

The 'Arquivo de Opinião' archive

Miguel Won
Café com o Arquivo.pt

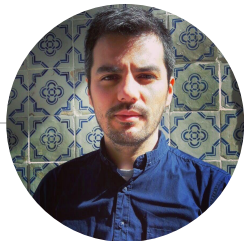


Presentation template by
SlidesCarnival

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

 inesc id
lisboa

 IN
CD
Infraestrutura
Nacional de
Computação
Distribuída



About me

Miguel Won

Postdoc researcher at INESC-ID

miguelwon@tecnico.ulisboa.pt

1

Introduction

Political Punditry



Political commentary

- 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



Opinion articles

- ◉ 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**



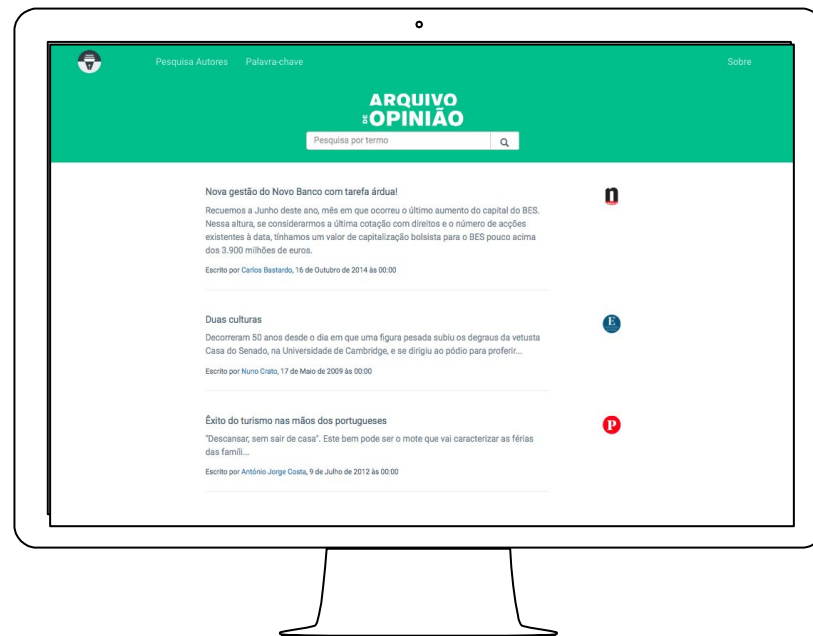
Memory

- 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



Arquivo de Opinião

- Digital archive of opinion articles



2

Collect & process

Arquivo construction



Data sources

- **Arquivo.pt**: web archive of .pt domain
- Opinion section (online)

The screenshot shows the GitHub homepage. At the top, there is a banner that says "Join GitHub today" with a "Sign up" button. Below the banner, the "APIs" section is visible, listing several APIs: "Arquivo.pt API (Full-text & URL search)", "CDX-server API (URL search)", and "Memento Timetravel API (URL search)". There are also sections for "Deprecated" and "Under development" APIs. On the right side, there is a "Pages" sidebar with a search bar and a list of links including "Home", "APIs", "Arquivo.pt API v.0.2 (beta version)", "Compile", "ConfigureSearch", "Install", "L2R4WAIR", "MainFeatures", and "Memento Time travel API".



Pipeline

URL identification

- Search for clues such as “opinioao”
- web crawling

Web scraping

- Title
- Author
- Publication date
- Body
- ...

