Searching images from the past

fernando.melo@fccn.pt

Arquivo.pt

https://arquivo.pt

From Portugal

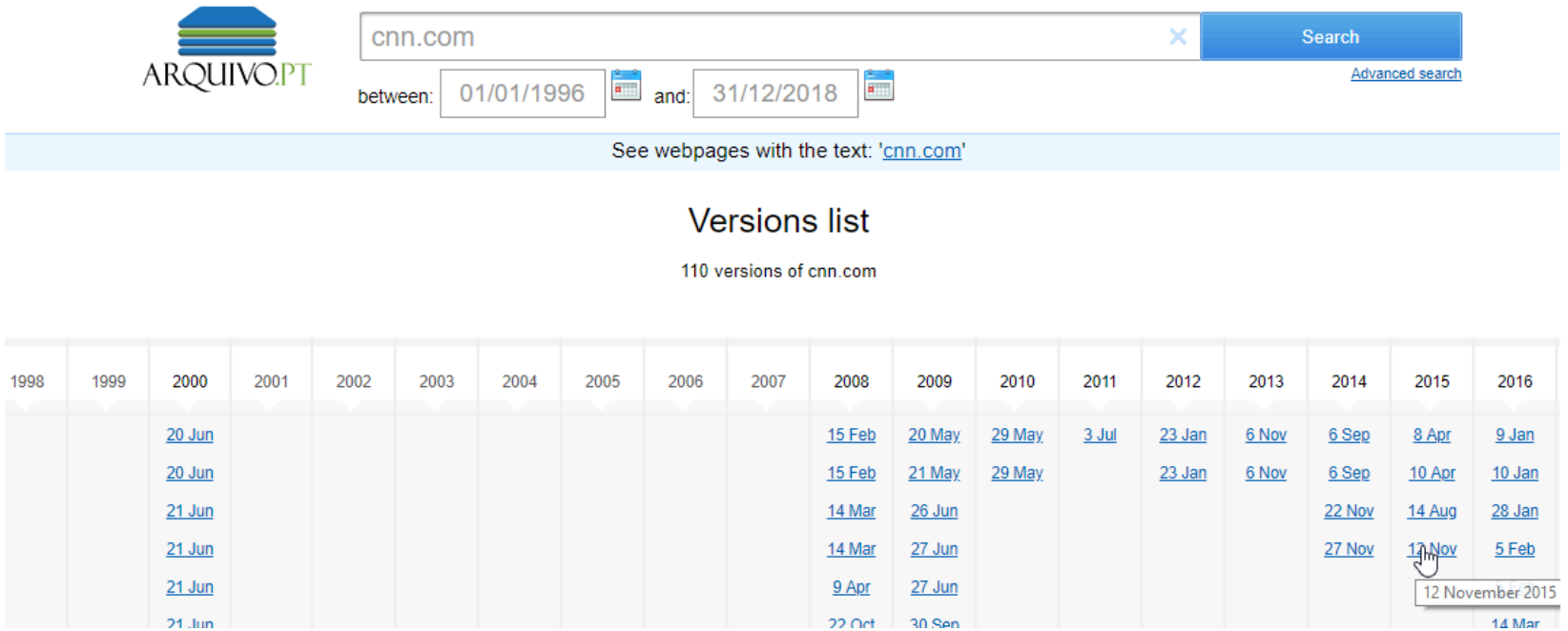
Publicly available web archive

Research infrastructure

Source code on Github github.com/arquivo

Free

URL Search


ARQUIVO.PT
cnn.com
Search
Advanced search
between: 01/01/1996
and: 31/12/2018
See webpages with the text: 'cnn.com'
Versions list
110 versions of cnn.com
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
20 Jun 15 Feb 20 May 29 May 3 Jul 23 Jan 6 Nov 6 Sep 8 Apr 9 Jan
20 Jun 15 Feb 21 May 29 May 23 Jan 6 Nov 6 Sep 10 Apr 10 Jan
21 Jun 14 Mar 26 Jun 22 Nov 14 Aug 28 Jan
21 Jun 14 Mar 27 Jun 27 Nov 12 Nov 5 Feb
21 Jun 9 Apr 27 Jun
12 November 2015

