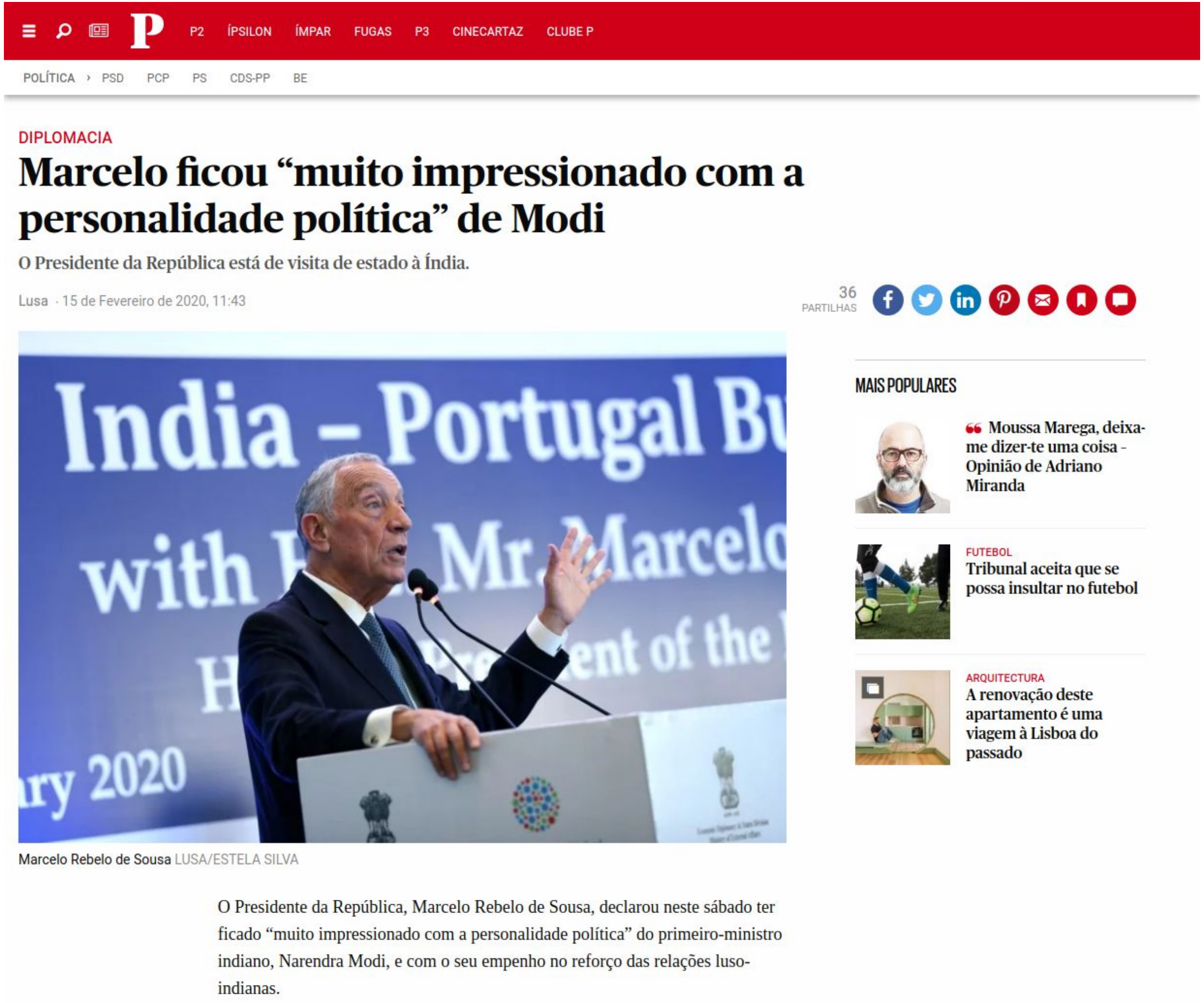
Arquivo.pt image search **2020 → 2021**

February 2nd 2021

André Mourão
R&D engineer
andre.mourao@fccn.pt

ARQUIVO.PT

FCT
Fundação
para a Ciência
e a Tecnologia



DIPLOMACIA

Marcelo ficou “muito impressionado com a personalidade política” de Modi

O Presidente da República está de visita de estado à Índia.

Lusa · 15 de Fevereiro de 2020, 11:43

Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

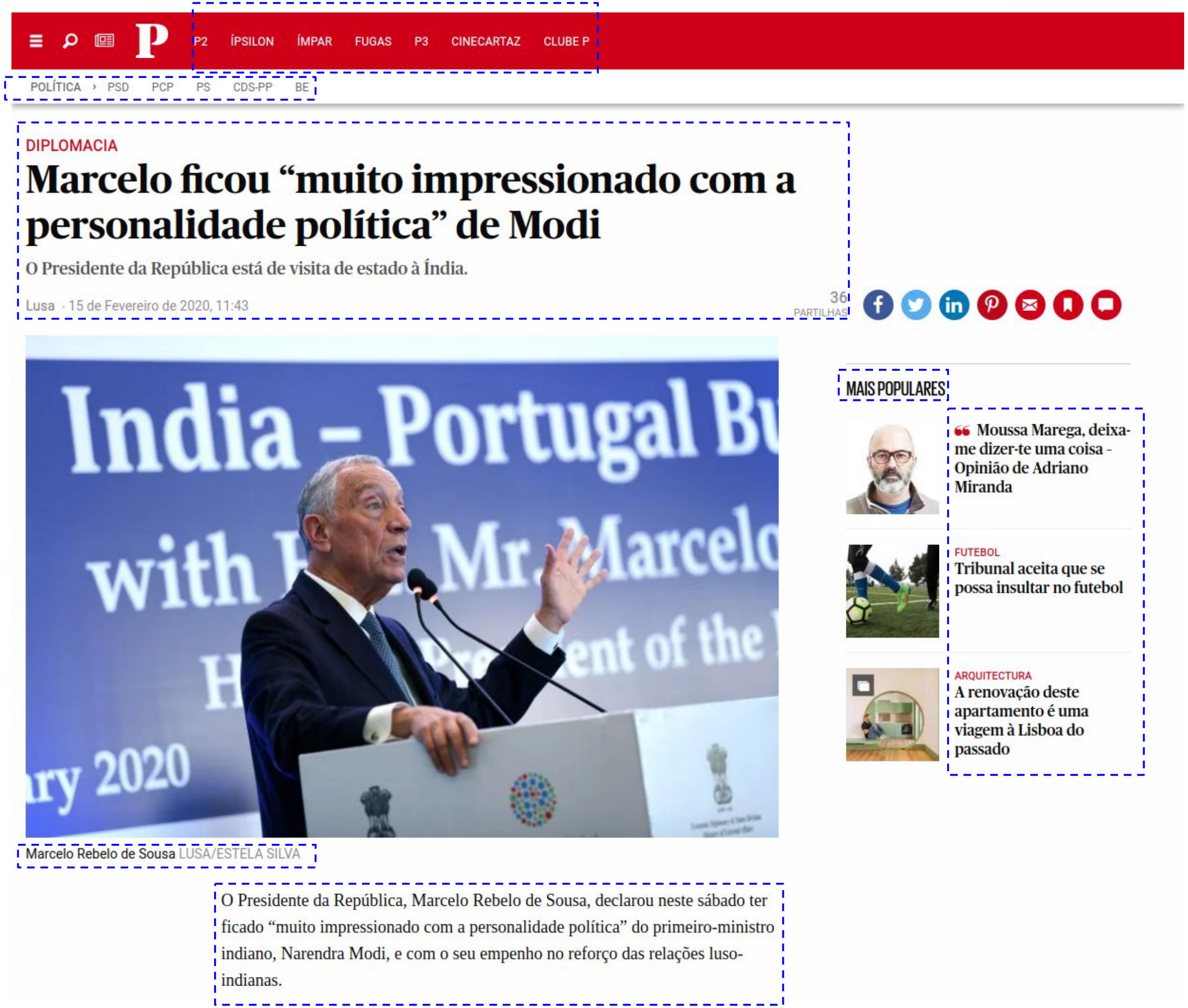
O Presidente da República, Marcelo Rebelo de Sousa, declarou neste sábado ter ficado “muito impressionado com a personalidade política” do primeiro-ministro indiano, Narendra Modi, e com o seu empenho no reforço das relações luso-indianas.

MAIS POPULARES

Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda

FUTEBOL
Tribunal aceita que se possa insultar no futebol

ARQUITECTURA
A renovação deste apartamento é uma viagem à Lisboa do passado



DIPLOMACIA

Marcelo ficou "muito impressionado com a personalidade política" de Modi

O Presidente da República está de visita de estado à Índia.

Lusa · 15 de Fevereiro de 2020, 11:43

Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

O Presidente da República, Marcelo Rebelo de Sousa, declarou neste sábado ter ficado "muito impressionado com a personalidade política" do primeiro-ministro indiano, Narendra Modi, e com o seu empenho no reforço das relações luso-indianas.

MAIS POPULARES

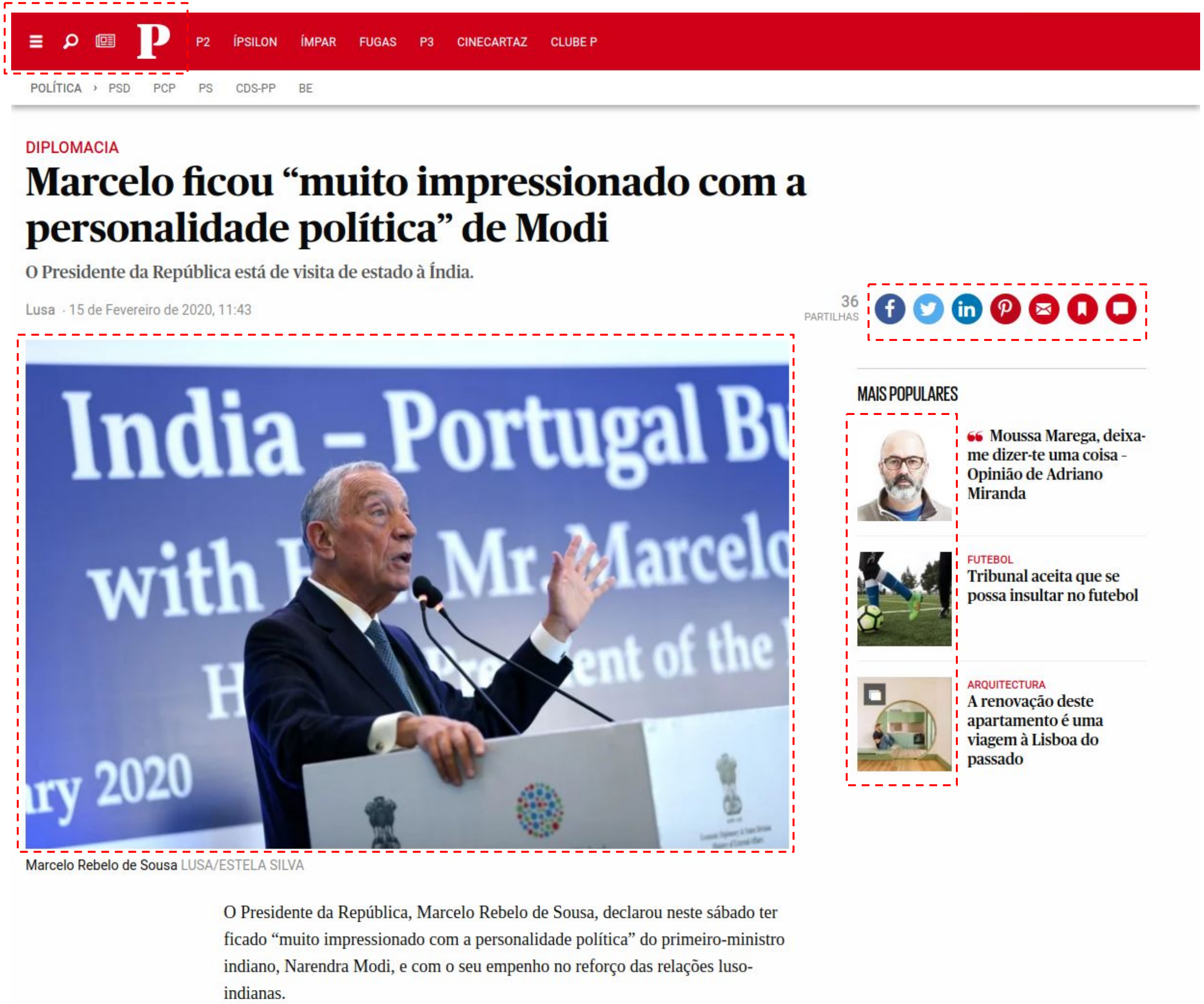
❝ Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda

FUTEBOL

Tribunal aceita que se possa insultar no futebol

ARQUITECTURA

A renovação deste apartamento é uma viagem à Lisboa do passado



P2 ÍPSILON ÍMPAR FUGAS P3 CINECARTAZ CLUBE P
POLÍTICA › PSD PCP PS CDS-PP BE

DIPLOMACIA

Marcelo ficou “muito impressionado com a personalidade política” de Modi

O Presidente da República está de visita de estado à Índia.

