The ‘Arquivo de Opinião’ archive

Miguel Won
TPDL 2018
About me

Miguel Won

2015-2018: FCT Postdoc researcher at INESC-ID in the field of Computational Social Science
miguelwon@tecnico.ulisboa.pt
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 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)**
Pipeline

**URL identification**
- Search for clues such as “opiniao”
- 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)
- 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.

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

<table>
<thead>
<tr>
<th></th>
<th>True Positive</th>
<th>False Positive</th>
<th>False Negative</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>952</td>
<td>221</td>
<td>236</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>846</td>
<td>208</td>
<td>230</td>
<td>0.81</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>921</td>
<td>243</td>
<td>261</td>
<td>0.79</td>
<td>0.78</td>
<td>0.79</td>
</tr>
<tr>
<td>4</td>
<td>939</td>
<td>209</td>
<td>327</td>
<td>0.82</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>892</td>
<td>213</td>
<td>252</td>
<td>0.81</td>
<td>0.78</td>
<td>0.79</td>
</tr>
<tr>
<td>Total</td>
<td>4550</td>
<td>1094</td>
<td>1306</td>
<td>0.81</td>
<td>0.78</td>
<td>0.79</td>
</tr>
</tbody>
</table>
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
Arquivo de Opinião

Opinion in the Portuguese media
Arquivo de Opinião

◉ Frontpage with a search engine
Search engine (mongo)

- Search for text of 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
Key-phrases

- Search indexed key-phrases (with autocomplete)

- Outputs
  - No. articles by time and source
  - Related (word embeddings)
85,530 Articles
3,571 Authors
30,000 key-phrases
9 Years of publications (2008-2016)
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