Text Search

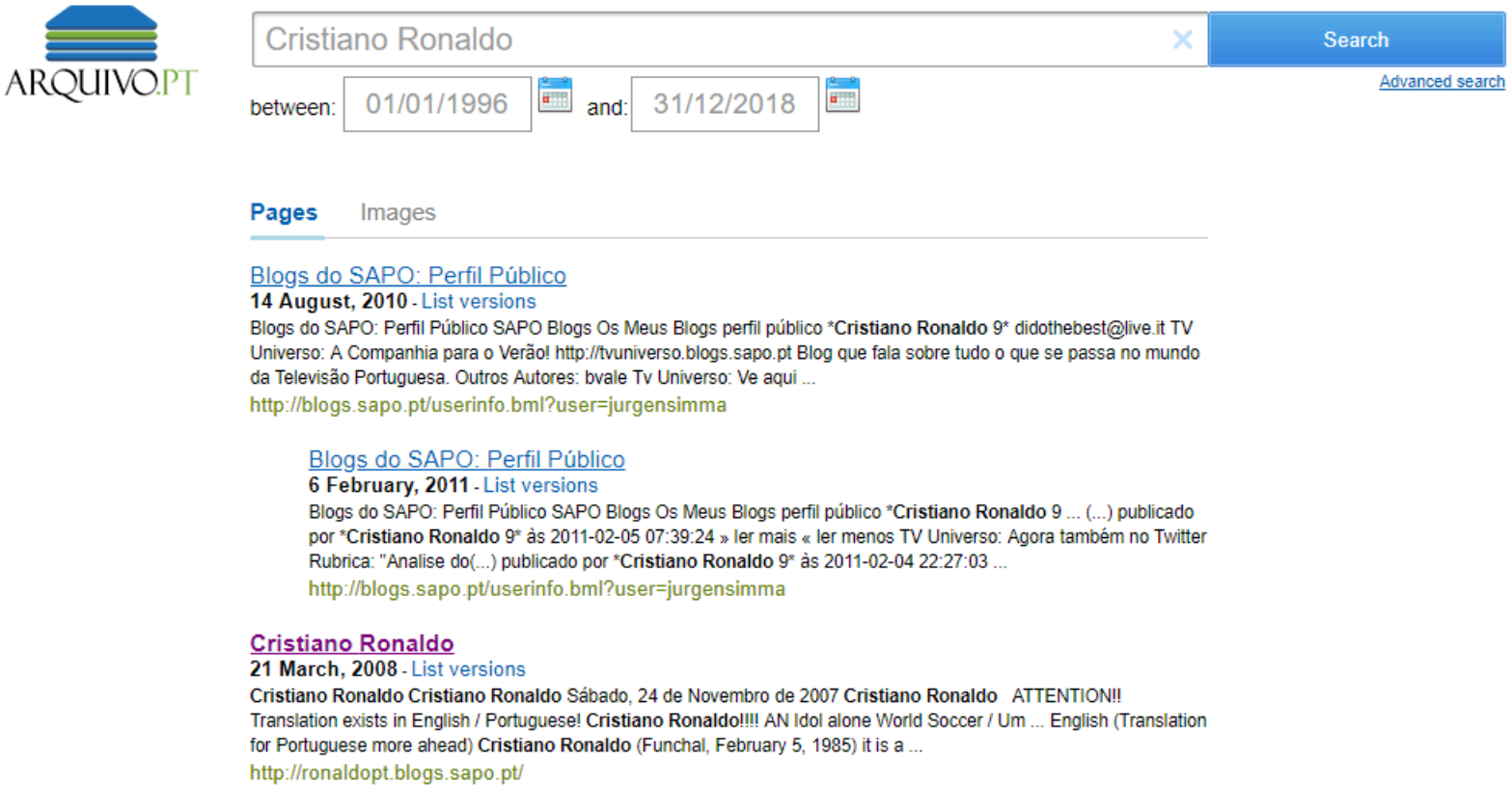

ARQUIVO.PT
Cristiano Ronaldo
Search
Advanced search
between: 01/01/1996
and: 31/12/2018
Pages
Images
Blogs do SAPO: Perfil Público
14 August, 2010 - List versions
Blogs do SAPO: Perfil Público SAPO Blogs Os Meus Blogs perfil público *Cristiano Ronaldo 9* didothebest@live.it TV Universo: A Companhia para 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 ...
http://blogs.sapo.pt/userinfo.bml?user=jurgensimma
Blogs do SAPO: Perfil Público
6 February, 2011 - List versions
Blogs do SAPO: Perfil Público SAPO Blogs Os Meus Blogs perfil público *Cristiano Ronaldo 9 ... (...) publicado por *Cristiano Ronaldo 9* às 2011-02-05 07:39:24 » ler mais « ler menos TV Universo: Agora também no Twitter Rubrica: "Analise do(...) publicado por *Cristiano Ronaldo 9* às 2011-02-04 22:27:03 ...
http://blogs.sapo.pt/userinfo.bml?user=jurgensimma
Cristiano Ronaldo
21 March, 2008 - List versions
Cristiano Ronaldo Cristiano Ronaldo Sábado, 24 de Novembro de 2007 Cristiano Ronaldo ATTENTION!! Translation exists in English / Portuguese! Cristiano Ronaldo!!!! AN Idol alone World Soccer / Um ... English (Translation for Portuguese more ahead) Cristiano Ronaldo (Funchal, February 5, 1985) it is a ...
http://ronaldopt.blogs.sapo.pt/

Image Search[new]