Lusa · 15 de Fevereiro de 2020, 11:43

Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

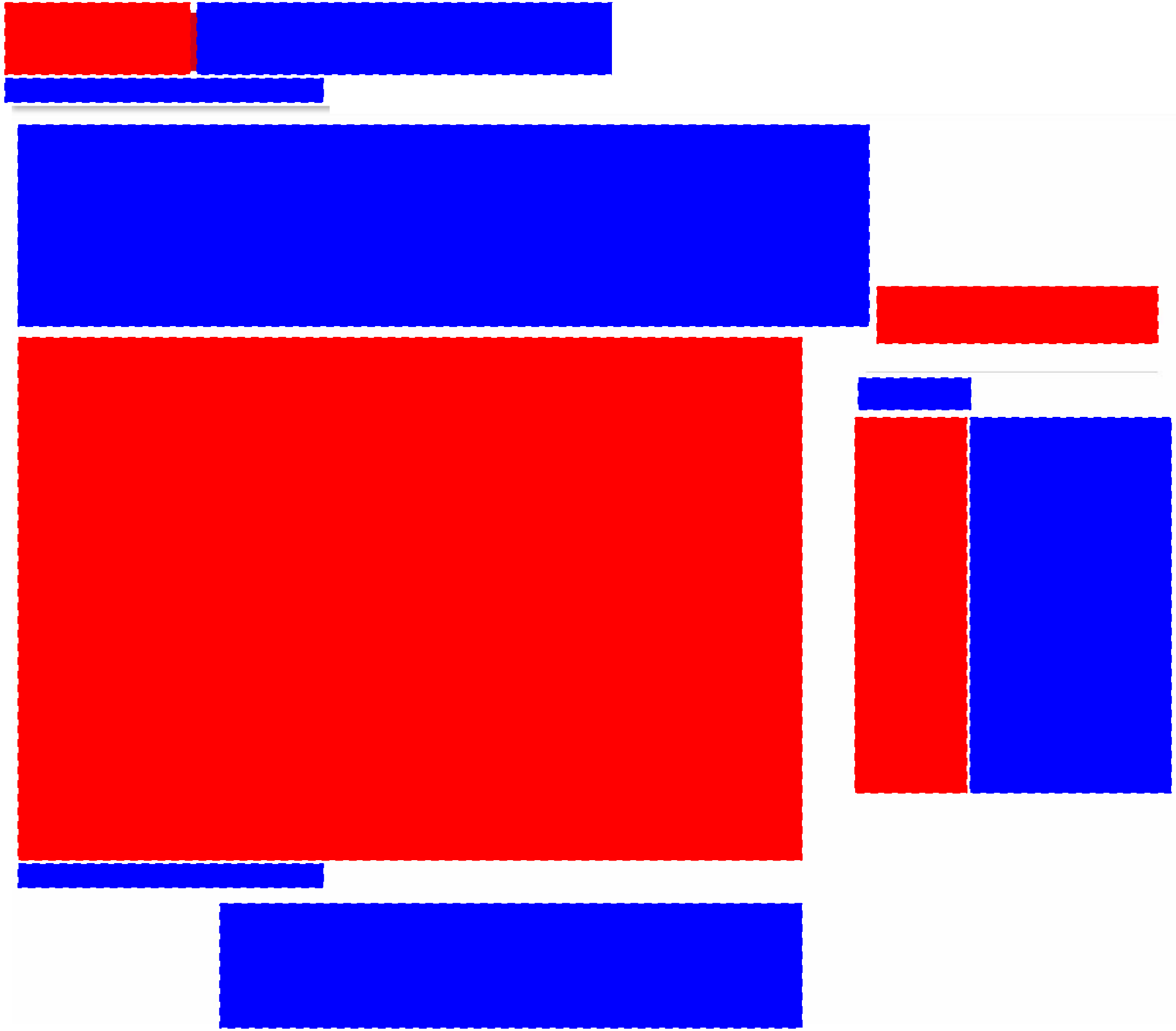
O Presidente da República, Marcelo Rebelo de Sousa, declarou neste sábado ter ficado “muito impressionado com a personalidade política” do primeiro-ministro indiano, Narendra Modi, e com o seu empenho no reforço das relações luso-indianas.

MAIS POPULARES

Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda

FUTEBOL
Tribunal aceita que se possa insultar no futebol

ARQUITECTURA
A renovação deste apartamento é uma viagem à Lisboa do passado



[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

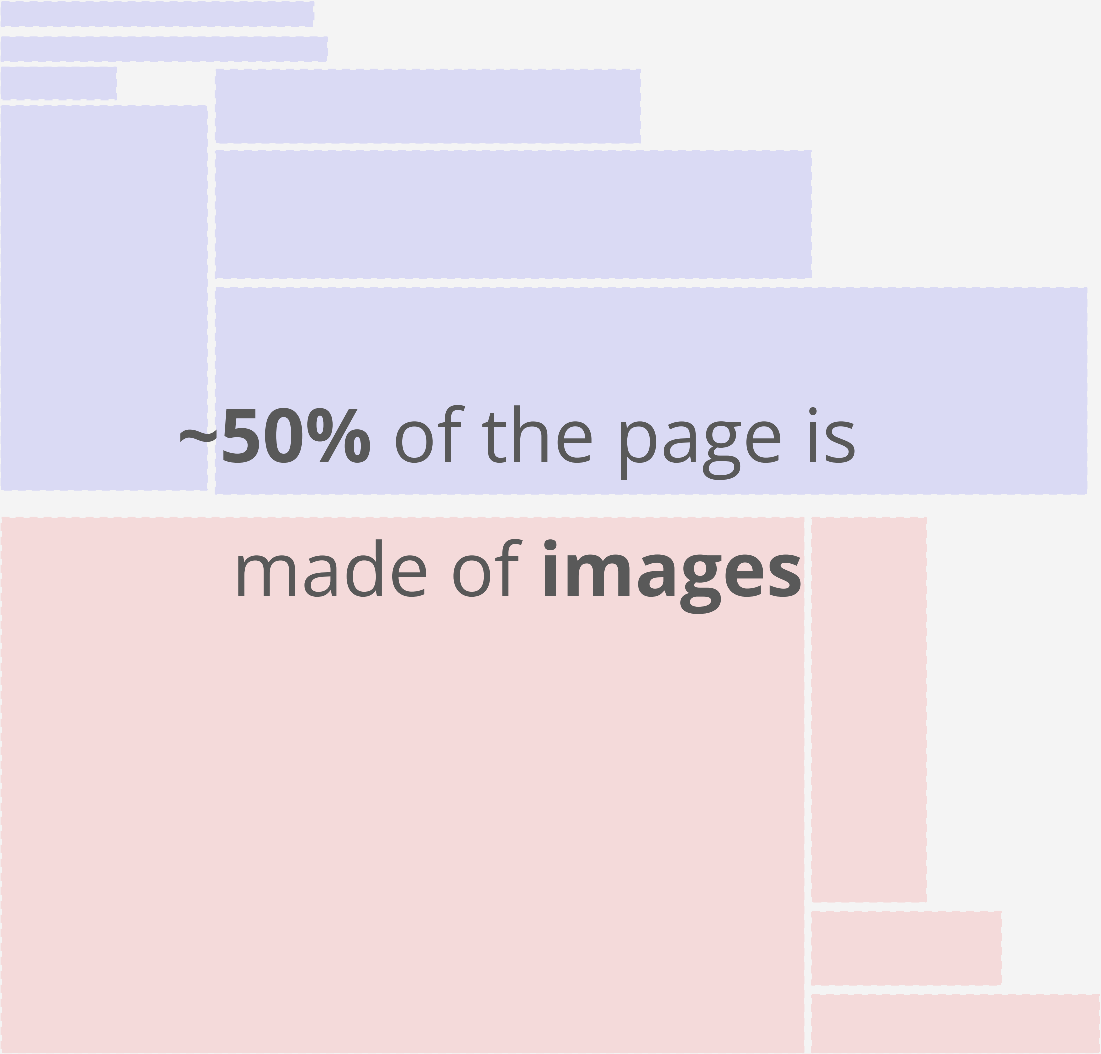
[redacted]

[redacted]

[redacted]

[redacted]

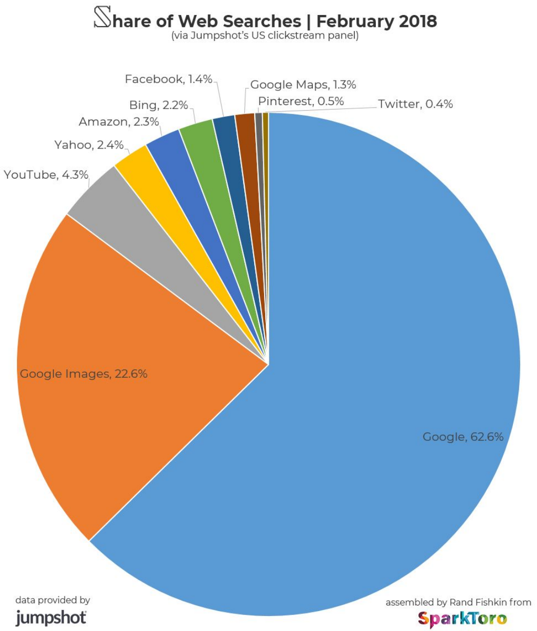
[redacted]

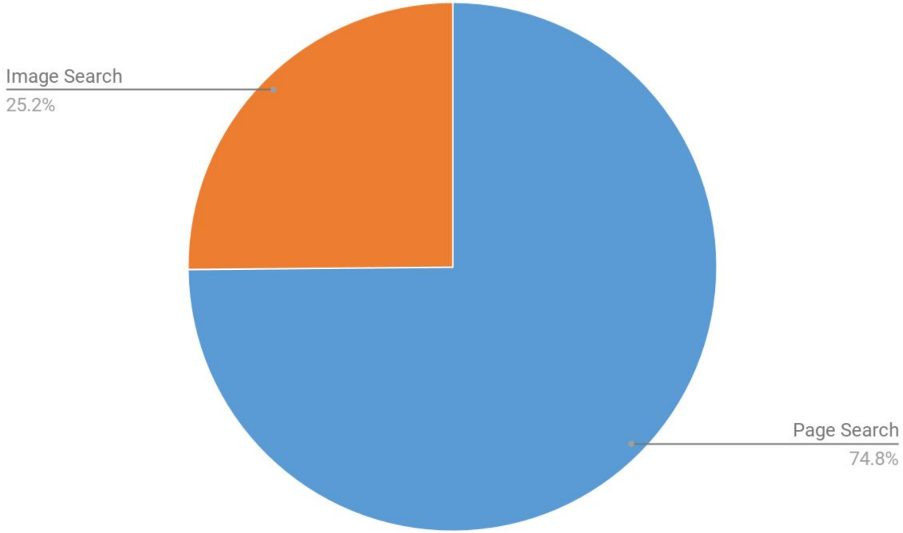


~50% of the page is

made of **images**

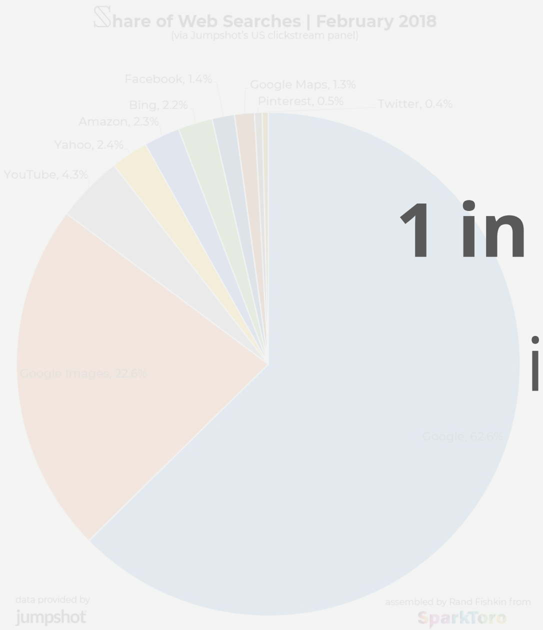
Why does image search matter?