Data Cleaning

- Name correction
- Remove html code
- Manual inspection




NLP

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



Pipeline

Tools:

- Python: nltk, re, scikit-learn, etc.
- Scrapy (web scraping)  Scrapy
- Lx-Tagger (pos tagging) 
- Stanford NER 

Web framework:

- Django 
- MongoDB 

3

NLP

NLP tasks



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

Location

Person



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



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



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

4

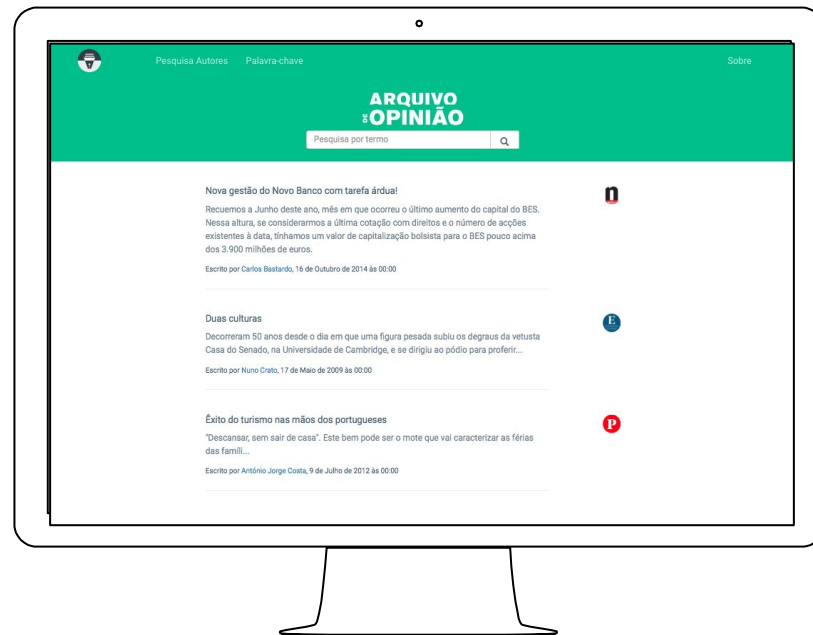
Arquivo de Opinião

Opinion in the Portuguese media



Arquivo de Opinião

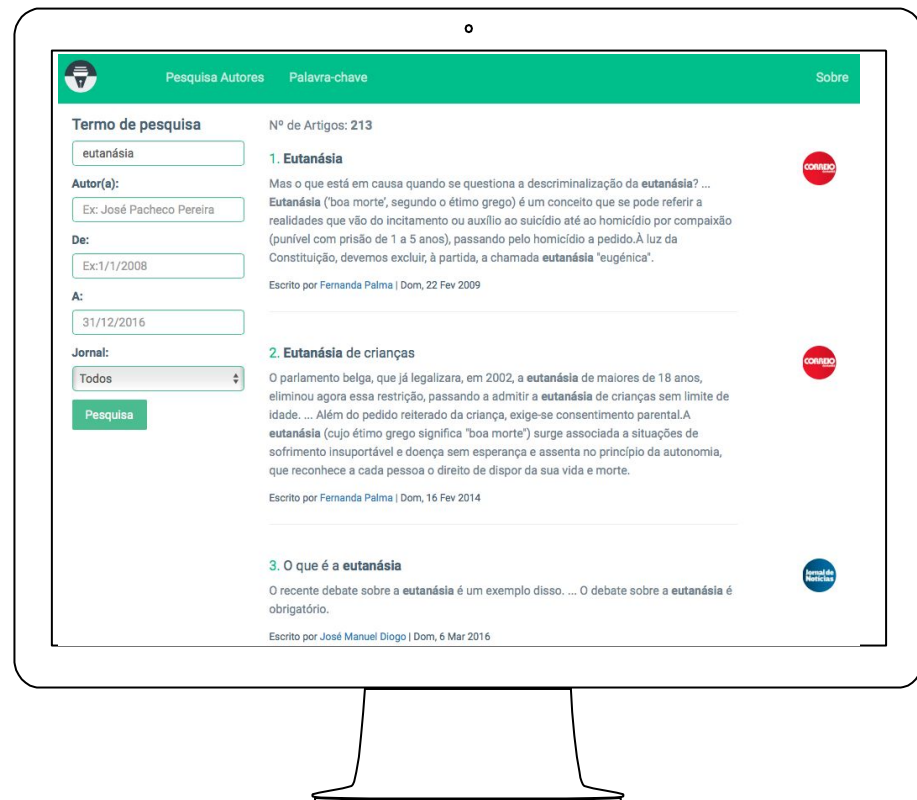
- Frontpage with a search engine





Search engine (mongo)

