The ‘Arquivo de Opinião’ archive

Miguel Won

TPDL 2018

Presentation template by SlidesCarnival

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

inesc id lisboa

INCD
Infraestrutura Nacional de Computação Distribuída

About me

Miguel Won

2015-2018: FCT Postdoc researcher at INESC-ID in the field of Computational Social Science

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: journalistic 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

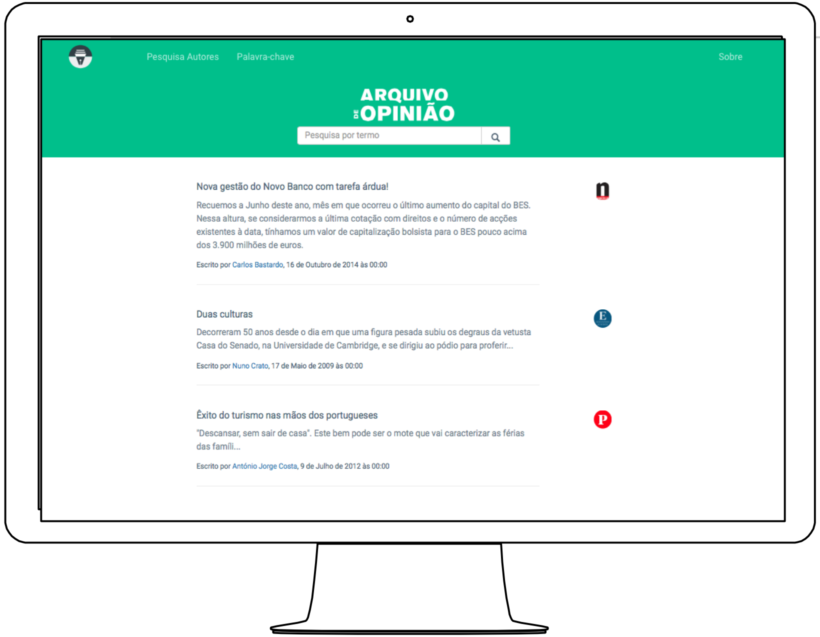
- Essential component of the public debate

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



Nova gestão do Novo Banco com tarefa árdua!

Recuemos 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, tínhamos um valor de capitalização bolsista para o BES pouco acima dos 3.900 milhões de euros.

Escrito por Carlos Bastardo, 16 de Outubro de 2014 às 00:00

Duas culturas

Decorreram 50 anos desde o dia em que uma figura pesada subiu os degraus da vetusta Casa do Senado, na Universidade de Cambridge, e se dirigiu ao pódio para proferir...

Escrito por Nuno Crato, 17 de Maio de 2009 às 00:00

Êxito do turismo nas mãos dos portugueses

"Descansar, sem sair de casa". Este bem pode ser o mote que vai caracterizar as férias das famíli...

Escrito por António Jorge Costa, 9 de Julho de 2012 às 00:00

2

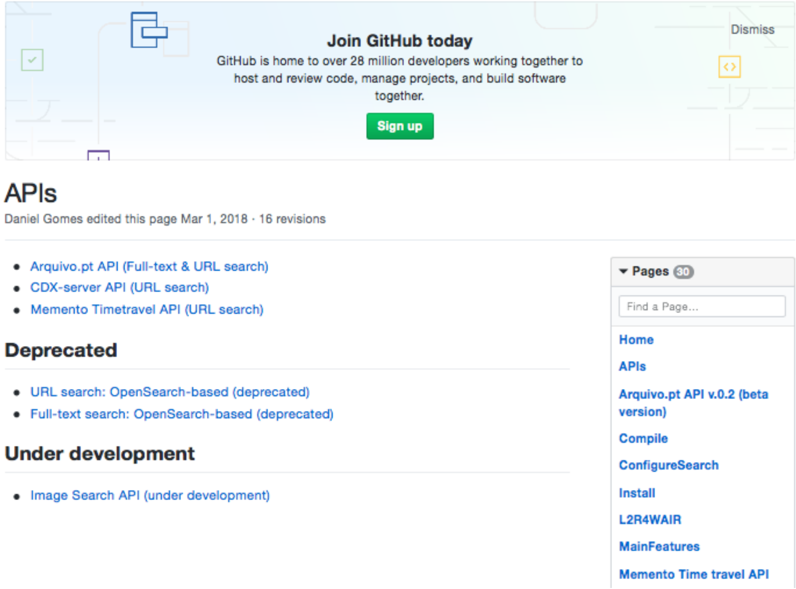
Collect & process

Arquivo construction

Data sources

- Arquivo.pt: web archive of .pt domain

- Opinion section (online)