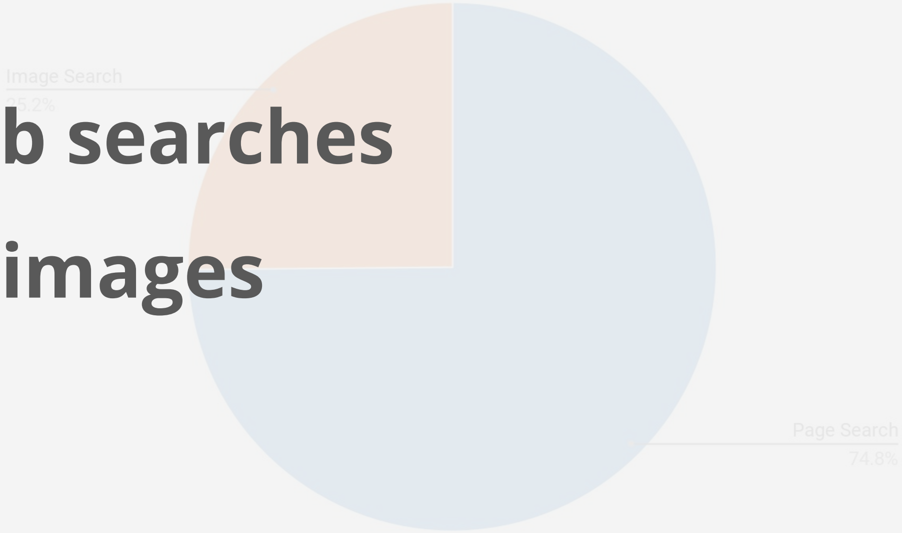




sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Why does image search matter?





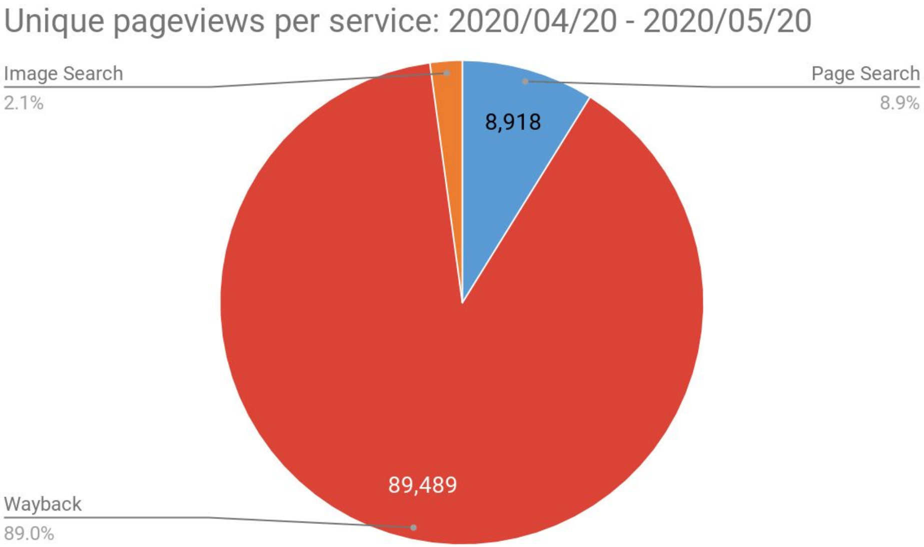
1 in 4 web searches

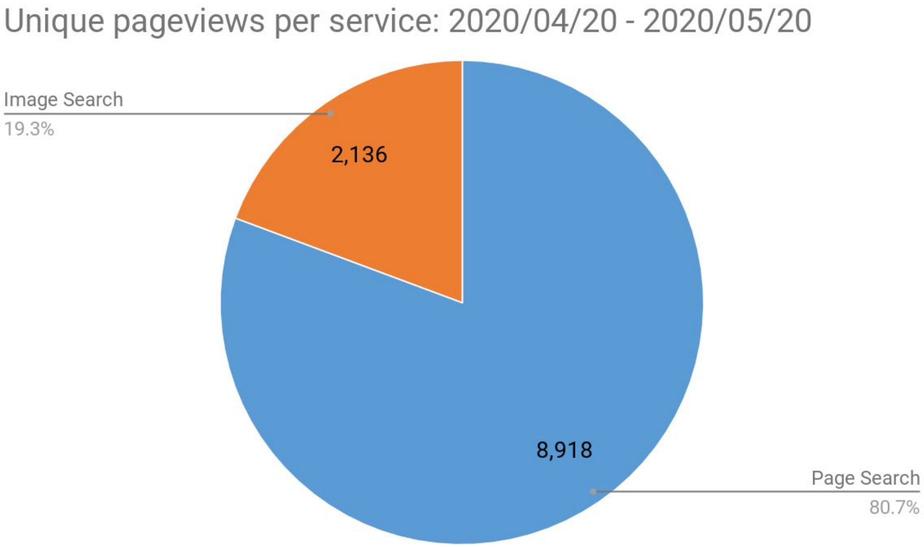
is for **images**

1 in 4 web searches is for images

sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

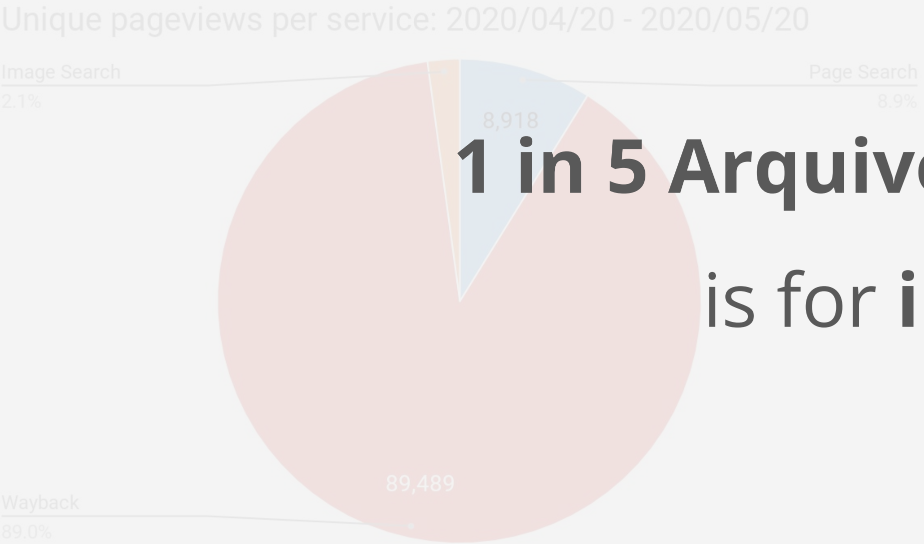
What about Arquivo.pt?

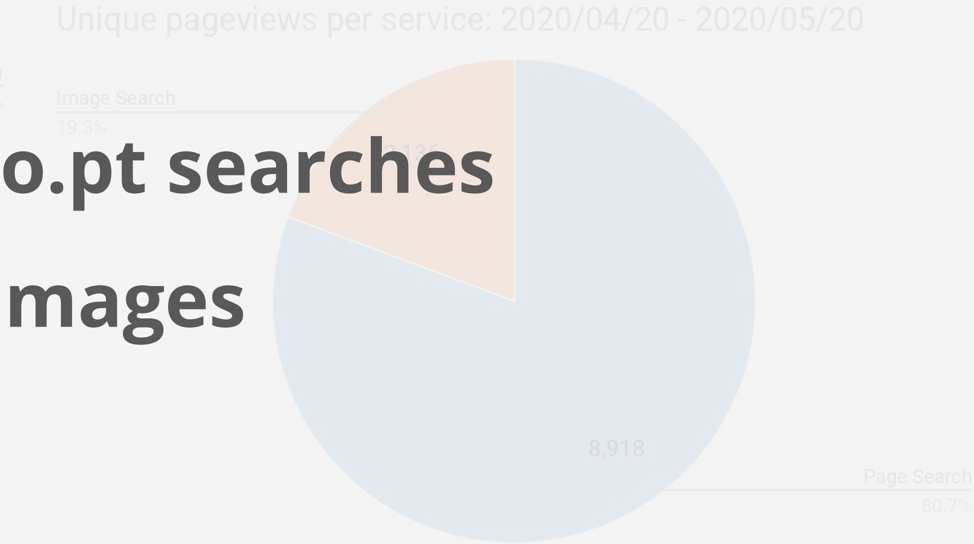




Arquivo.pt Google Analytics

What about Arquivo.pt?


Unique pageviews per service: 2020/04/20 - 2020/05/20

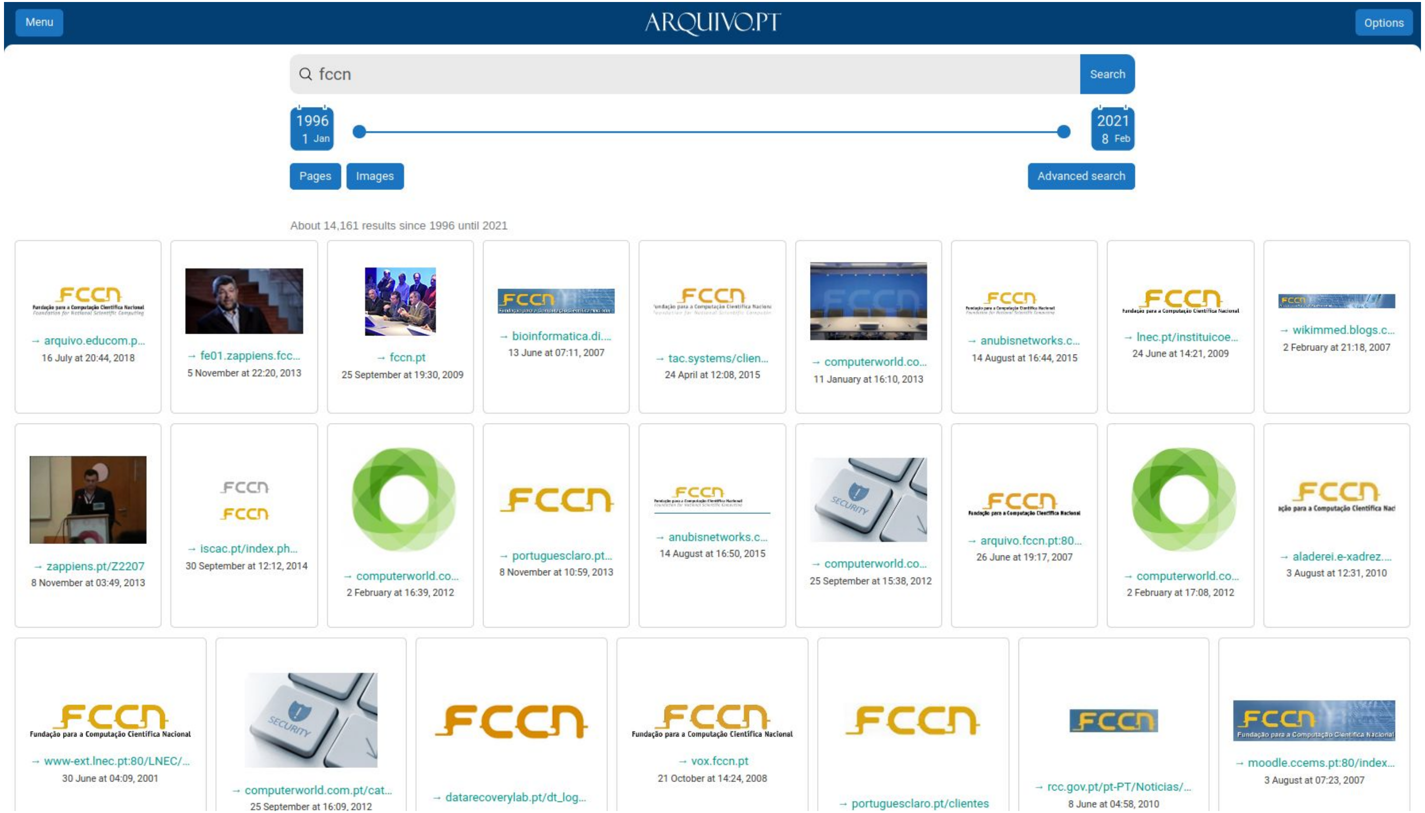

Unique pageviews per service: 2020/04/20 - 2020/05/20

1 in 5 Arquivo.pt searches

is for **images**

Arquivo.pt Google Analytics

Arquivo.pt Image Search



fccn

About 14,161 results since 1996 until 2021

→ arquivo.educom.p...
16 July at 20:44, 2018

→ fe01.zappiens.fcc...
5 November at 22:20, 2013

→ fccn.pt
25 September at 19:30, 2009

→ bioinformatica.di....
13 June at 07:11, 2007

→ tac.systems/clien...
24 April at 12:08, 2015

→ computerworld.co...
11 January at 16:10, 2013

→ anubisnetworks.c...
14 August at 16:44, 2015

→ lnec.pt/instituicoe...
24 June at 14:21, 2009

→ wikimmed.blogs.c...
2 February at 21:18, 2007

→ zappiens.pt/Z2207
8 November at 03:49, 2013

→ iscac.pt/index.ph...
30 September at 12:12, 2014

→ computerworld.co...
2 February at 16:39, 2012

→ portuguesclaro.pt...
8 November at 10:59, 2013

→ anubisnetworks.c...
14 August at 16:50, 2015

→ computerworld.co...
25 September at 15:38, 2012

→ arquivo.fccn.pt:80...
26 June at 19:17, 2007

→ computerworld.co...
2 February at 17:08, 2012

→ aladerei.e-xadrez....
3 August at 12:31, 2010

→ www-ext.lnec.pt:80/LNEC/...
30 June at 04:09, 2001

→ computerworld.com.pt/cat...
25 September at 16:09, 2012

→ datarecoverylab.pt/dt_log...

