

A transnational and cross-lingual crawl of the European Parliamentary Elections 2019

Ivo.Branco@fccn.pt

RicardoBasilio@fcsh.unl.pt

Daniel.Gomes@fccn.pt

ARQUIVO.PT

Cristiano Ronaldo

1996
1 Jan

2021
8 Mar

Pages

Images

Advanced search

About 72,340,359 results since 1996 until 2021

→ blogs.sapo.pt/userinfo.bml?user=jurgensimma

Blogs do SAPO: Perfil Público

14 August, 2010

Blogs do SAPO: Perfil Público SAPO Blogs Os Meus Blogs perfil público *Cristiano **Ronaldo** 9* didothebest@live.it TV Universo: A Companhia para o Verão! <http://tvuniverso.blogs.sapo.pt> Blog que fala sobre tudo o que se passa no mundo da Televisão Portuguesa. Outros Autores: bvale Tv Universo: Ve aqui ...

ARQUIVO.PT

Cristiano Ronaldo

1996
1 Jan


2021
8 Mar

Pages

Images


Advanced search

About 43,466 results since 1996 until 2021



→ oikos.pt:80/galeria/album01/DS...

30 November at 10:05, 2006



→ 10mais.blogs.sapo.pt

24 October at 01:24, 2009

Motivation

European Parliamentary Elections

- International relevance
- Cross border
- Cross lingual

Collaboration

1	Sheet description
2	This page contains contributes from Europe to the collaborative list, before European
3	Title: Contribute From Europe
4	Date: April 23 to May 22, 2019
5	
6	Collaborative list of seeds
7	European Elections 2019
8	These websites will be preserved by Arquivo.pt Your help is fundamental! Help us by adding more websites to the following list so we don't lose a part of our History
9	Websites about the 2019 European Elections (e.g: news and sections in the media, blogs, think pie You can create new sections. Or you can simply add your suggestion here:
10	
11	
12	Official Websites:
13	http://www.europarl.europa.eu/at-your-service/en/be-heard/elections
14	https://www.european-elections.eu/why-you-should-vote
15	https://yourvotematters.eu/en/
16	https://yourvotematters.eu/en/page/about-the-european-parliament-and-ep-elections-2019
17	https://administracion.gob.es/pag_Home/etencionCiudadana/Elecciones_locales_europeas_2019.htm



Contributions per Country

Country	#
Slovakia	114
Portugal	114
France	78
Other countries	226
Not mentioned	91
Total	680

Automatic translation



PT - Portuguese	BG - Bulgarian	HR - Croatian	CS - Czech	DA - Danish	NL - Dutch	EN	ET
eleições parlamento europeu 2019	изборите за Е	izbori za Europ	Volby do Evrop	valg til Europa-	Europese verki	European Parli	Eu
eleições europeias 2019	Европейски из	Europski izbori	Volby do Evrop	Europæiske va	Europese verki	European elect	Eu
abstenção europeias 2019	Европейската	Europska apsti	Evropská zdrže	Europæisk afst	Europese onth	European abst	Eu
resultados eleitorais 2019	резултатите о	Rezultati izbora	Výsledky voleb	valgresultater 2	verkiezingsuits	election results	val
vencedores europeias 2019	Европейските	Europski pobje	Evropští vítězo	Europæiske vir	Europese winn	European winn	Eu
brexit eleições europeias 2019	brexit европей	brexit europski	Volby brexit ev	valget brexit eu	brexit Europese	brexit European	bre
partidos eleições europeias 2019	партиа европ	stranke europs	strany Evropsk	parties europæ	partijen Europe	parties Europe	po
eleições europeias blogs	блогове избор	blogovi europsi	blogy volby do	valget blogs eu	blogs Europese	blogs European	blo

- Automatic translation to the 24 EU official languages
- Publications Office of the European Union reviewed the translations



Automatic search

```
import argparse
import os
import time

# encoding=utf8
import sys
reload(sys)
sys.setdefaultencoding('utf8')

from azure.cognitiveservices.search.websearch import WebSearchAPI
from msrest.authentication import CognitiveServicesCredentials

parser = argparse.ArgumentParser(description="Bing WebSearch API v7")
parser.add_argument(dest='queries_list_file', nargs='?', default="queries_list.txt",
                    help="File with list of queries to be performed.")
parser.add_argument(dest='results_file', nargs='?', default="query_results.tsv", help="")
parser.add_argument(dest='results_number', nargs='?', type=int, default=20, help="Number of results to return")
parser.add_argument(dest='subscription_key_env_name', nargs='?', default="AZURE_KEY",
                    help="Environment variable with Bing Service Subscription Key")

args = parser.parse_args()

subscription_key = os.environ[args.subscription_key_env_name]

client = WebSearchAPI(CognitiveServicesCredentials(subscription_key))
```