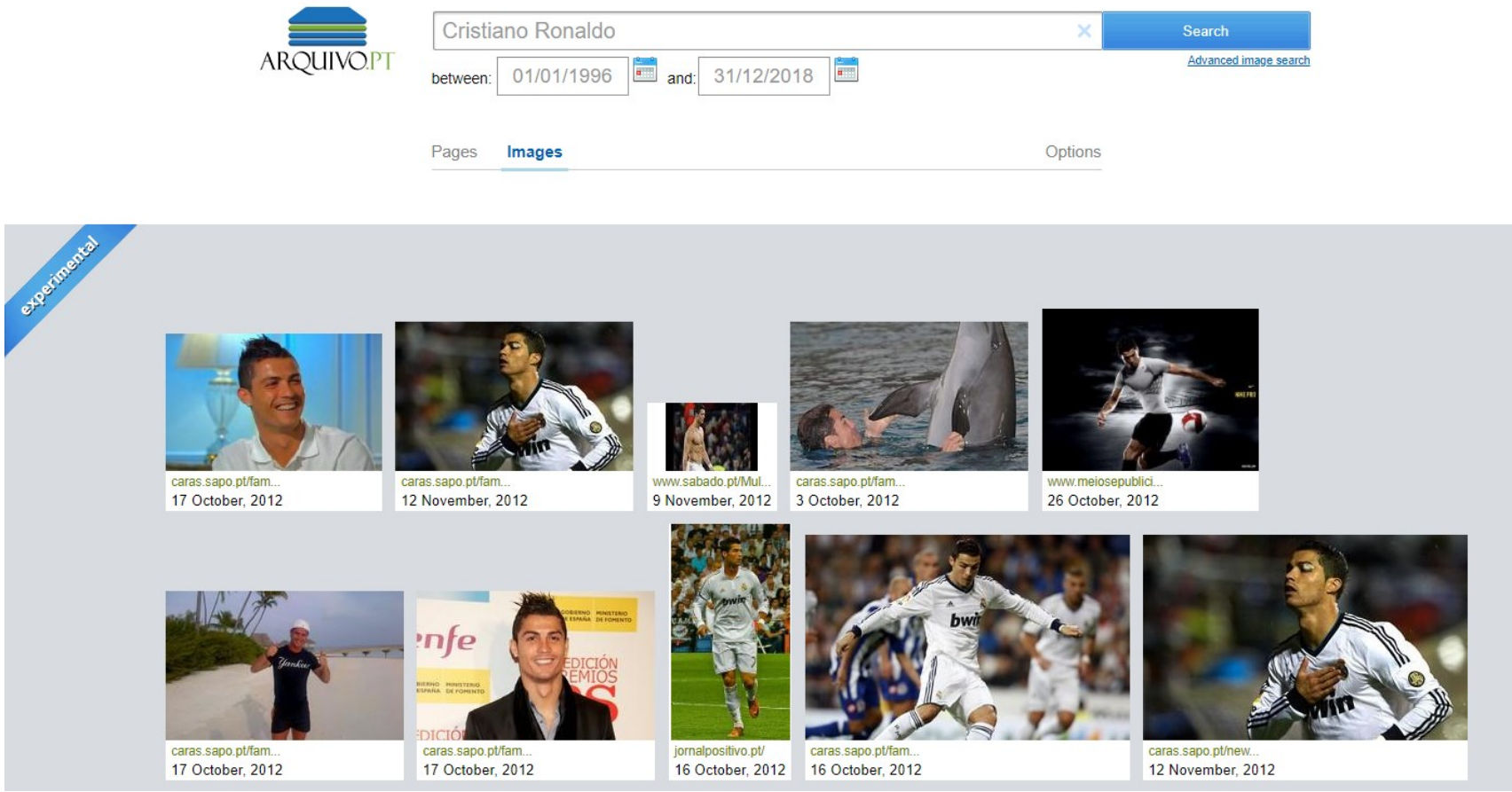


Image Search - viewer

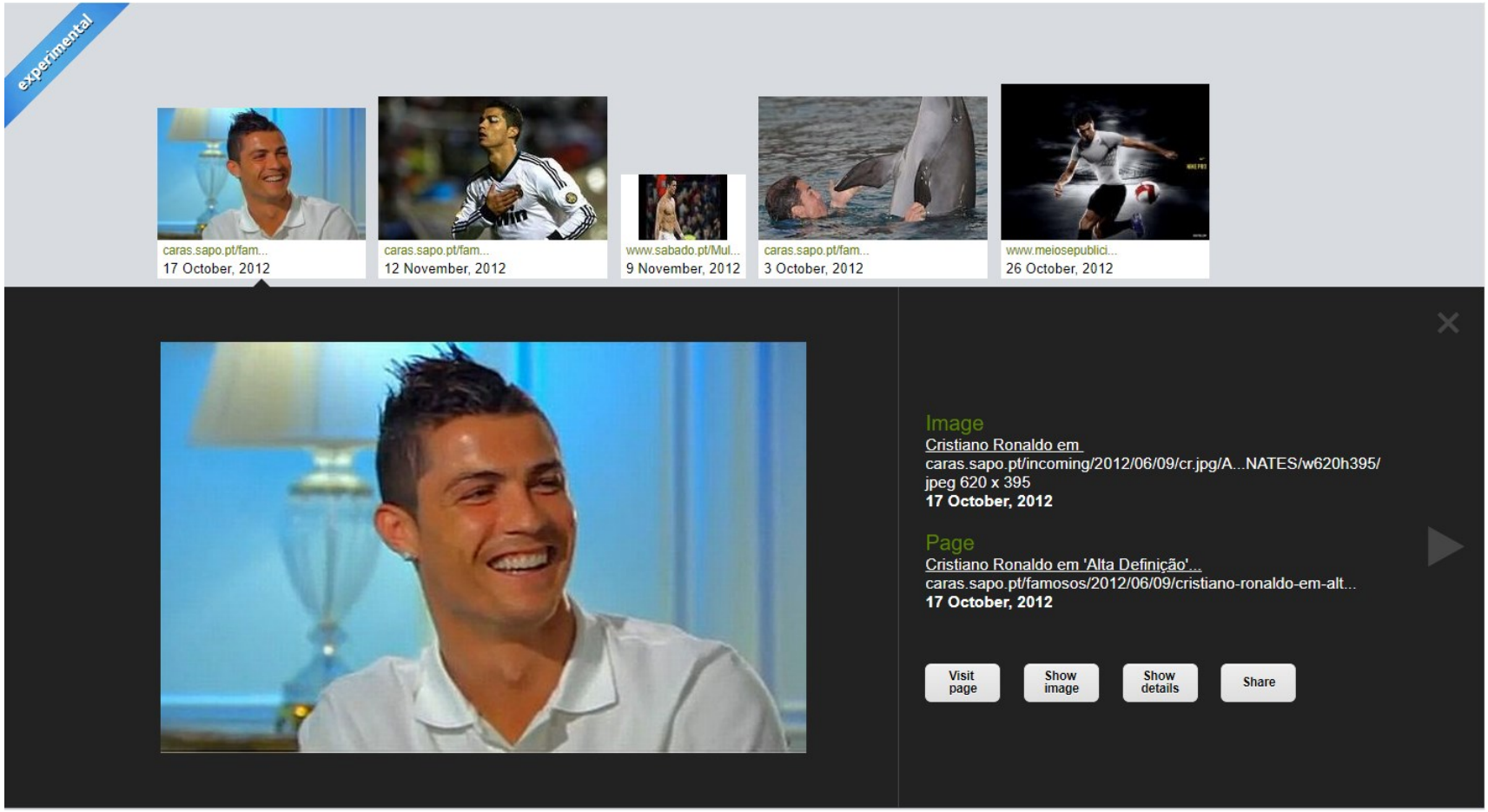


Image Search - **estimate**

6 000 000 000 files

~15% images

900 million searchable images

We need more servers!!!

Image Search - **reality**

17 million searchable images

Unique images within a collection

From **1996** to **2017**

Size greater than **50px width** and **50px height**