→ vox.fccn.pt
21 October at 14:24, 2008

→ portuguesclaro.pt/clientes

→ rcc.gov.pt/pt-PT/Noticias/...
8 June at 04:58, 2010

→ moodle.ccems.pt:80/index...
3 August at 07:23, 2007

Arquivo.pt APIs

- Arquivo.pt makes **8,000+ million pages** and **1,800+ million images** available for visualization and search:
  - Archived web pages -> **Text Search API**/Memento/CDX Server
  - Text and metadata search -> **Text Search API**
  - Image search -> **Image Search API**
- Available to the general public without registration
- Open Source
- https://github.com/arquivo/pwa-technologies/wiki/APIs

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	22 million
Collection count	90
(W)ARCs	3 million
(W)ARC sizes	334 TB
Total collected files	6,000 million
Total collected images	1,602 million
Oldest image date	15/04/1994
Newest image date	14/11/2019

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22 million**
Collection count	90
(W)ARCs	3 million
(W)ARC sizes	334 TB
Total collected files	6,000 million
Total collected images	**1,602 million**
Oldest image date	15/04/1994
Newest image date	14/11/2019

Arquivo.pt Image Search (as of March 2021)

Indexed images	**1,862 million**
Unique images	**584 million**
Collection count	115
(W)ARCs	5 million
(W)ARC sizes	520 TB
Total collected files	8,500 million
Total collected images	2,408 million
Oldest image date	15/04/1994
Newest image date	14/11/2020

Opportunities for improvement

- Lack of image specific metadata
  - 43% (10,163,080 images) without imgAlt or imgTitle
- Why is the difference between collected and indexed so large?
- Only the oldest page per image is indexed
- Search result ranking does not take image popularity into account

Potential solutions

- Index all* pages that mention an image
- Solve relative URL issues and find images in more places on page
- Remove MongoDB and use only Hadoop/HDFS

- Extract image caption from text surrounding images

- Use correct Solr types
- Extract metadata and use it for ranking
  - Number of times an image appear on page; number of times its metadata changes

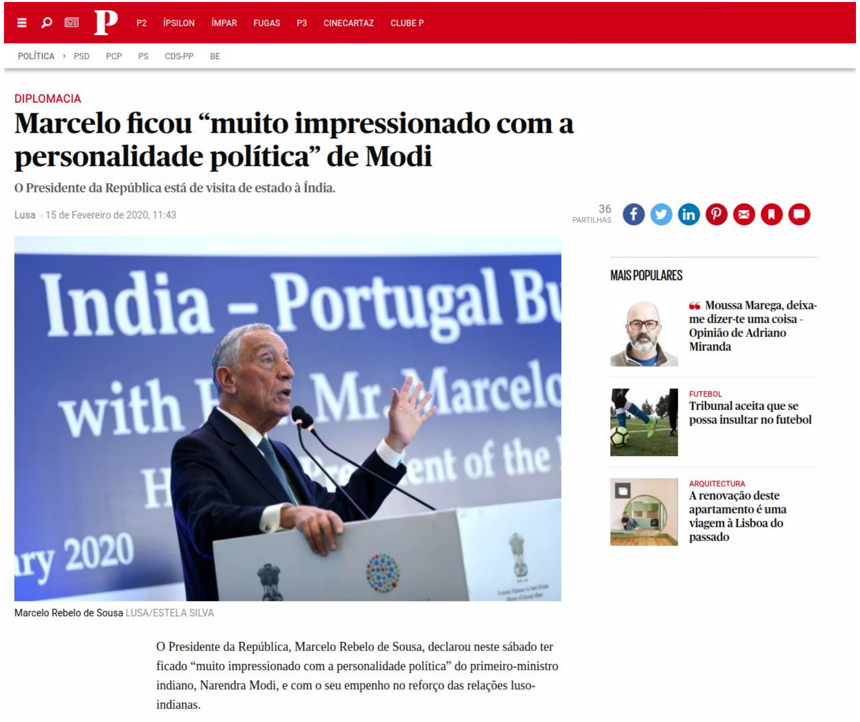
From images to metadata

- Image search is only as good as the associated metadata
- If we only look into the (W)ARC image records, we only have information about the image URL and image date, which is not very informative
- Where can we find this information?

From images to metadata

- Image search is only as good as the associated metadata
- If we only look into the (W)ARC image records, we only have information about the image URL and image date, which is not very informative
- Where can we find this information?
- HTML PAGES!

The anatomy of a webpage



DIPLOMACIA

Marcelo ficou “muito impressionado com a personalidade política” de Modi

O Presidente da República está de visita de estado à Índia.

Lusa · 15 de Fevereiro de 2020, 11:43

36 PARTILHAS

Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

O Presidente da República, Marcelo Rebelo de Sousa, declarou neste sábado ter ficado “muito impressionado com a personalidade política” do primeiro-ministro indiano, Narendra Modi, e com o seu empenho no reforço das relações luso-indianas.

MAIS POPULARES

Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda

FUTEBOL

Tribunal aceita que se possa insultar no futebol

ARQUITECTURA

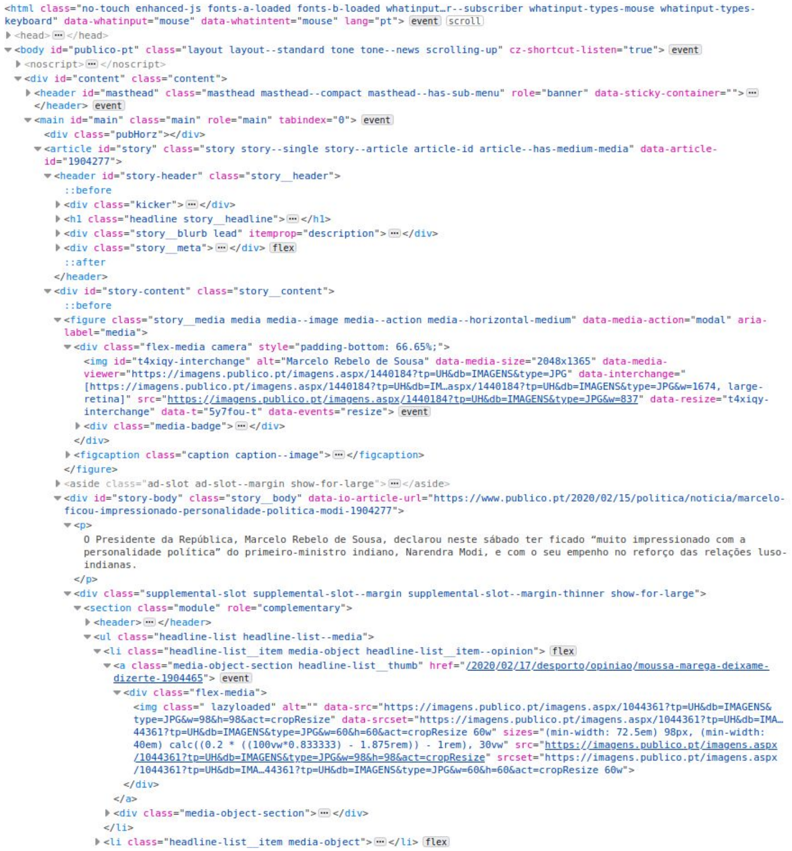
A renovação deste apartamento é uma viagem à Lisboa do passado

News regarding Marcelo’s visit to India

- Main image:
  - Marcelo Rebelo de Sousa, President of Portugal giving a speech as a part of the state visit to India
  - Caption: “Marcelo Rebelo de Sousa LUSA/ESTELA SILVA”
- Secondary images:
  - Person/Author of opinion piece
  - Soccer practice “stock photo”
  - Renovation of an apartment
  - Social network share icons
  - Público’s Logo
  - Other navigational buttons
- Links to external images
- Images as a CSS background

publico.pt/2020/02/15/politica/noticia/marcelo-ficou-impressionado-personalidade-politica-modi-1904277

Inside the skeleton



```
<html class="no-touch enhanced-js fonts-a-loaded fonts-b-loaded whatinput…r--subscriber whatinput-types-mouse whatinput-types-
keyboard" data-whatinput="mouse" data-whatintent="mouse" lang="pt"> event scroll
 <head>…</head>
 <body id="publico-pt" class="layout layout--standard tone tone--news scrolling-up" cz-shortcut-listen="true"> event
  <noscript>…</noscript>
  <div id="content" class="content">
   <header id="masthead" class="masthead masthead--compact masthead--has-sub-menu" role="banner" data-sticky-container="">…
   </header> event
   <main id="main" class="main" role="main" tabindex="0"> event
    <div class="pubHorz"></div>
    <article id="story" class="story story--single story--article article-id article--has-medium-media" data-article-
    id="1904277">
     <header id="story-header" class="story__header">
       ::before
      <div class="kicker">…</div>
      <h1 class="headline story__headline">…</h1>
      <div class="story__blurb lead" itemprop="description">…</div>
      <div class="story__meta">…</div> flex
       ::after
     </header>
     <div id="story-content" class="story__content">
       ::before
      <figure class="story__media media media--image media--action media--horizontal-medium" data-media-action="modal" aria-
      label="media">
       <div class="flex-media camera" style="padding-bottom: 66.65%;">
         <img id="t4xiqy-interchange" alt="Marcelo Rebelo de Sousa" data-media-size="2048x1365" data-media-
         viewer="https://imagens.publico.pt/imagens.aspx/1440184?tp=UH&db=IMAGENS&type=JPG" data-interchange="
         [https://imagens.publico.pt/imagens.aspx/1440184?tp=UH&db=IM…aspx/1440184?tp=UH&db=IMAGENS&type=JPG&w=1674, large-
         retina]" src="https://imagens.publico.pt/imagens.aspx/1440184?tp=UH&db=IMAGENS&type=JPG&w=837" data-resize="t4xiqy-
         interchange" data-t="5y7fou-t" data-events="resize"> event
        <div class="media-badge">…</div>
       </div>
       <figcaption class="caption caption--image">…</figcaption>
      </figure>
      <aside class="ad-slot ad-slot--margin show-for-large">…</aside>
      <div id="story-body" class="story__body" data-io-article-url="https://www.publico.pt/2020/02/15/politica/noticia/marcelo-
      ficou-impressionado-personalidade-politica-modi-1904277">
       <p>
         O Presidente da República, Marcelo Rebelo de Sousa, declarou neste sábado ter ficado “muito impressionado com a
         personalidade política” do primeiro-ministro indiano, Narendra Modi, e com o seu empenho no reforço das relações luso-
         indianas.
       </p>
       <div class="supplemental-slot supplemental-slot--margin supplemental-slot--margin-thinner show-for-large">
        <section class="module" role="complementary">
         <header>…</header>
         <ul class="headline-list headline-list--media">
          <li class="headline-list__item media-object headline-list__item--opinion"> flex
           <a class="media-object-section headline-list__thumb" href="/2020/02/17/desporto/opiniao/moussa-marega-deixame-
           dizerte-1904465"> event
            <div class="flex-media">
              <img class=" lazyloaded" alt="" data-src="https://imagens.publico.pt/imagens.aspx/1044361?tp=UH&db=IMAGENS&
              type=JPG&w=98&h=98&act=cropResize" data-srcset="https://imagens.publico.pt/imagens.aspx/1044361?tp=UH&db=IMA…
              44361?tp=UH&db=IMAGENS&type=JPG&w=60&h=60&act=cropResize 60w" sizes="(min-width: 72.5em) 98px, (min-width:
              40em) calc((0.2 * ((100vw*0.833333) - 1.875rem)) - 1rem), 30vw" src="https://imagens.publico.pt/imagens.aspx
              /1044361?tp=UH&db=IMAGENS&type=JPG&w=98&h=98&act=cropResize" srcset="https://imagens.publico.pt/imagens.aspx
              /1044361?tp=UH&db=IMA…44361?tp=UH&db=IMAGENS&type=JPG&w=60&h=60&act=cropResize 60w">
            </div>
           </a>
           <div class="media-object-section">…</div>
          </li>
          <li class="headline-list__item media-object">…</li> flex
```