Join GitHub today
GitHub is home to over 28 million developers working together to host and review code, manage projects, and build software together.
Sign up
Dismiss

APIs

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

- Arquivo.pt API (Full-text & URL search)
- CDX-server API (URL search)
- Memento Timetravel API (URL search)

Deprecated

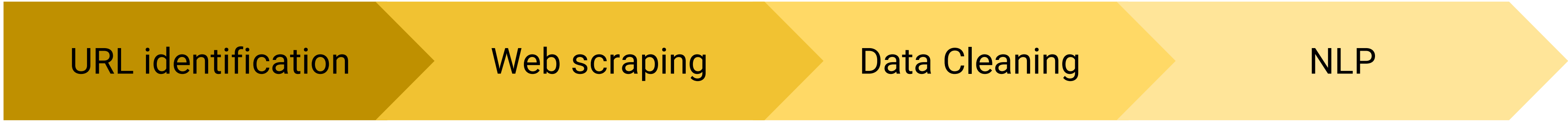
- URL search: OpenSearch-based (deprecated)
- Full-text search: OpenSearch-based (deprecated)

Under development

- Image Search API (under development)

Pages 30
Find a Page…
Home
APIs
Arquivo.pt API v.0.2 (beta version)
Compile
ConfigureSearch
Install
L2R4WAIR
MainFeatures
Memento Time travel API

Pipeline



URL identification

Web scraping

Data Cleaning

NLP

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

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

- Name correction
- Remove html code
- Manual inspection

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

Pipeline

Tools:

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

Web framework:

- Django
- MongoDB

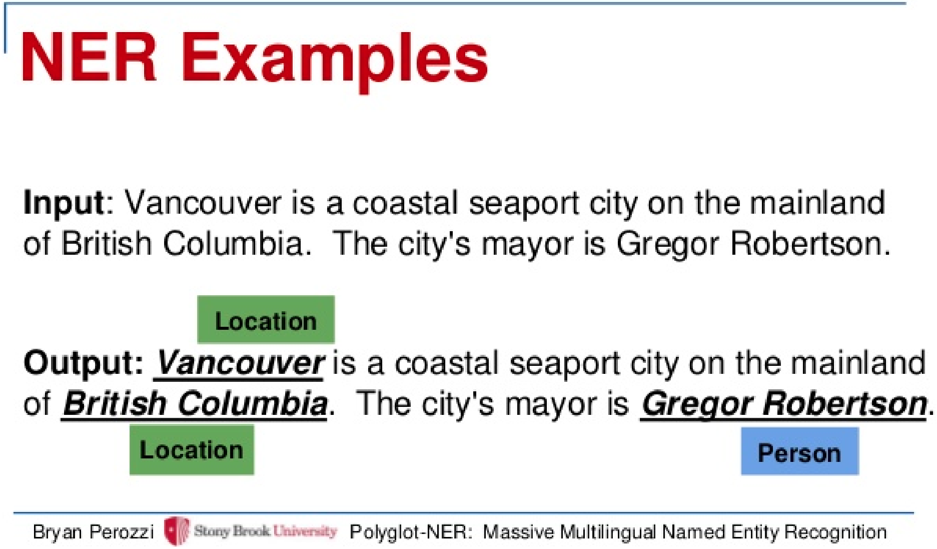
3

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

Bryan Perozzi Stony Brook University Polyglot-NER: Massive Multilingual Named Entity Recognition