Each image **has to link to** an archived web page that contains the image.

Image Search **Workflow**

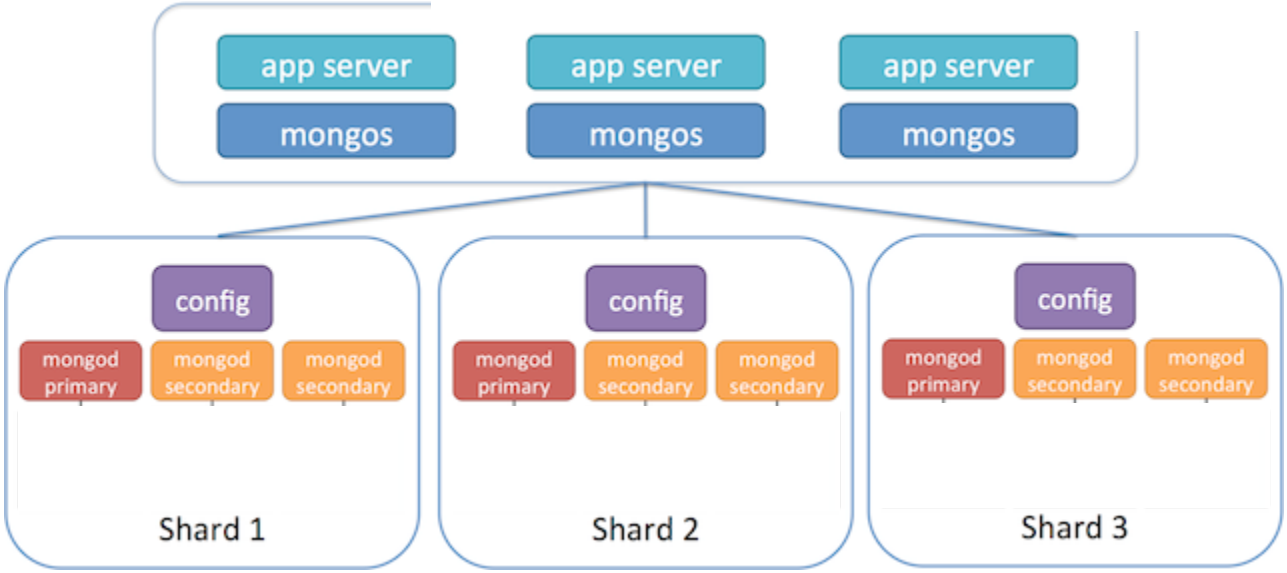
Workflow

1. Create image indexes from ARC/WARC files
2. Image classification
3. Solr indexing

1. Create image

Hadoop 3 cluster

MongoDB sharde



1. Create image indexes - steps

1.1 Extract all ARC/WARC **image** records

store Image Record

--image thumbnail

--image attributes

1.2 Extract all ARC/WARC **html** records

- Extract **<img>** tags in each html record
- If image exists in the database

store Image Index

--page URL

--page timestamp

--page title

2. Image classification - **infrastructure**

2 Tesla P4 GPUs

2. Image classification - **step**

Automatically classify images from step 1

Safe for work:

- Value from 0.000 to 1.000
- Greater than 0.500 (considered safe for work)
- Less than 0.500 (may have explicit content)

Add more classifiers in the future

3. SOLR indexing - **infrastructure**

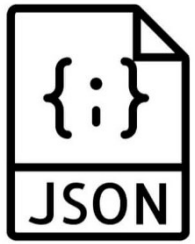
2 Servers with Apache Solr

~ 2-3GB of Ram

~ 400 GB of disk space (indexes)

Configure Solr Cloud in a near future

3. SOLR indexing - **step**

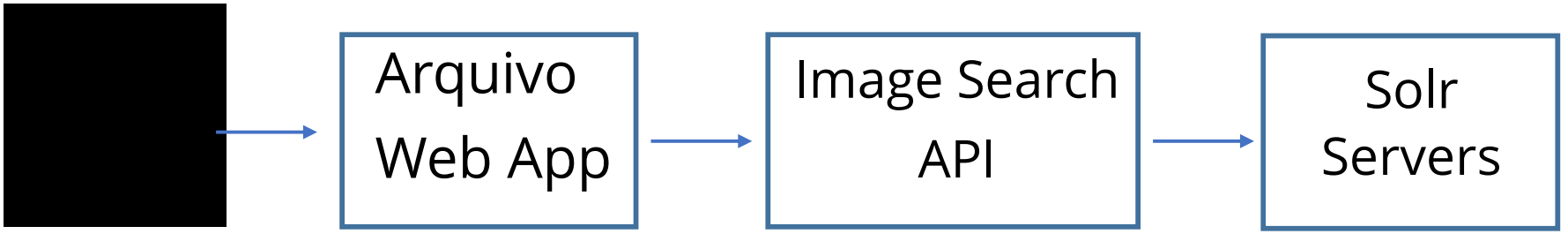


Index JSON image records

3. SOLR index – document


```
{
 "collection": "EAWP12",
 "pageURL": "http://www.acorianooriental.pt/noticia/fc-porto-bate-rio-ave-por-3-1-no-arranque-da-i-liga-video",
 "imgWidth": 540,
 "pageTstamp": 20160817114400,
 "imgAlt": "epa05478211 Rio Ave's Pedrinho (L) and Wakaso (2-R) in action against FC Porto's Corona (C) during
  the Portuguese First League soccer match between Rio Ave and Porto at Arcos Stadium, in Vila do Conde,
  Portugal, 12 August 2016. ",
 "imgHeight": 360,
 "pageTitle": "FC Porto bate Rio Ave por 3-1 no arranque da I Liga (vídeos) - Açoriano Oriental",
 "imgSrc": "http://www.acorianooriental.pt/image_cache/images/view/_video/25474.jpg",
 "imgMimeType": "image/jpeg",
 "imgTstamp": 20160817114410,
 "safe":  0.986
}
```

Image Search - architecture



arquivo.pt/api

Image Search API

Image Search API – **q** parameter

arquivo.pt/imagesearch?**q=soccer**

Search for images
related with word
soccer

Image Search API – **response header**


```
{
"serviceName": "Arquivo.pt - image search service.",
"linkToService": "https://arquivo.pt/images.jsp",
"linkToDocumentation": "https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1-(beta)",
"linkToMoreFields": "https://arquivo.pt/imagesearch?q=soccer&prettyPrint=true&
more=imgThumbnailBase64,imgSrcURLDigest,imgDigest,pageProtocol,pageHost,pageImages,safe",
"nextPage": "https://arquivo.pt/imagesearch?q=soccer&prettyPrint=true&offset=50",
"previousPage": "https://arquivo.pt/imagesearch?q=soccer&prettyPrint=true&offset=0",
"totalItems": 9403,
"numberOfResponseItems": 50,
"offset": 0,
"responseItems": [ ]
}
```