Finding images in pages

- <img> tag attributes
- <a> tag attributes
- Inline CSS background images
- Inline base64 images
- Images set by JS
- <figure>, <picture>

Takeways for finding images in pages

- <img> tag attributes
  - everything in *src* (regardless of image extension)
  - **other attributes that match list of image extensions**
- **<a> tag attributes**
  - *href* that match list of image extensions
- **Inline CSS background images**
  - *background-url:* that match list of image extensions
- **Base64 images**
- **Fixed relative URL solver**

Finding images in pages results

- <img> tag attributes
- <a> tag attributes
- Inline CSS background images
- Inline base64 images
- Images set by JS
- <figure>, <picture>

	Percentage of references
<img>	90.6%
<a>	8.7%
css	0.7%

	Percentage of references
Normal images	99.9%
base64	0.1%

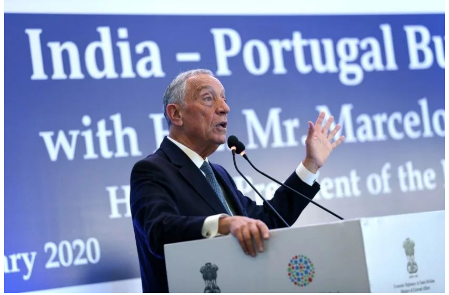
From page metadata to image metadata

The following attributes are common to all images that show up in a page:

- Page Title
  - Page title attribute; it is used to provide additional information about an HTML page;
- Page URL Tokens
  - The keywords of the URL of the HTML page that contains the image.

But this general information may not be relevant to all images

Matching images to HTML `<img>` tags



```
pageTitle="Marcelo ficou “muito
impressionado com a personalidade política”
de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02
15 politica noticia marcelo ficou (...)"
```

```
pageTitle="Marcelo ficou “muito
impressionado com a personalidade política”
de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02
15 politica noticia marcelo ficou (...)"
```

```
pageTitle="Marcelo ficou “muito
impressionado com a personalidade política”
de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02
15 politica noticia marcelo ficou (...)"
```

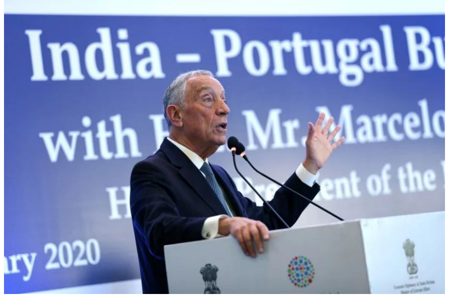
Metadata: <img> tag attributes

We select all <img> tags in the html and extract the following metadata:

- imgSrcTokens
  - an image by a URL, which often includes the filename of the image
- imgTitle
  - it provides additional information about the image;
- imgAlt
  - it provides alternative information about an image if a user cannot view it;

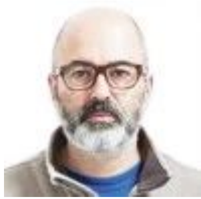
Matching images to HTML `<img>` tags

ARQUIVO.PT



pageTitle="Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02 15 politica noticia marcelo ficou (...)"

imgAlt="Marcelo Rebelo de Sousa"
imgTitle=""
imgSrcTokens="imagens publico pt imagens aspx 1440184"



pageTitle="Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02 15 politica noticia marcelo ficou (...)"

imgAlt=""
imgTitle=""
imgSrcTokens="imagens publico pt imagens aspx 1044361"

pageTitle="Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO"
pageSrcTokens="https www publico pt 2020 02 15 politica noticia marcelo ficou (...)"

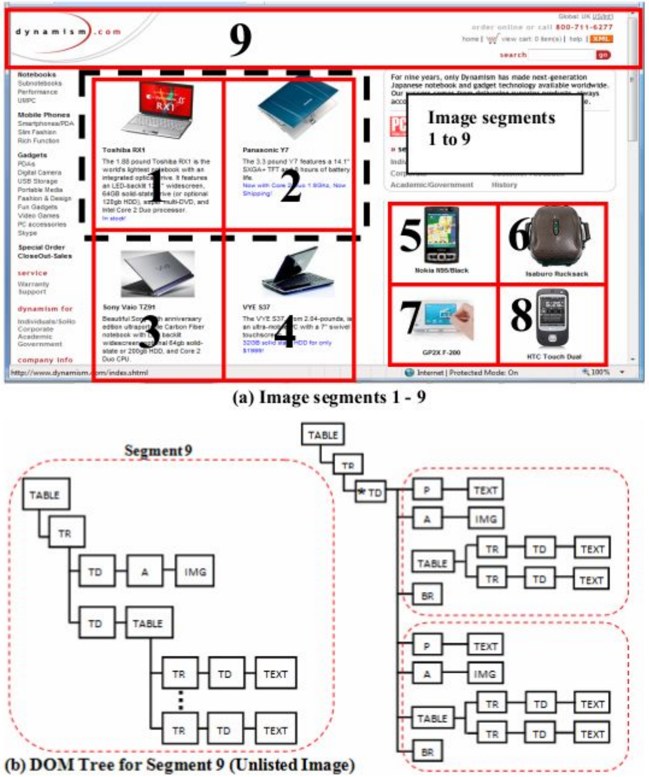
imgAlt=""
imgTitle=""
imgSrcTokens="imagens publico pt imagens aspx 735549">

Metadata: <img> tag attributes

We select all <img> tags in the html and extract the following metadata:

- imgSrcTokens
  - an image by a URL, which often includes the filename of the image
- imgTitle
  - it provides additional information about the image;
- imgAlt
  - it provides alternative information about an image if a user cannot view it;
- **imgCaption**
  - **portion of the HTML page text that is closest to the image**

Finding an image caption

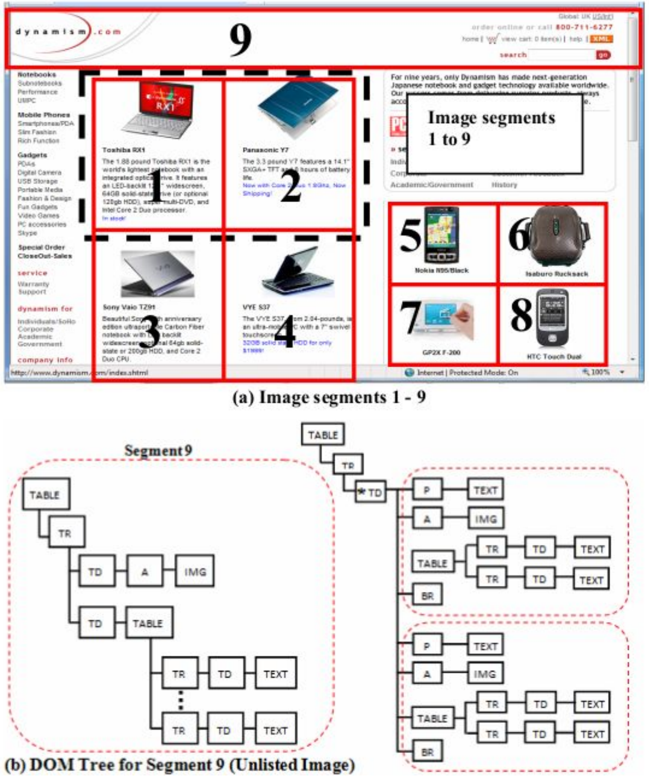


(a) Image segments 1 - 9

(b) DOM Tree for Segment 9 (Unlisted Image)

Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information. 649-652. 10.1145/1631272.1631379.

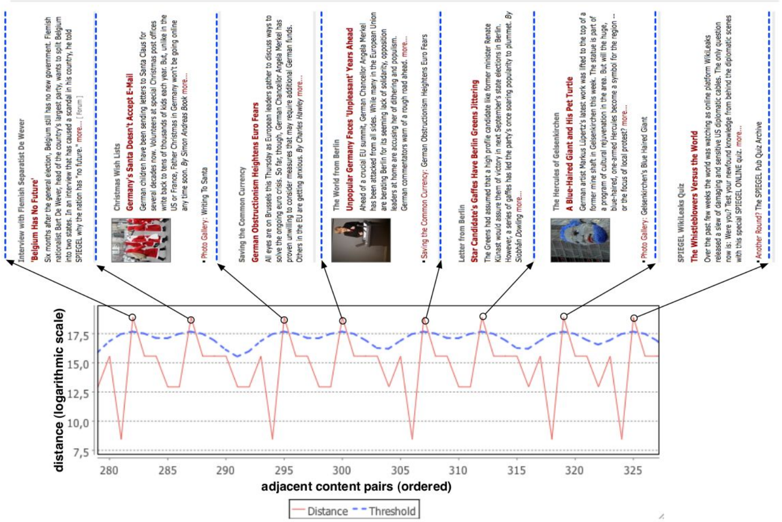
Finding an image caption



(a) Image segments 1 - 9

(b) DOM Tree for Segment 9 (Unlisted Image)

Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information. 649-652. 10.1145/1631272.1631379.



Sadet, Alcic & Conrad, Stefan. (2011). A Clustering-based Approach to Web Image Context Extraction. MMEDIA - International Conferences on Advances in Multimedia.

Image caption extraction

I arrived at the following method

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages

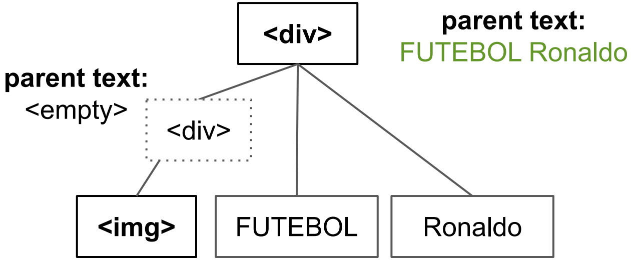
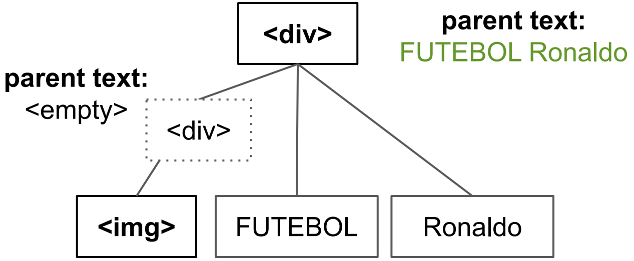


Image caption extraction

I arrived at the following method

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages



Previous and next node text

- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog

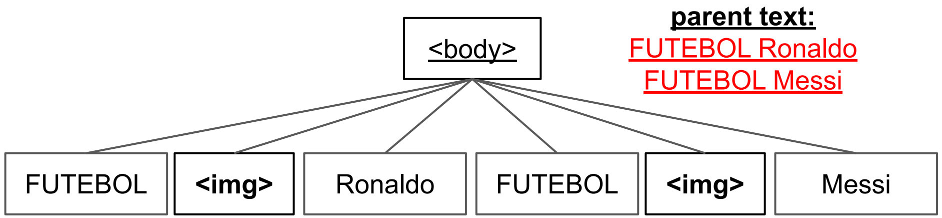
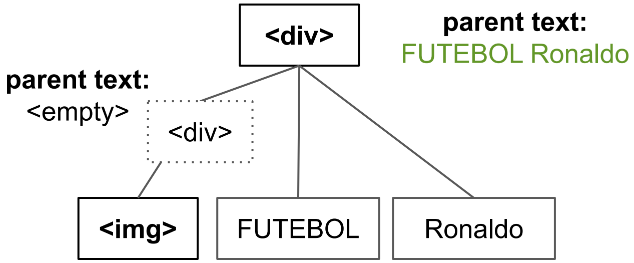


Image caption extraction

I arrived at the following method

First parent with text

- Default method
- Works well for images in boxes or *reasonably* structured pages