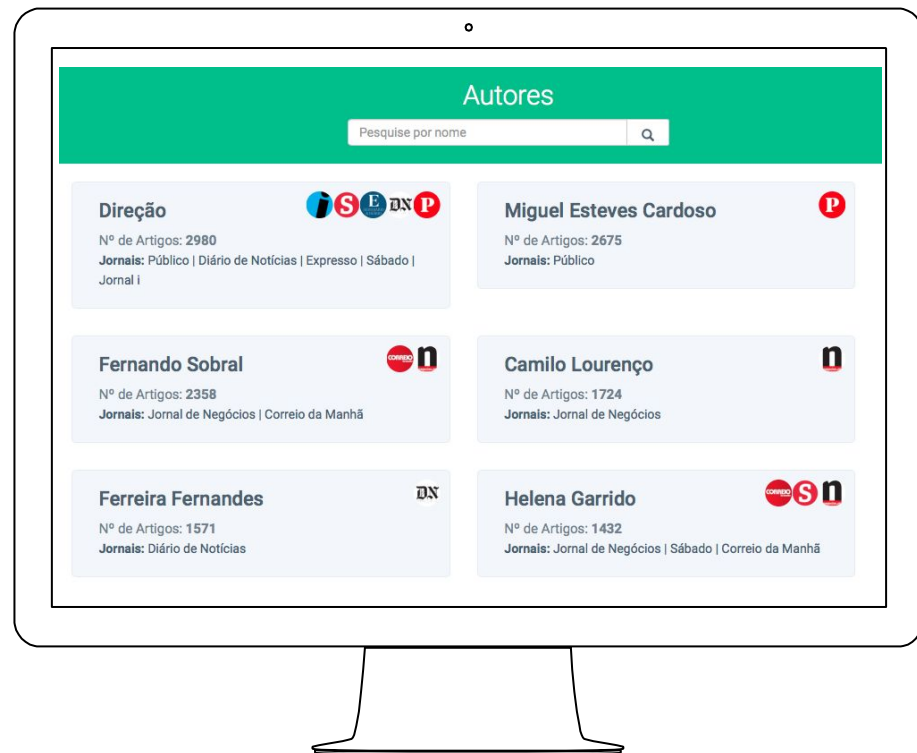
- Search for text or phrase
- Filters:
 - author
 - time interval
 - source





Author

- Search for author
- 3500 available authors





Author (cont.)

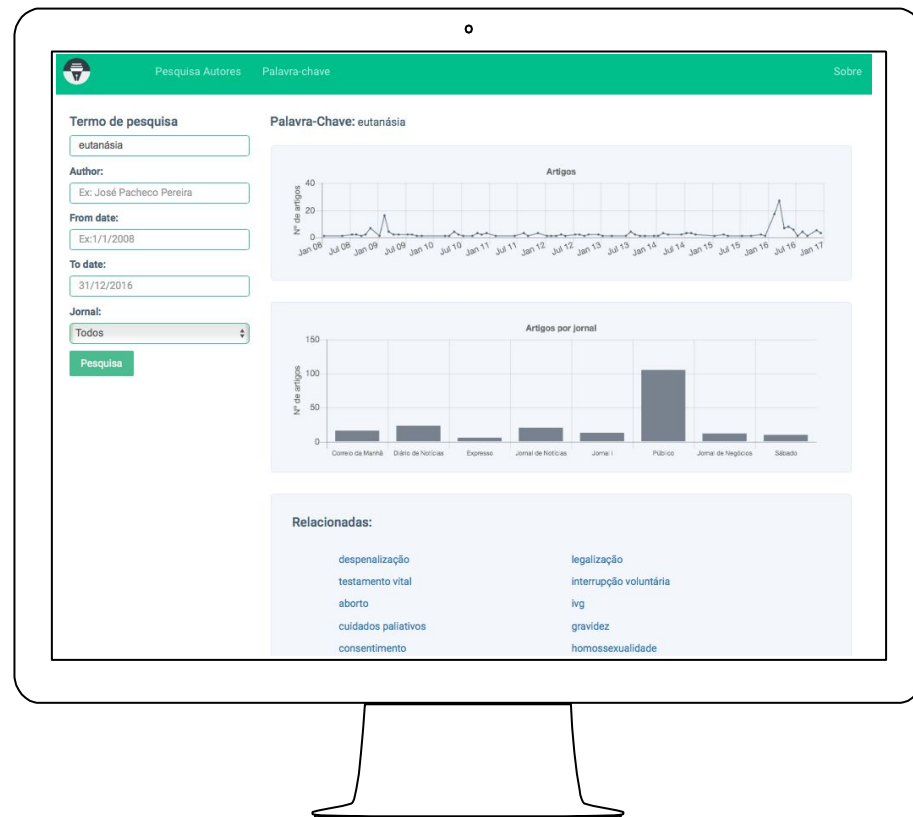
- Each author has its page
- Key-phrases cloud
- Mentioned entities:
 - Persons names
 - Locations
 - Organizations