Imagesearch API – **response item**


```
{
  "imgAlt": "soccer",
  "imgTstamp": 20110513171504,
  "imgWidth": 90,
  "imgTitle": "soccer",
  "pageTstamp": 20110513150752,
  "pageTitle": "2011 Maio | Jornal do Centro",
  "pageURL": "http://www.jornaldocentro.pt/?m=201105",
  "collection": "FAWP5",
  "imgMimeType": "image/jpeg",
  "imgSrc": "http://www.jornaldocentro.pt/wp-content/uploads/soccer-90x65.jpg",
  "imgHeight": 65,
  "imgLinkToArchive": "https://arquivo.pt/wayback/20110513171504/http://www.jornaldocentro.pt
  /wp-content/uploads/soccer-90x65.jpg",
  "pageLinkToArchive": "https://arquivo.pt/wayback/20110513150752/http://www.jornaldocentro.pt
  /?m=201105"
},
```

Image Search API – **type** parameter

http://arquivo.pt/imagesearch?q=Euro 2004**&type=png**

Search for images
related with words
Euro 2004,
in **PNG** image format.

Image Search API – **size** parameter

http://arquivo.pt/imagesearch?q=Euro 2004**&size=sm**

Search for images, with size **small** related with words **Euro 2004**,

size=md (medium image size)
size=lg (large image)

Future Work

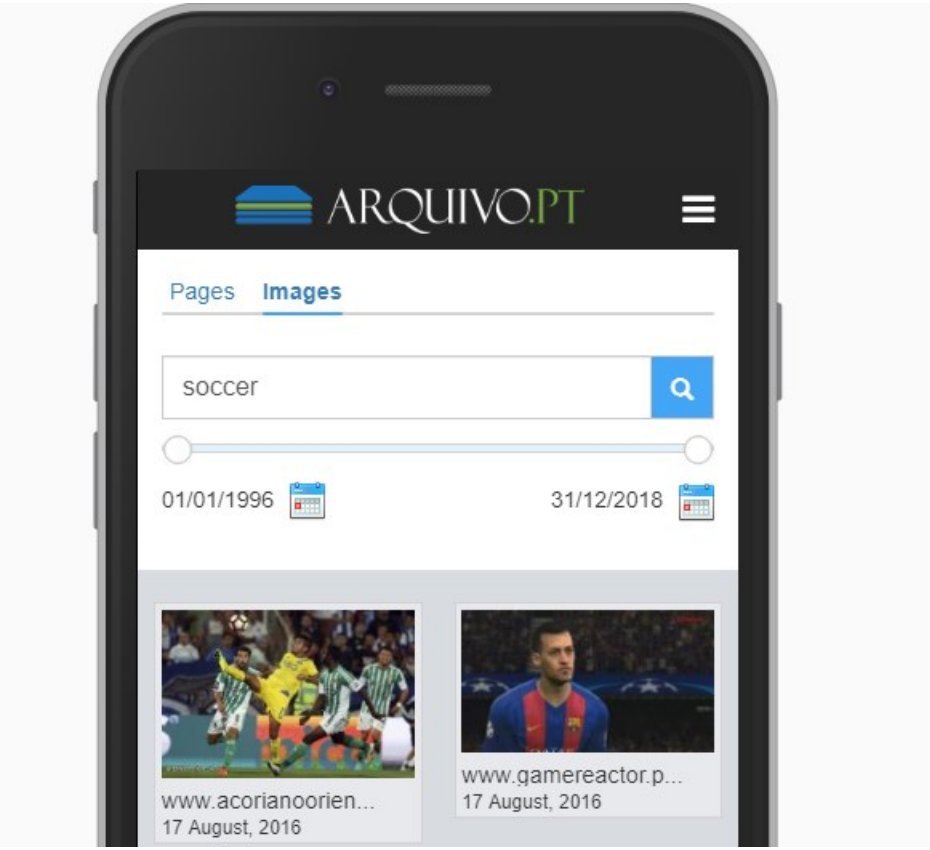
Future work – **scene classification**

http://places2.csail.mit.edu/

Predictions:

- **Type of environment**: indoor
- **Scene categories**: television_studio (0.927)
- **Scene attributes**: no horizon, enclosed area, man-made, indoor lighting, cloth, working, wood, glossy, congregating

Future work – **mobile** version



Future work – **mobile** version

Thank you

Try our APIs and send us feedback
arquivo.pt/api

Fernando Melo < fernando.melo@fccn.pt >