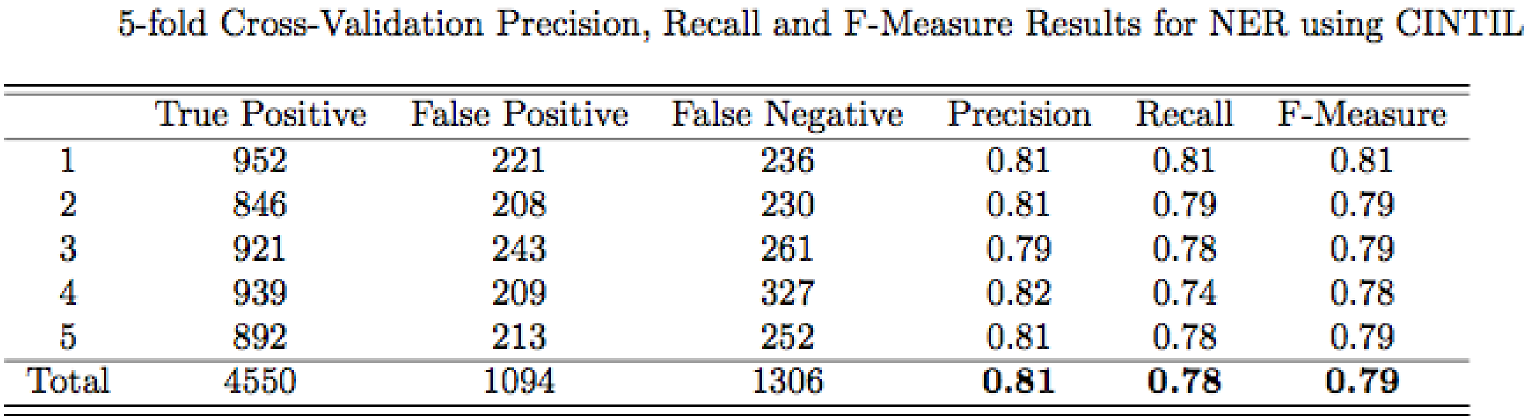
Named Entity Recognition (cont.)

- 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

“Automatic extraction of relevant key-phrases for the study of issue competition”, work in progress with Bruno Martins (INESC-ID) and Filipa Raimundo (ICS)

- 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

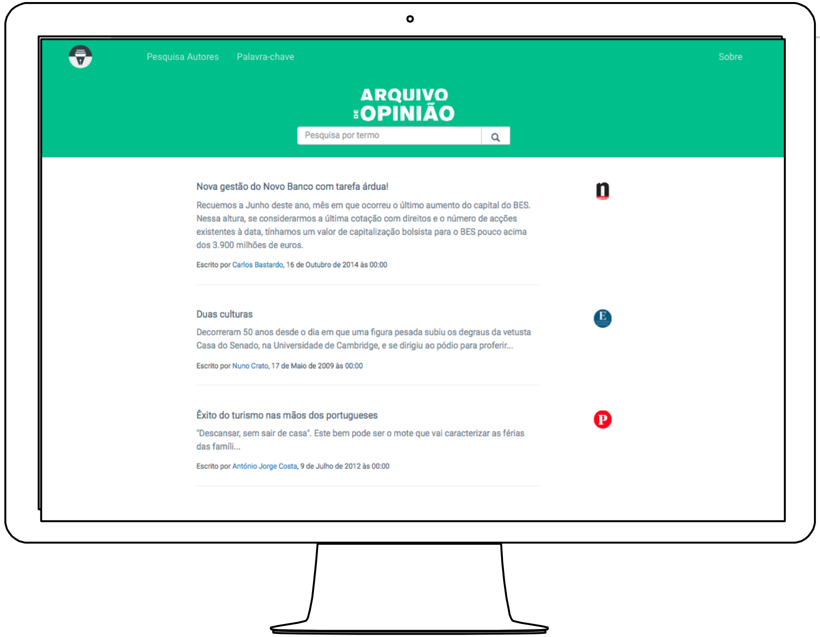
4

Arquivo de Opinião

Opinion in the Portuguese media

Arquivo de Opinião

- Frontpage with a search engine



Nova gestão do Novo Banco com tarefa árdua!

Recuemos 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, tínhamos um valor de capitalização bolsista para o BES pouco acima dos 3.900 milhões de euros.

Escrito por Carlos Bastardo, 16 de Outubro de 2014 às 00:00

Duas culturas

Decorreram 50 anos desde o dia em que uma figura pesada subiu os degraus da vetusta Casa do Senado, na Universidade de Cambridge, e se dirigiu ao pódio para proferir...