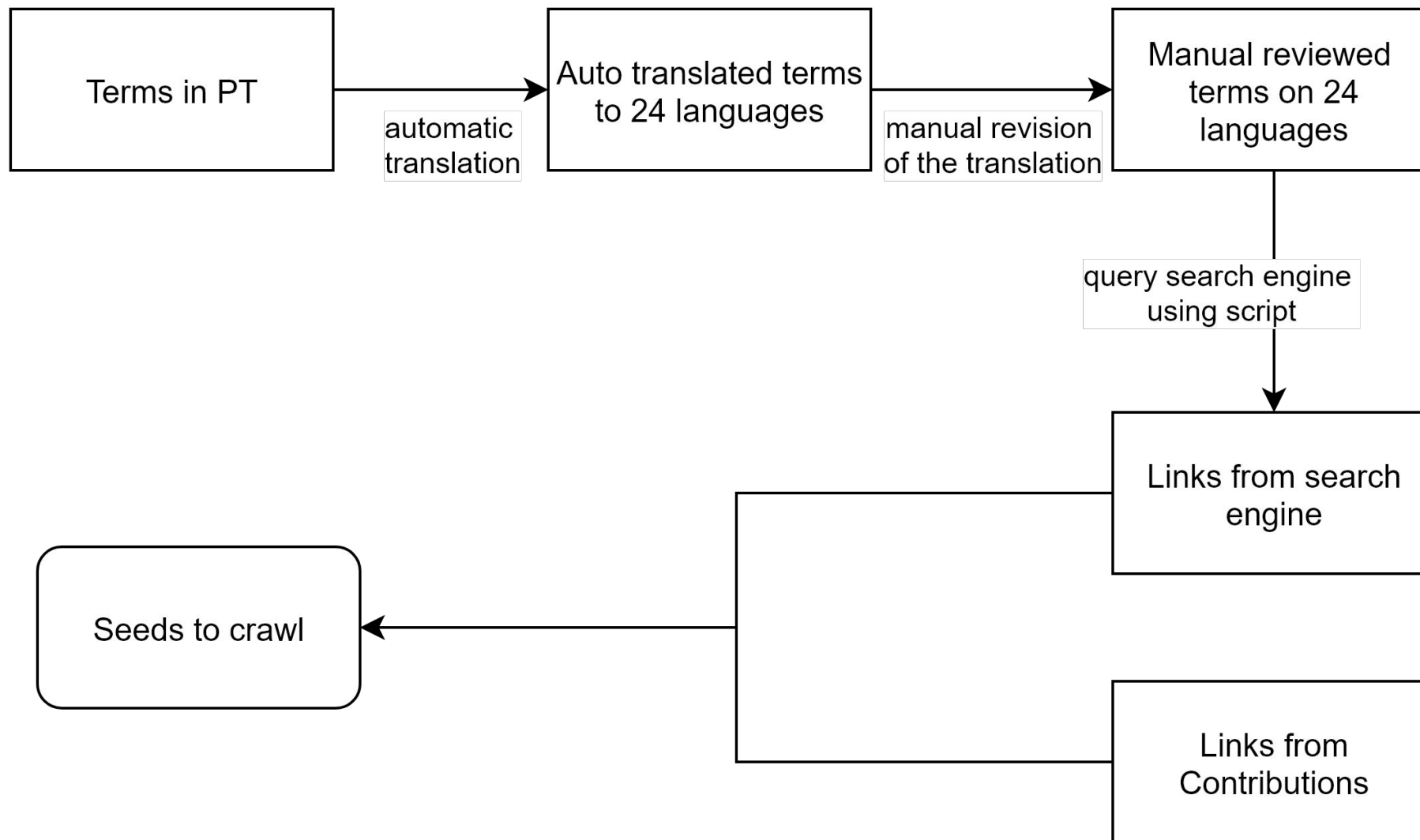
```
with open(args.queries_list_file, mode='r') as input_file:
    for line in input_file:
        query = line.rstrip()
        print("Querying for: {}".format(query))
        try:
            web_data = client.web.search(query=query, count=args.results_number)

            if web_data.web_pages.value:
                print("Webpage Results #{}".format(len(web_data.web_pages.value)))

                with open(args.results_file, mode='a') as output_file:
                    for i, result in enumerate(web_data.web_pages.value):
                        print("Writing result: {}".format(result.url))
                        output_file.write("{}\t{}\t{}\t{}\n".format(query, i+1, result.url, result.snippet))

        except Exception as err:
            print("Exception {}".format(err))
```