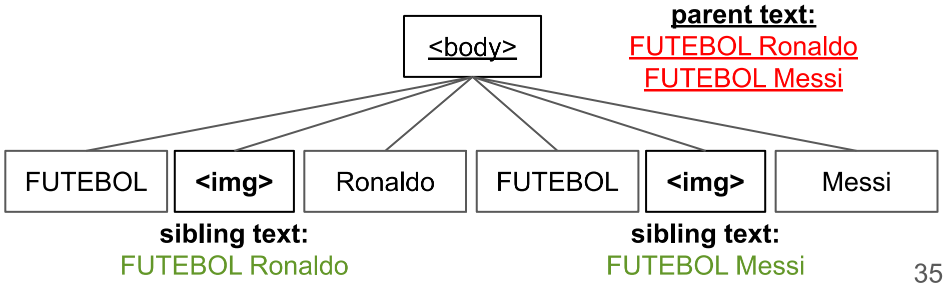


parent text:
FUTEBOL Ronaldo

parent text:
<empty>

Previous and next node text

- Used if the first parent with text is at the level of the page with more siblings
- List of images as in a blog

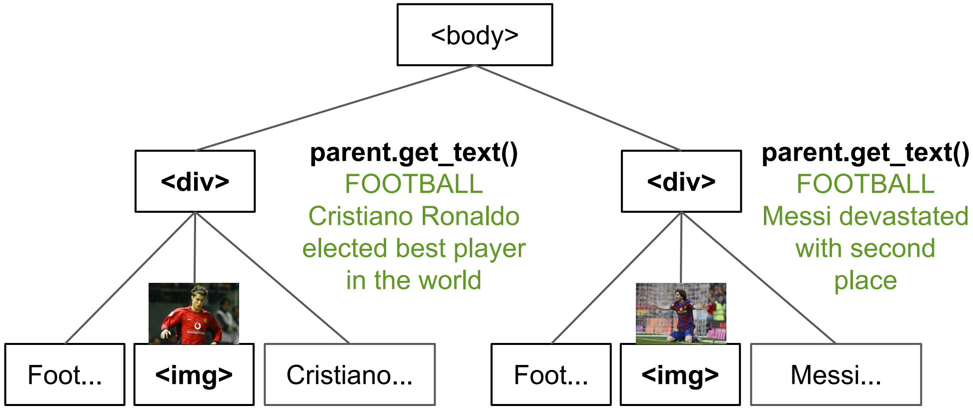


parent text:
FUTEBOL Ronaldo
FUTEBOL Messi

sibling text:
FUTEBOL Ronaldo

sibling text:
FUTEBOL Messi

Associar palavras do elemento *parent* do HTML à imagem (legendas)



<body>

<div>

parent.get_text()
FOOTBALL
Cristiano Ronaldo
elected best player
in the world

Foot...

<img>

Cristiano...

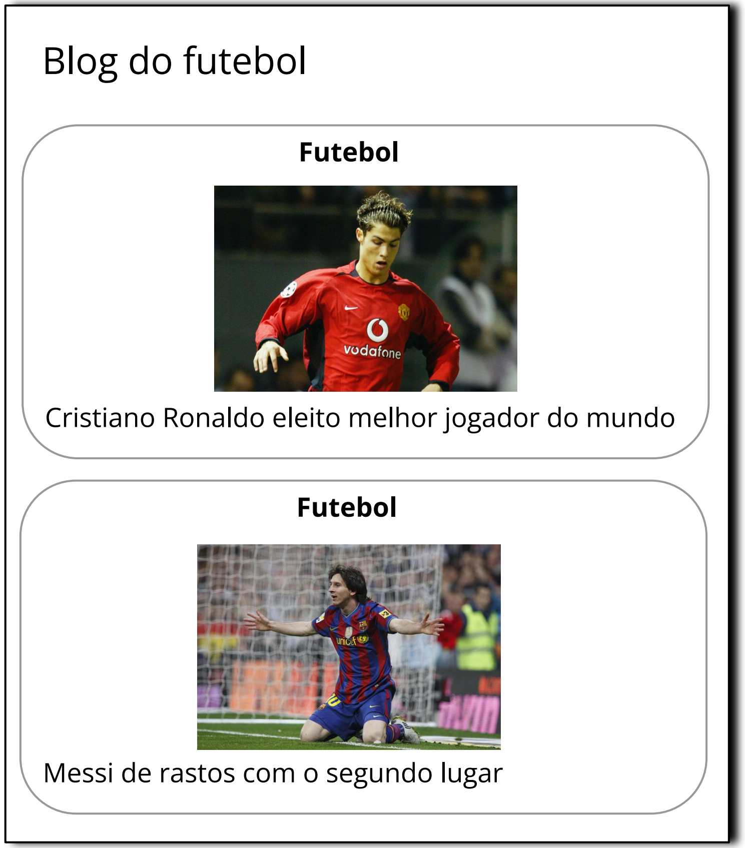
<div>

parent.get_text()
FOOTBALL
Messi devastated
with second
place

Foot...

<img>

Messi...



Blog do futebol

Futebol

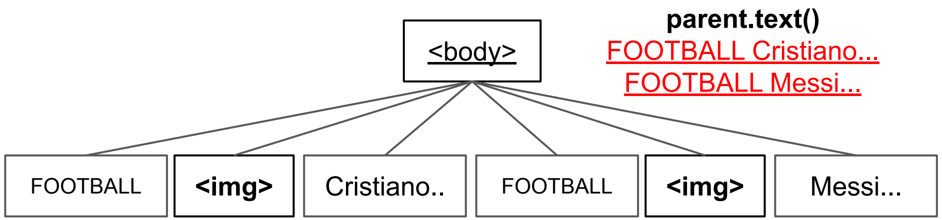
Cristiano Ronaldo eleito melhor jogador do mundo

Futebol

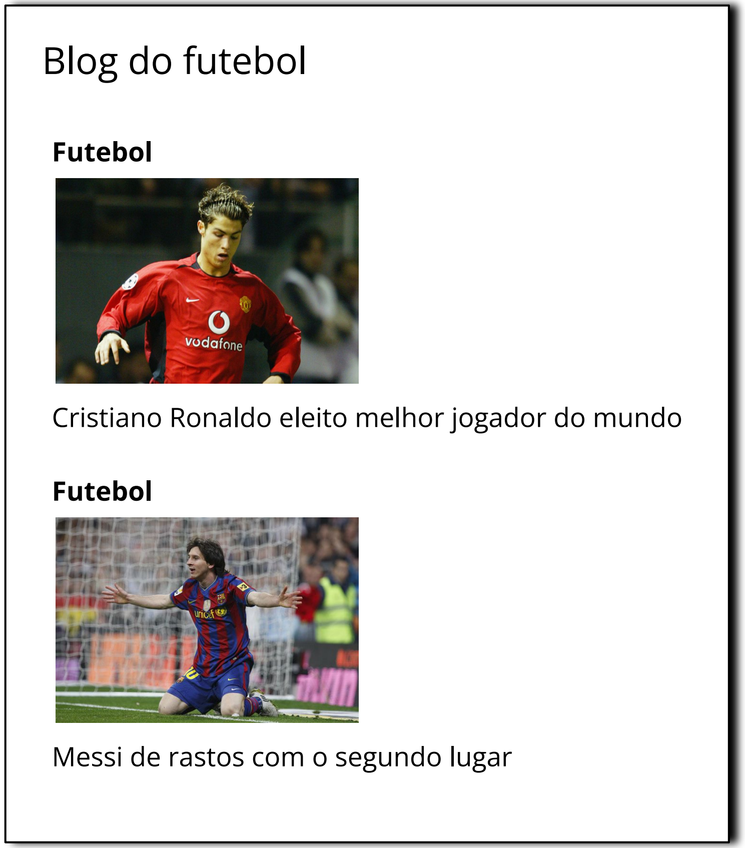
Messi de rastos com o segundo lugar

Texto do *parent* **funciona** em páginas com HTML correctamente estruturado

Hipótese falha em páginas com estrutura *“flat”*



Texto do elemento *parent* **falha** em páginas mal estruturadas (sem separação entre tipos de conteúdo semântico)



Blog do futebol

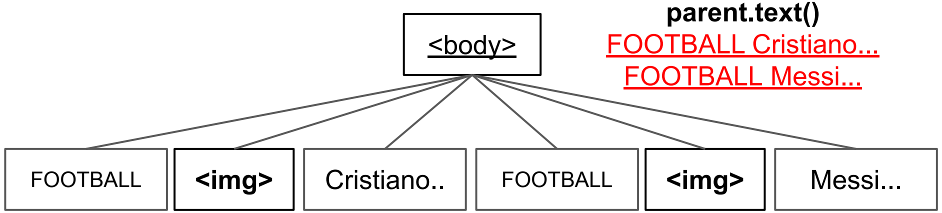
Futebol

Cristiano Ronaldo eleito melhor jogador do mundo

Futebol

Messi de rastos com o segundo lugar

Solução método híbrido: parent.text() OR sibling.text()

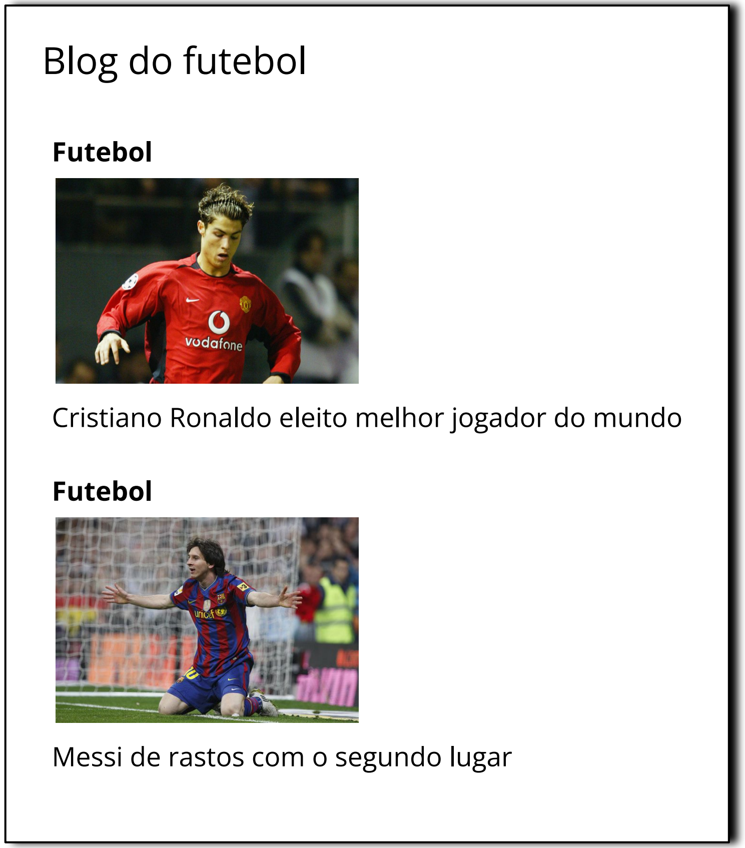


parent.text()
FOOTBALL Cristiano...
FOOTBALL Messi...

sibling.get_text()
FOOTBALL
Cristiano Ronaldo
elected best player
in the world

sibling.get_text()
FOOTBALL
Messi devastated
with second place

- **Páginas normais:** *text* do *parent*
- **Páginas com estrutura *flat*:** *text* dos nós adjacentes (*siblings*)



Blog do futebol

Futebol

Cristiano Ronaldo eleito melhor jogador do mundo

Futebol

Messi de rastos com o segundo lugar

Matching images to HTML `<img>` tags

pageTitle="Marcelo ficou “muito impressionado com a personalidade política” de Modi (...)"

pageSrcTokens="(...) marcelo ficou (...)"

imgAlt="Marcelo Rebelo de Sousa"

imgTitle=""

imgSrcTokens="image ns publico pt imagens aspx (...)"

imgCaption="Marcelo Rebelo de Sousa LUSA/ESTELA SILVA"

pageTitle="Marcelo ficou “muito impressionado (...)

pageSrcTokens="(...)"

imgAlt=""

imgTitle=""

imgSrcTokens="image ns publico pt imagens aspx (...)"

imgCaption="FUTEBOL Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda"

pageTitle="Marcelo ficou “muito impressionado (...)"

pageSrcTokens="(...)"

imgAlt=""

imgTitle=""

imgSrcTokens="image ns publico pt imagens aspx (...)"

imgCaption="ARQUITECTURA A renovação deste apartamento é uma viagem à Lisboa do passado"

Metadata: <a> tag attributes (new)

An alternative way to find images on the page is find direct links to images

To do this, we select links (<a>) that point to files to with image extensions, and extract the following metadata:

- **imgSrcTokens**
- **imgFilename**
- **imgCaption (<a> anchor text)**
  - **The text inside the link is used as the image caption**

Metadata: CSS image attributes (new)

An increasingly popular way of placing images on the web is through the use of the CSS background attribute, which places the image inside an HTML element (usually a div)

These images are referenced through a CSS background:url(‘<url>’)

- **imgSrcTokens**
- **imgFilename**

Image metadata takeways

- Extracted image caption for images found in <img>
- Used anchor text as image caption for images found in <a>
- Used only page metadata for *css* images

Metadata Open Questions

How many images have explicit metadata?