Escrito por Nuno Crato, 17 de Maio de 2009 às 00:00

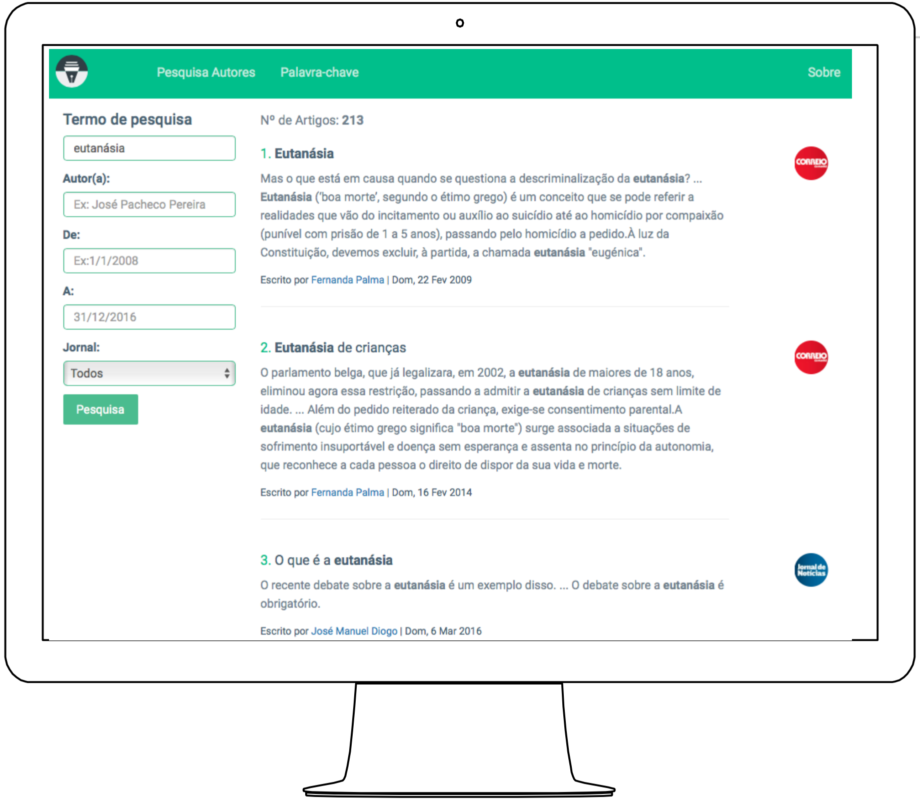
Êxito do turismo nas mãos dos portugueses

"Descansar, sem sair de casa". Este bem pode ser o mote que vai caracterizar as férias das famíli...

Escrito por António Jorge Costa, 9 de Julho de 2012 às 00:00

Search engine (mongo)

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



Termo de pesquisa

eutanásia

Autor(a): Ex: José Pacheco Pereira

De: Ex:1/1/2008

A: 31/12/2016

Jornal: Todos

Nº de Artigos: **213**

1. **Eutanásia**

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".

Escrito por Fernanda Palma | Dom, 22 Fev 2009

2. **Eutanásia** de crianças

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.