Schema



Crawls

1. Pre-elections crawl using Heritrix
 - collaborative list
 - couple of sentences on a search engine
2. Post-elections automatic translation + query search engine
 - 12 terms per language on 24 languages
3. Manual revision of the translations + query search engine

Review post-elections crawl:

- increase the number from 12 to 40 search terms per language
- run a high quality crawl for each seed

Crawls

4. Brozzler \Rightarrow to increase quality

- run a single page crawl on each seed
- uses a real browser to better preserve CSS+JS

5. Browsertrix

- for youtube videos
- because lack compatibility on Brozzler crawler and pywb replay

6. Heritrix

- final crawl

arquivo.pt/ee2019

- full text page search
- image search
- page replay

Português

RESEARCH
European Elections 2019



Search pages from the past

[About this project](#)

EE2019 Numbers

- Total seeds/preserved sites: 12147
- From: 2019-05-22
- Until: 2019-07-17
- Disk space: 4.8 TB
- 99 millions collected files



The screenshot shows the ARQUIVO.PT search interface. At the top is a dark blue header with a hamburger menu icon on the left and the text 'ARQUIVO.PT' on the right. Below the header is a search bar with the placeholder text 'Palavras ou URL' and a magnifying glass icon on the right. Under the search bar is a date range selector. It features two calendar icons: the left one shows '1996' and '1 Jan', and the right one shows '2021' and '9 Mar'. A horizontal blue line with two dots connects the two dates. Below the date range are three buttons: 'Páginas', 'Imagens', and 'Pesquisa avançada'.

arquivo.pt/api

- available for automatic programs
- APIs
- Arquivo.pt API
- CDX
- Memento
- Image search

```
{
  "serviceName": "Arquivo.pt - the Portuguese web-archive",
  "linkToService": "https://arquivo.pt",
  "next_page": "https://arquivo.pt/textsearch?q=european%20elections%202019&from=19960101000000&to=20210309235959&offset=10&maxItems=10&siteSearch=&type=&collection=&trackingId=alc2075e4d79c35f0577_3870d3688288fe05bab5",
  "estimated_nr_results": 1420573,
  "request_parameters": {
    "offset": 0,
    "dedupValue": 2,
    "from": "19960101000000",
    "to": "20210309235959",
    "type": [],
    "collection": [],
    "dedupField": "site",
    "q": "european elections 2019",
    "maxItems": 10,
    "siteSearch": []
  },
  "response_items": [
    {
      "title": "European Elections 2019",
      "originalURL": "https://www.european-elections.eu/",
      "linkToArchive": "https://arquivo.pt/wayback/20190228180836/https://www.european-elections.eu/",
      "tstamp": "20190228180836",
      "contentLength": 16149,
      "digest": "14bdfb2c4b8cacadcde916b1a6cc7075",
      "mimeType": "text/html",
      "encoding": "UTF-8",
      "date": "1551377316",
      "linkToScreenshot": "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20190228180836%2Fhttps%3A%2F%2Fwww.european-elections.eu%2F",
      "linkToNoFrame": "https://arquivo.pt/noFrame/replay/20190228180836/https://www.european-elections.eu/",
      "linkToExtractedText": "https://arquivo.pt/textextracted?m=https%3A%2F%2Fwww.european-elections.eu%2F%2F20190228180836",
      "linkToThumbnail": "https://arquivo.pt/thumbnail/20190228180836/https://www.european-elections.eu/"
    }
  ]
}
```

Conclusion

Using this method we have combined:

Human expertise + Automation

⇒ Maximize coverage of a cross-lingual event

⇒ using very limited amount of resources

This process can be applied easily to any kind of transnational event.

COVID-19 outbreak

- IIPC Collaborative Collection
- Automatically queried a web search engine
 - For each portuguese news site search “coronavirus site:news-site.pt”
 - use 10 first search results
- Result on a spreadsheet with
 - query input
 - position 1-10
 - page URL
 - page title
- Manual revision by web curator team member
 - remove some sites
 - by default single page crawl
 - other increase the depth to + 1 hop



Thank you !

More information

arquivo.pt/ee2019 > [About this project](#)

Português

RESEARCH
European Elections 2019



Search pages from the past

[About this project](#)