- Current measurements show over **99%** of the images have metadata (imgCaption)

But how can we measure the quality of this metadata?

- IMG_00123.jpg is not a very helpful entry

What to do when this metadata is missing (CSS or orphan image records) or wrong?

- Page title, top-k terms….
- Deep Image Captioning/Classification techniques

Examine the quality of the metadata

Encodings on the internet

“Lote para construÃ§Ã£o de moradia IncluÃ projecto a aprovar Ã�rea do lote --Â» 360 m2 Com frente de 20 metros”

“EspectÃ¡culos a nÃ£o perder 09/04/2009 | Sem ComentÃ¡rios | Concertos”

“Mobidogs Sempre sonhou ter um cÃ£o, um companheiro simpÃ¡tico que esteja ao seu lado para partilhar as suas alegrias e desgostos.... 4.00EUR”

```
public static String decode(byte[] arcRecordBytes) throws IOException {
    String recordEncoding = ImageSearchIndexingUtil .guessEncoding(arcRecordBytes);
    InputStream is = new ByteArrayInputStream(arcRecordBytes);
    String html = IOUtils.toString(is, recordEncoding);
    //if chars in UTF8_MISMATCH were detected, the page is in UTF_8 but encoded in ISO_8859_1
    //if we re-encode the string, the accented chars will be correctly represented
    if (ImageSearchIndexingUtil.UTF8_MISMATCH.matcher(html).find()){
        byte[] b = html.getBytes(StandardCharsets .ISO_8859_1);
        String newHtml = new String(b, StandardCharsets .UTF_8);
        //if the chars are detected again, the page is beyond repair and the initial encoding is used
        if (!ImageSearchIndexingUtil .UTF8_MISMATCH.matcher(newHtml).find()){
            html = newHtml;
        }
    }

    return html;
}
```

```java
public static String decode(byte[] arcRecordBytes) throws IOException {
   String recordEncoding = ImageSearchIndexingUtil .guessEncoding(arcRecordBytes);
   InputStream is = new ByteArrayInputStream(arcRecordBytes);
   String html = IOUtils.toString(is, recordEncoding);
   //if chars in UTF8_MISMATCH were detected, the page is in UTF_8 but encoded in ISO_8859_1
   //if we re-encode the string, the accented chars will be correctly represented
   if (ImageSearchIndexingUtil .UTF8_MISMATCH.matcher(html).find()){
       byte[] b = html.getBytes( StandardCharsets .ISO_8859_1);
       String newHtml = new String(b, StandardCharsets .UTF_8);
       //if the chars are detected again, the page is beyond repair and the initial encoding is used
       if (!ImageSearchIndexingUtil.UTF8_MISMATCH.matcher(newHtml).find()){
           html = newHtml;
       }
   }

   return html;
}
```

“Lote para construção de moradia Incluí projecto a aprovar área do lote 360 m2 Com frente de 20 metros”

“Espectáculos a não perder 09/04/2009 | Sem Comentários | Concertos”

“Mobidogs Sempre sonhou ter um cão, um companheiro simpático que esteja ao seu lado para partilhar as suas alegrias e desgostos.... 4.00EUR”

Pages with wrong encoding in AWP4

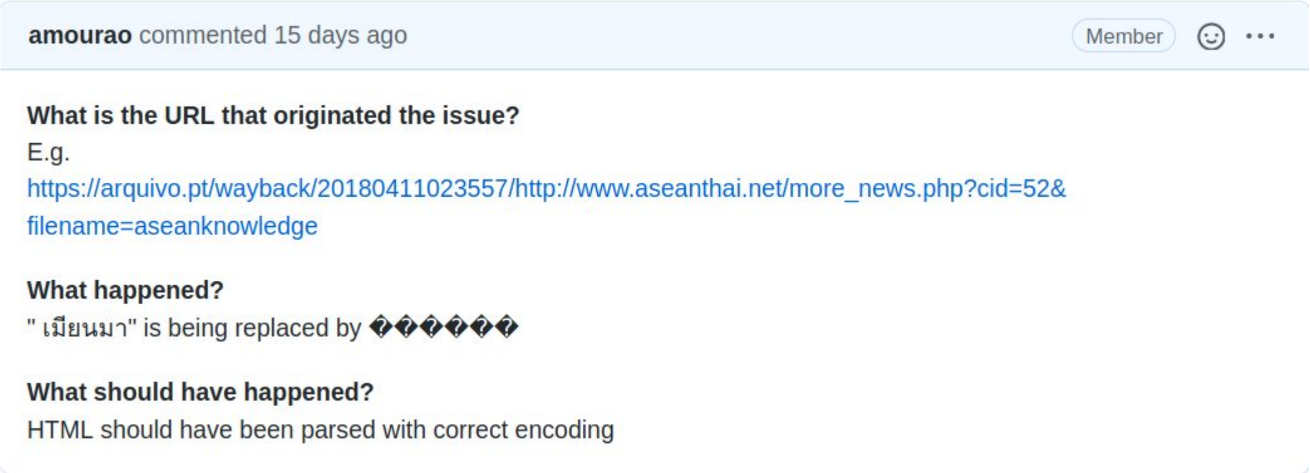
Detector Mozilla: 160524

Detector Tika: 366202

Detector Mozilla + meu fix: 9665

But this does not solve all encoding issues

- https://github.com/arquivo/pwa-technologies/issues/1059



amourao commented 15 days ago

What is the URL that originated the issue?
E.g.
https://arquivo.pt/wayback/20180411023557/http://www.aseanthai.net/more_news.php?cid=52&filename=aseanknowledge

What happened?
" เมียนมา" is being replaced by ������

What should have happened?
HTML should have been parsed with correct encoding

- Won’t fix, only affects 1.4% of all extracted results, mostly in “exotic” encodings

Indexing Architecture

(W)ARC examples

https://imagens.publico.pt/(...)

https://imagens.publico.pt/imagens.aspx/1440184

https://imagens.publico.pt/(...)

https://imagens.publico.pt/imagens.aspx/1044361

https://imagens.publico.pt/(...)

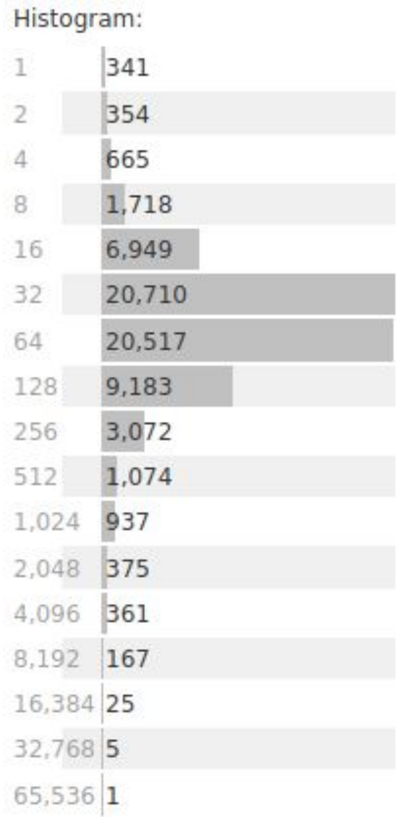
```
<img alt="Marcelo Rebelo de
Sousa"
id="t4xiqy-interchange" (…)
src="https://imagens.public
o.pt/imagens.aspx/1440184">

<img alt="" (…)
src="https://imagens.public
o.pt/imagens.aspx/1044361">
```

(W)ARCs, HTML pages and images

During crawling, pages and images are added to a queue

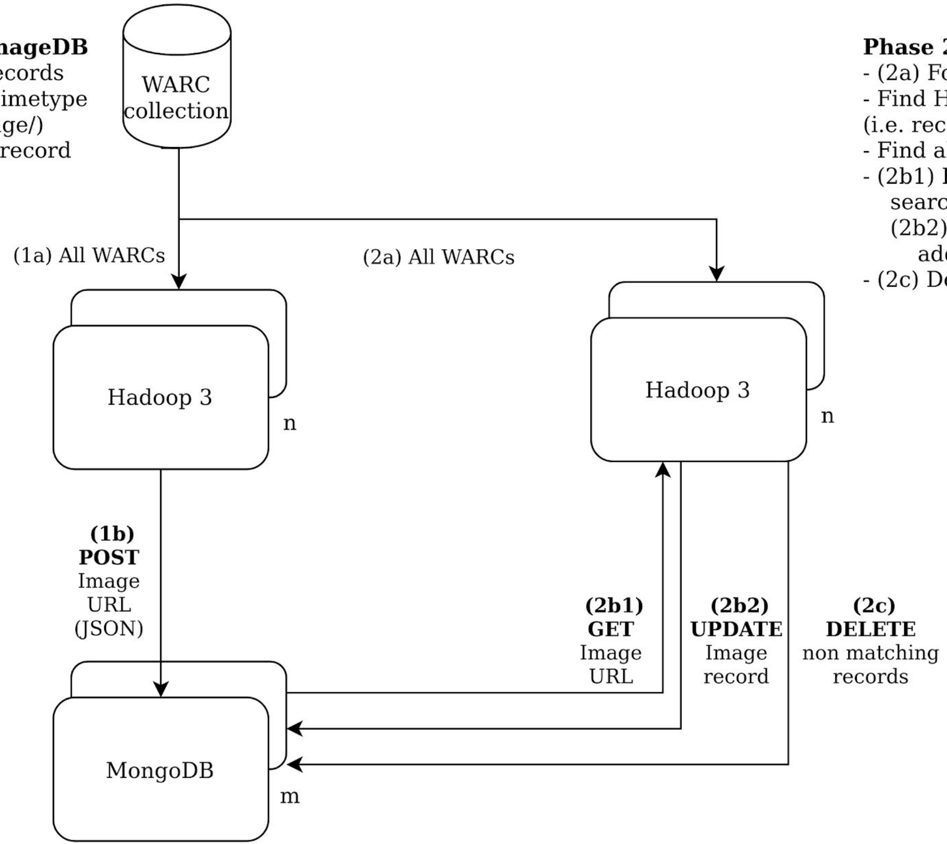
Best case scenario, HTML and images are crawled together

- HTML and Images may be in different (W)ARCs
- HTML may reference non-crawled images
- Multiple versions of the image and HTML may be crawled
- URL may not be consistent across HTML images and img src
- Images may be referenced on more than one page
  - Same URL
  - Different URL, Same digest
  - Different URL, Different digest, near duplicate (e.g. resized versions)
  - Images may also change for the same url (Same URL, Different digest)



Old Map Reduce

Phase 1: CreateImageDB
- (1a) Find image records
(i.e. records with mimetype
that starts with image/)
- (1b) Create JSON record



Phase 2: IndexImages
- (2a) For all WARCS
- Find HTML records
(i.e. records with mimetype that starts with text/html)
- Find all image tags in that html page
- (2b1) For each imgtag
    search image in DB
    (2b2) If image found
        add new index for that image to the outputs.
- (2c) Delete non matching records

Problems

- (W)ARCs are downloaded and parsed twice (images and HTML)
- Only the oldest page for each image is stored in the index
- MongoDB bottlenecks Hadoop parsing
- No logging is performed for what is happening
    - Failed (W)ARCs
    - Images parsed
    - ....

