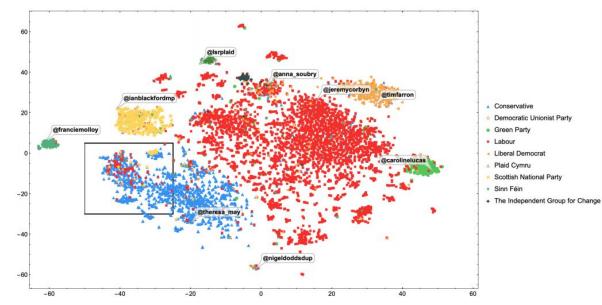


Pesquisar	Arqui	vo de Opini	ião				Sign up
DN Eco Expresso IOnline	Jornal Económico	Jornal de Negócios	JN	Observador	Público	Sábado	Visão
Helena Garrido							
Jornais: DX 🕒 🕐	Posicionam	ento Twitter					
Nº de artigos: 123							
Tendências do autor		۰.			• •		•
António Costa(121) Covid-19 +(100) Marcelo Rebelo de Sousa(90)						• .	
Crise(85)				· •.			
Europa(30) Alemanhã(24)		"PS			"Hele	na Garrido*	
Trump(20)				۰.			
Pandemia(12)				Pedro Sar	ntana Lopes		
							۰.
				PSD			
						•	
				•			
Recentes Populares							
Anne-Marie			-	Autore	es próxin	105	
			DN	António C			
Juntos venceremos o con	nbate contra o	Covid 19		Jose Man Autor 3	uel Fernande	S	
January 1, 2014				Autor 4			



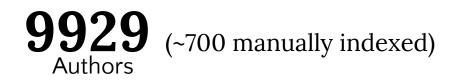
Twitter political position

Miguel Won and Jorge Fernandes; *Political space as an embedding space. From Twitter retweets to party labels*



Miguel Won Café com o Arquivo.pt









- Political commentary is an important section of newspaper media
- A digital archive of this type of memory contributes to a better public debate
- *Arquivo de Opinião* main objective is to offer a digital online archive of the political opinion published in the main Portuguese newspapers
- All data was processed in order to extract additional information (NLP)
- M. Won. "*Political opinions on the past Web*" In Daniel Gomes Last Editor (Ed.), *The Past Web*. To be published.



Acknowledgements

This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the scholarship with reference <u>SFRH/BPD/104176/2014</u>, as well as through the INESC-ID multi-annual funding from the PIDDAC programme, which has the reference <u>UID/CEC/50021/2013</u>, and FEDER under the project 22153-01/SAICT/2016