The screenshot displays the author profile for Rui Tavares. At the top, there are search filters for 'Pesquisa Autores' and 'Palavra-chave'. The author's name 'Rui Tavares' is prominently displayed, along with 'Nº de Artigos: 829' and 'Jornais: Público'. Below this, there are date range filters (De: (ex:1/1/2008) and A: (ex:31/12/2016)) and a 'Filtrar' button. A key-phrase cloud is visible, with 'pedro passos coelho' being the largest word. Other words include 'portugal', 'parlamento europeu', 'criso', 'governo', 'esquerda', 'união europeia', 'europa', 'reino unido', 'paises', 'paulo portas', 'milhões de euros', 'manuel alegre', 'estado', 'passos coelho', 'jósé sócrates', 'democracia', 'cavaco silva', 'banco', 'comissão europeia', 'mundo', 'trampo', 'pessoas', 'euro', 'país', 'almbedo', 'dias laureiro', 'BIOS', and 'passos coelho'. On the right side, there are sections for 'Pessoas' (listing Pedro Passos Coelho, Cavaco Silva, and Passos Coelho with their respective article counts), 'Locais' (listing Portugal, Europa, EUA, Grécia, and Espanha with counts), and 'Organizações' (listing Parlamento Europeu, União Europeia, Portugal, BE, and União with counts). Two article excerpts are shown below, each with a red 'P' icon indicating mentioned entities. The first excerpt is titled '1. Razões de esperança' and discusses optimism about 2017. The second is titled '2. Eu vi 2017, e vai ser só homens' and discusses opinions on women's debate.



Key-phrases

- Search indexed key-phrases (with autocomplete)
- Outputs
 - No. articles by time and source
 - Related (word embeddings)