Escrito por Fernanda Palma | Dom, 16 Fev 2014

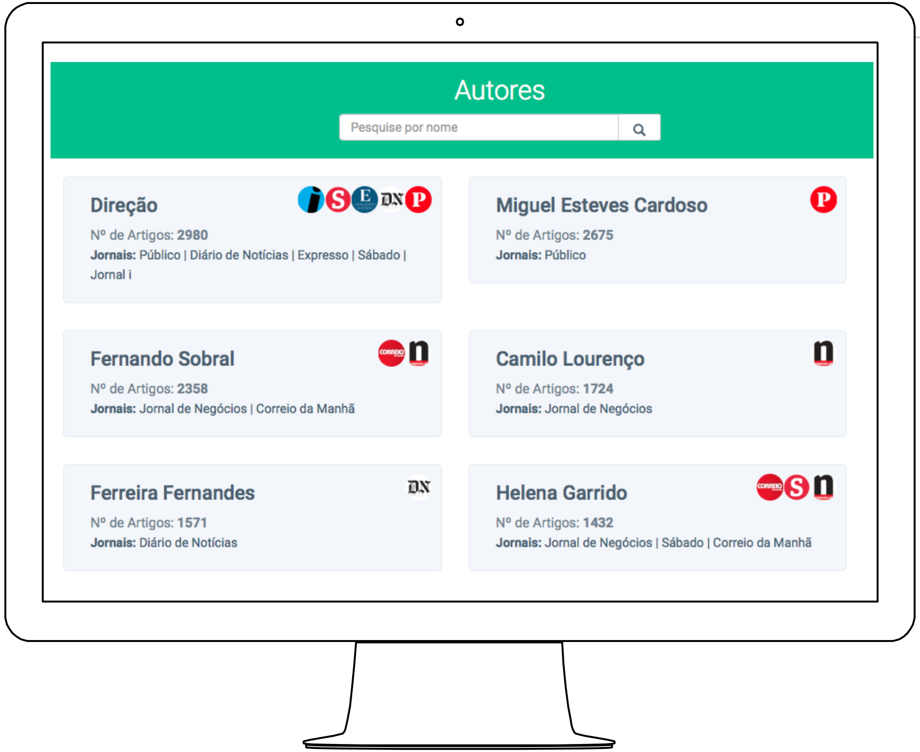
3. O que é a **eutanásia**

O recente debate sobre a **eutanásia** é um exemplo disso. ... O debate sobre a **eutanásia** é obrigatório.

Escrito por José Manuel Diogo | Dom, 6 Mar 2016

Author

- Search for author
- ~3500 available authors



Autores

Direção
Nº de Artigos: **2980**
Jornais: Público | Diário de Notícias | Expresso | Sábado | Jornal i

Miguel Esteves Cardoso
Nº de Artigos: **2675**
Jornais: Público

Fernando Sobral
Nº de Artigos: **2358**
Jornais: Jornal de Negócios | Correio da Manhã

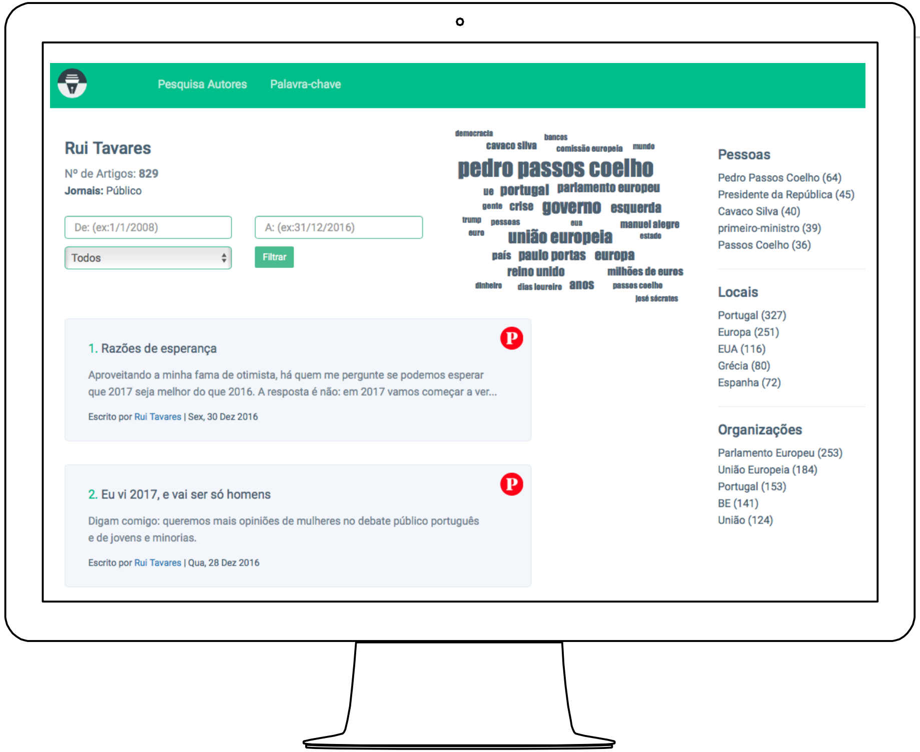
Camilo Lourenço
Nº de Artigos: **1724**
Jornais: Jornal de Negócios

Ferreira Fernandes
Nº de Artigos: **1571**
Jornais: Diário de Notícias

Helena Garrido
Nº de Artigos: **1432**
Jornais: Jornal de Negócios | Sábado | Correio da Manhã

Author (cont.)