85 530

Articles

3571

Authors

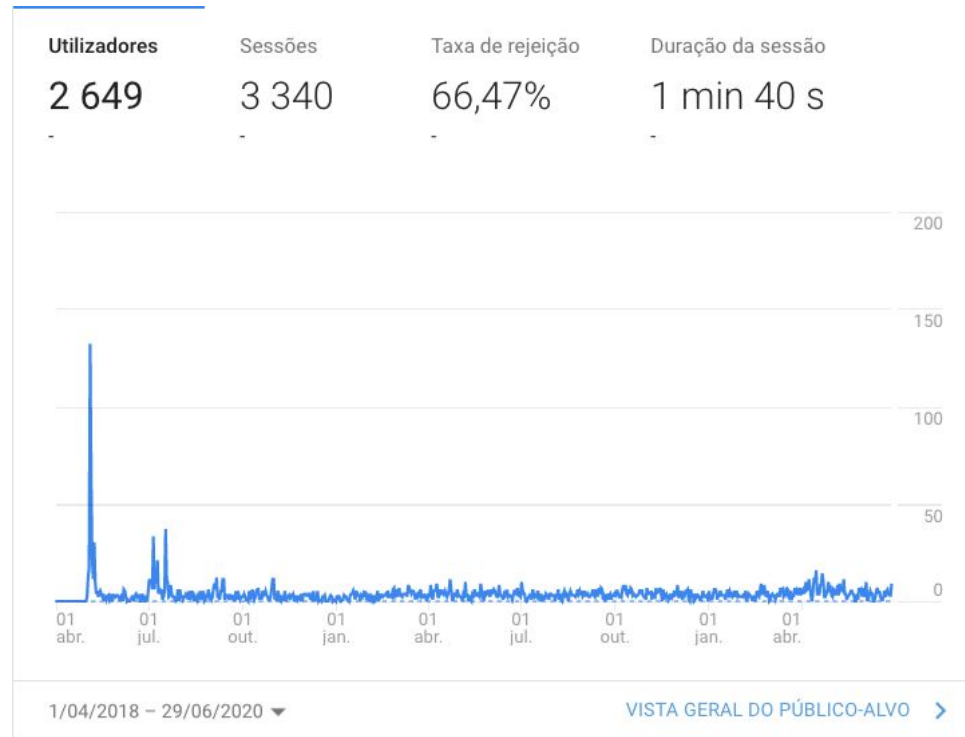
30 000

key-phrases

9

Years of publications (2008-2016)





5

Arquivo de Opinião II

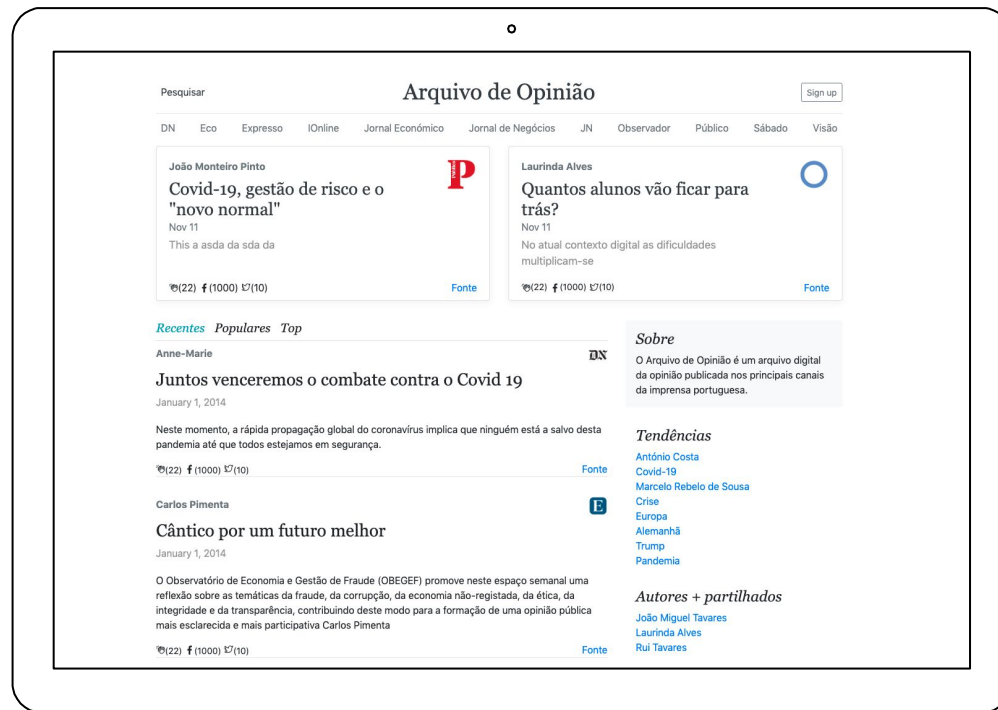


Version 2.0

- ◉ 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?



Homepage





Author

The screenshot shows the profile page for Helena Garrido on the Arquivo de Opinião website. The page includes a search bar, navigation links for various news outlets, and a list of journals where she has published. A 'Tendências do autor' section lists topics like António Costa and Covid-19. A 'Posicionamento Twitter' scatter plot shows her position relative to other authors like PS and PSD. Below, there are sections for 'Recentes Populares' and 'Autores próximos'.

Arquivo de Opinião Sign up

DN Eco Expresso IOnline Jornal Económico Jornal de Negócios JN Observador Público Sábado Visão

Helena Garrido

Jornais:

Nº de artigos: 123

Tendências do autor

- [António Costa](#)(121)
- [Covid-19](#) +(100)
- [Marcelo Rebelo de Sousa](#)(90)
- [Crise](#)(85)
- [Europa](#)(30)
- [Alemanhã](#)(24)
- [Trump](#)(20)
- [Pandemia](#)(12)

Posicionamento Twitter

Recentes Populares

Anne-Marie

Juntos venceremos o combate contra o Covid 19

January 1, 2014

Neste momento, a rápida propagação global do coronavírus implica que ninguém está a salvo

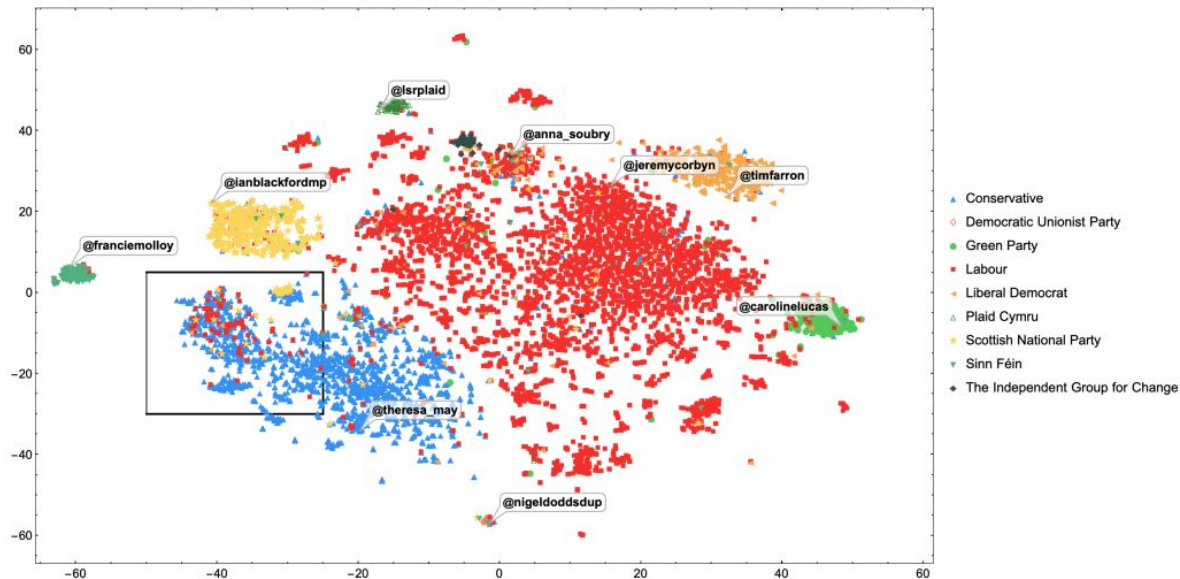
Autores próximos

- António Costa
- José Manuel Fernandes
- Autor 3
- Autor 4
- Autor 5
- Autor 6



Twitter political position

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



166163

Articles

9929 (~700 manually indexed)
Authors

12
Years of publications (2008-2020)





Final remarks

- 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 SFRH/BPD/104176/2014, as well as through the INESC-ID multi-annual funding from the PIDDAC programme, which has the reference UID/CEC/50021/2013, and FEDER under the project 22153-01/SAICT/2016