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



Rui Tavares

Nº de Artigos: 829
Jornais: Público

1. Razões de esperança

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

Digam comigo: queremos mais opiniões de mulheres no debate público português e de jovens e minorias.

Escrito por Rui Tavares | Qua, 28 Dez 2016

Pessoas

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

Locais

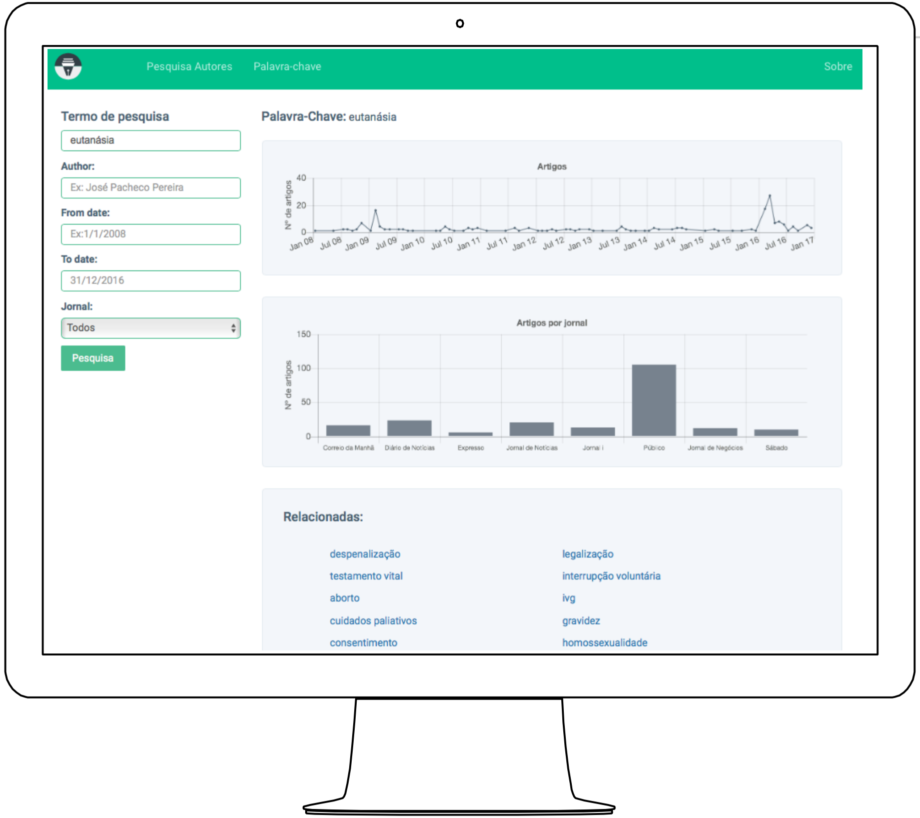
Portugal (327)
Europa (251)
EUA (116)
Grécia (80)
Espanha (72)

Organizações

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

Key-phrases

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



Termo de pesquisa

eutanásia

Author:

Ex: José Pacheco Pereira

From date:

Ex:1/1/2008

To date:

31/12/2016

Jornal:

Todos

Palavra-Chave: eutanásia

Relacionadas:

despenalização
testamento vital
aborto
cuidados paliativos
consentimento

legalização
interrupção voluntária
ivg
gravidez
homossexualidade

85 530

Articles

3571

Authors

30 000

key-phrases

9

Years of publications (2008-2016)

5

Next steps and final remarks

Version 2.0

- Add additional sources: Observador, O Jornal Económico
- 2016-Present
- Social Media:
  - Authors pages
  - Shares, likes, etc.
  - Networks
- Real time monitoring (daily, weekly?)
- Add more NLP metrics: topic modeling, sentiment, etc.

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)
- Future work will be carried out towards the inclusion of external data, in particular from social media

Arquivo.pt awards (3rd place)

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