Solutions

- Parse HTML and image records in the same process
- Store all relevant pages for an image
- Rely on Hadoop/HDFS to match images to pages across WARCs
    - Use image URL as Hadoop key

Extract images with metadata

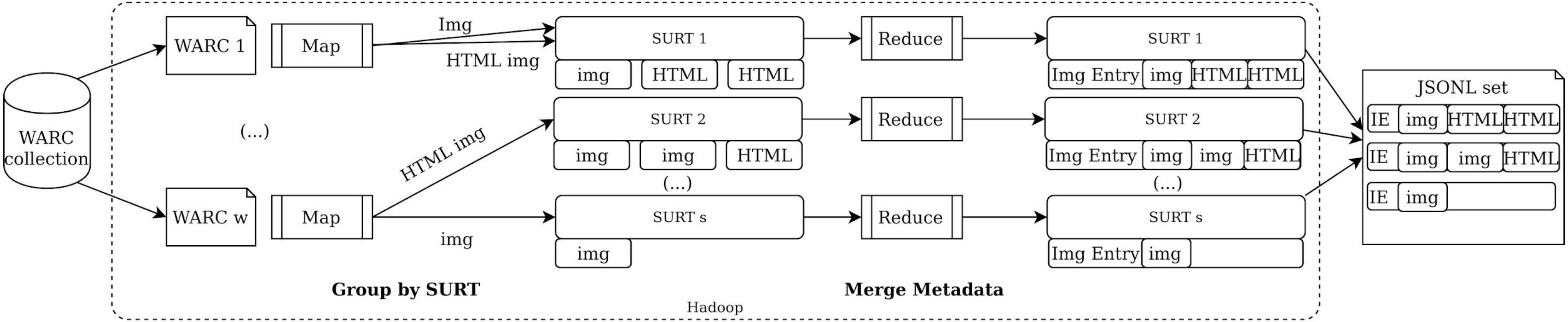
Group by SURT

- For each record in the WARC
    - Find all image entries, extract metadata (height, width, img digest) and add them to the image SURT reduce list
    - Find all *<img>/<a>/css* entries in the HTML, extract metadata and add them to the image SURT reduce list

Merge metadata

- For each entry in the SURT, merge metadata for that record to produce a single record for each unique image (measured by digest)
    - This includes keeping *img title* and *alt* entries for images that show up in more than one page
    - Only add pages if they offer new image metadata *(title, alt or caption)*!

Map Reduce: Extract images and metadata



How to deal with duplicate information?

- The amount of data produced by this step is huge!
- Generating a lot of documents for indexing
- But most of this information is duplicate
  - Images and pages that were crawled at different times but have not changed
  - References to the images that have the same caption/metadata

Deduplication potential solutions

- After careful examination, we arrived at the 3 deduplication scenarios:
  a. every page-image pair is a document
  b. the oldest page that references the image is the canonical document
  c. oldest page information and image specific information from all pages
     - keep reference to oldest page
     - Add all new image specific information (title, alt, caption) to the document
     - replace oldest page reference if a new oldest document shows up

What users want in image search?

- Assumptions
  - Users use page search to find pages
  - Users do not want to see duplicate images on the search results
  - Users do not use image search to find pages
  - No need to find all the pages that contain a given image
  - When an image appears on more than one page, finding the oldest page best matches the information need of a web archive user
    - Finding the page that better matches the image is not necessary
  - Technical details (imgAlt, ...) are rarely accessed by users
- One Solr document per image with all page information
  - Store page metadata for the oldest page
  - Store image specific metadata from all pages in a combined field
  - Remove fields that do not matter
- Expected a decrease of 25-50% in index size

Images in multiple pages

Indexed:

```
"id":"a4236137f5455edd8436a5d122c366fa62ba91139f45895f447565b3a6b926bb",
"imgSrc":"http://bp3.blogger.com/_v4YQruRZWLI/RxTDBM9FL5I/AAAAAAAAALY/YJTbzjdOKH0/s320/2.jpg",
"pageTstamp":"2008-02-15T08:40:21Z",
"imgTstamp":"2008-02-23T09:36:42Z",
"pageURL":"http://www.worksfromthecave.blogspot.com/",
"collection":["AWP1"],
"caption":["great"],
```

To index:

```
"id":"a4236137f5455edd8436a5d122c366fa62ba91139f45895f447565b3a6b926bb",
"imgSrc":"http://bp4.blogger.com/_v4YQruRZWLI/RxTDBM9FL5I/AAAAAAAAALY/YJTbzjdOKH0/s320/2.jpg",
"pageTstamp":"2004-02-15T08:40:21Z",
"imgTstamp":"2009-02-23T09:36:42Z",
"pageURL":"http://www.worksfromthecave.sapo.pt/",
"collection":["AWP3"],
"caption":["fantastic"],
```

Images in multiple pages

Indexed:

```
"id":"a4236137f5455edd8436a5d122c366fa62ba91139f45895f447565b3a6b926bb",
"imgSrc":"http://bp3.blogger.com/_v4YQruRZWLI/RxTDBM9FL5I/AAAAAAAAALY/YJTbzjdOKH0/s320/2.jpg",
"pageTstamp":"2008-02-15T08:40:21Z",
"imgTstamp":"2008-02-23T09:36:42Z",
"pageURL":"http://www.worksfromthecave.blogspot.com/",
"collection":["AWP1"],
"caption":["great"],
```

To index:

```
"id":"a4236137f5455edd8436a5d122c366fa62ba91139f45895f447565b3a6b926bb",
"imgSrc":"http://bp4.blogger.com/_v4YQruRZWLI/RxTDBM9FL5I/AAAAAAAAALY/YJTbzjdOKH0/s320/2.jpg",
"pageTstamp":"2004-02-15T08:40:21Z",
"imgTstamp":"2009-02-23T09:36:42Z",
"pageURL":"http://www.worksfromthecave.sapo.pt/",
"collection":["AWP3"],
"caption":["fantastic"],
```

Images in multiple pages

Final:

```
"id":"a4236137f5455edd8436a5d122c366fa62ba91139f45895f447565b3a6b926bb",
"imgSrc":"http://bp4.blogger.com/_v4YQruRZWLI/RxTDBM9FL5I/AAAAAAAAALY/YJTbzjdOKH0/s320/2.jpg",
"pageTstamp":"2004-02-15T08:40:21Z",
"imgTstamp":"2009-02-23T09:36:42Z",
"pageURL":"http://www.worksfromthecave.sapo.pt/",
"collection":["AWP1", "AWP3"],
"caption":["great", "fantastic"]],
```

Deduplication selected solution

- After careful examination, we arrived at the 3 deduplication scenarios:
    a. every page-image pair is a document
    b. the oldest page that references the image is the canonical document
    c. **oldest page information and image specific information from all pages**
        - keep reference to oldest page
        - Add all new image specific information (title, alt, caption) to the document
        - replace oldest page reference if a new oldest document shows up

Group by digest

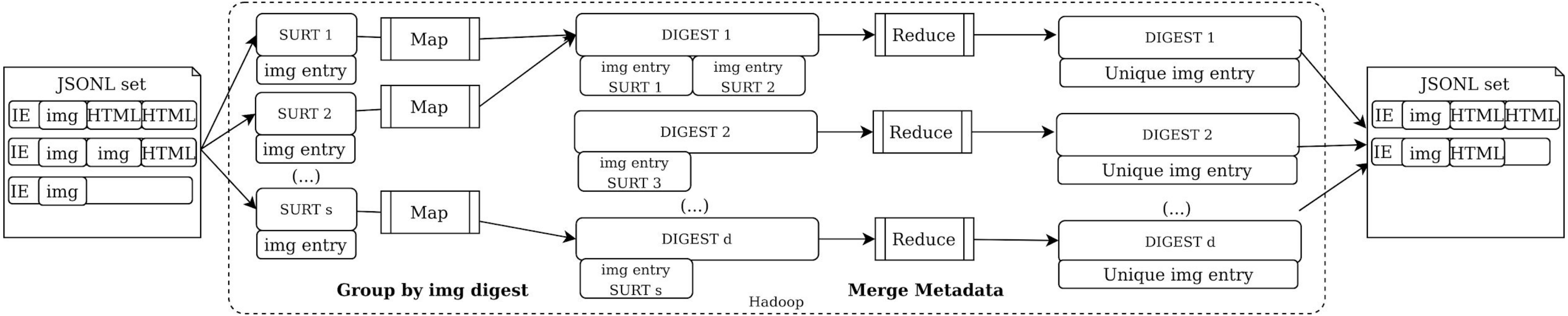
Group by Digest

- For each image record in the JSONL, send it to the matching Digest list

Merge metadata

- For each entry in the Digest list, merge metadata for that record to produce an unique record for each image
  - Similar to the previous merge by SURT step
- If there are multiple Digests for the same URL:
  - Pages are updated to represent the data of the image that is closest in capture time
  - Additional image information is added to imgAlt, Title and Caption fields

Map Reduce: Group by digest



Duplicates across collections

- Hadoop processing is performed across per collection
  - To better manage computing resources (e.g. HDFS disk space)
  - Thus, deduplication is only performed on a per-collection basis
- We added an extra “group by digest” step when sending docs to Solr

Summary

1. Find all images in records and find image references in pages
2. Group by SURT
    a. store only pages that have new metadata
3. Regroup by Digest
    a. create new records for images with multiple digests
4. Find best image entry for each image reference
5. Send to Solr

Popularity fields

- Extracting multiple versions of each image and pages opens up a world of possibilities!
    - Find how individual pages and images evolve over time (change digests)!
    - Images that appear in more than one page are more or less relevant?
    - Images that change metadata often are less relevant?
    - ….

Metadata: Popularity fields

matchingImages

- number of times the image was crawled (by image content digest)

matchingPages

- number of times the image was referenced on <*img*> tags, css or JS

imagesInOriginalPage

- number of images in the oldest page

imageMetadataChanges

- number of times that the image metadata (alt, title or caption) changes

pageMetadataChanges

- number of times that the page metadata (title) changes

Takeways

- **Faster WARC parsing**
  - **Fixed two times pass** and WARC download errors
  - (**3 ms -> less than 0.5 ms** per image)
- **A lot more images found!**
  - We will see how many in the following slides...
- **Multiple pages per image** (current ratio: **~2 per image**)
- **Removed unneeded bottlenecks** (MongoDB)
- **Logging the indexing process**
  - Hadoop counters for errors
  - Metadata counters for images found and collected

My predictions in May 2020

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22 million**
Collection count	90
(W)ARCs	3 million
(W)ARC sizes	334 TB
Total collected files	6,000 million
Total collected images	**1,602 million**
Oldest image date	15/04/1994
Newest image date	14/11/2019
Daily page views	~87

Tested collections - number of images

Collection	Old Parser	New Parser	Diff New to Current	Ratio vs New
AWP24	865,589	14,133,997	+13,268,408	16.33
AWP15	552,275	26,127,269	+25,574,994	47.31
FAWP26	213,527	1,562,617	+1,349,090	7.32
Tomba	169,308	1,076,967	+907,659	6.36
BlogsSapo2018	71,668	752,679	+681,011	10.50
Weblog	6,336	87,252	+80,916	13.77
DinisAlves2018	1,215	1,216	+1	1.00
DEM-IST	191	360	+169	1.88
BlocoEsquerda	15	16	+1	1.07

Takeways

~200-650 million images

1,880,124 -> 43,742,373

~9-28x more images

[illegible] -> [illegible]

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

~400-1,300 million pages (2/image)

~18-56x more pages

Takeways

~200-650 million images

1,880,124 -> 43,742,373

~9-28x more images

23,589,395 -> ???,???,???

~400-1,300 million pages (2/image)

~18-56x more pages

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	225,868	3,414,742	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

Takeways

654 million images

1,880,124 -> 43,742,373

23,589,395 -> ???,???,???

29x more images

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

1,252 million pages (1.91/image)

55x more pages

But Arquivo.pt kept growing on 2020

Takeways

+ 317 million images in one year (2019)

1,880,124 -> 43,742,373

48% growth

23,589,395 -> [illegible]

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	[illegible]	652,203,512	27.65x

+ 610 million pages in one year (2019)

49% growth

Takeways

971 million images

1,880,124 -> 43,742,373

42x more images

23,589,395 -> ???,???,???

1,862 million pages (1.91/image)

81x more pages

Indexed images	1,880,124	23,589,395	548,823,437	23.27x
Crawl/Collection count	9	88	427,703,203	18.13x
(W)ARCS	[illegible]	[illegible]	[illegible]0,669	28.03x
(W)ARC sizes	21.43 TB	336.47 TB	686,806,771	29.12x
Total collected files	408,230,995	6,086,768,283	652,203,512	27.65x

Impact of deduplication

	Number of documents
a	1,862 million image-page pair documents
b	584 million unique documents (971 million before deduplication across collections)
c	**584 million** documents, containing information from all 1,862 million image-page pairs

How will we index these **584 million** documents?

Current Solr indexing architecture

ARQUIVO.PT

Current image index has **31 million** documents

(22,881,688 plus some special crawls we added in 2020)

on one 20 core, 40 thread server with 512 GB RAM*

* one server per branch, two redundant branches

running Solr 6.3 with a 735 GB index

What to do with new data?

Our indexing process resulted in

584 million documents

(expected index size: ~720GB)

Where will we fit all this data?

Arquivo.pt response time guidelines

The 355 rule

- **3 responses per second**
- With an average query time **below 5 seconds**
- For **5 concurrent users**

- We are currently performing these experiments

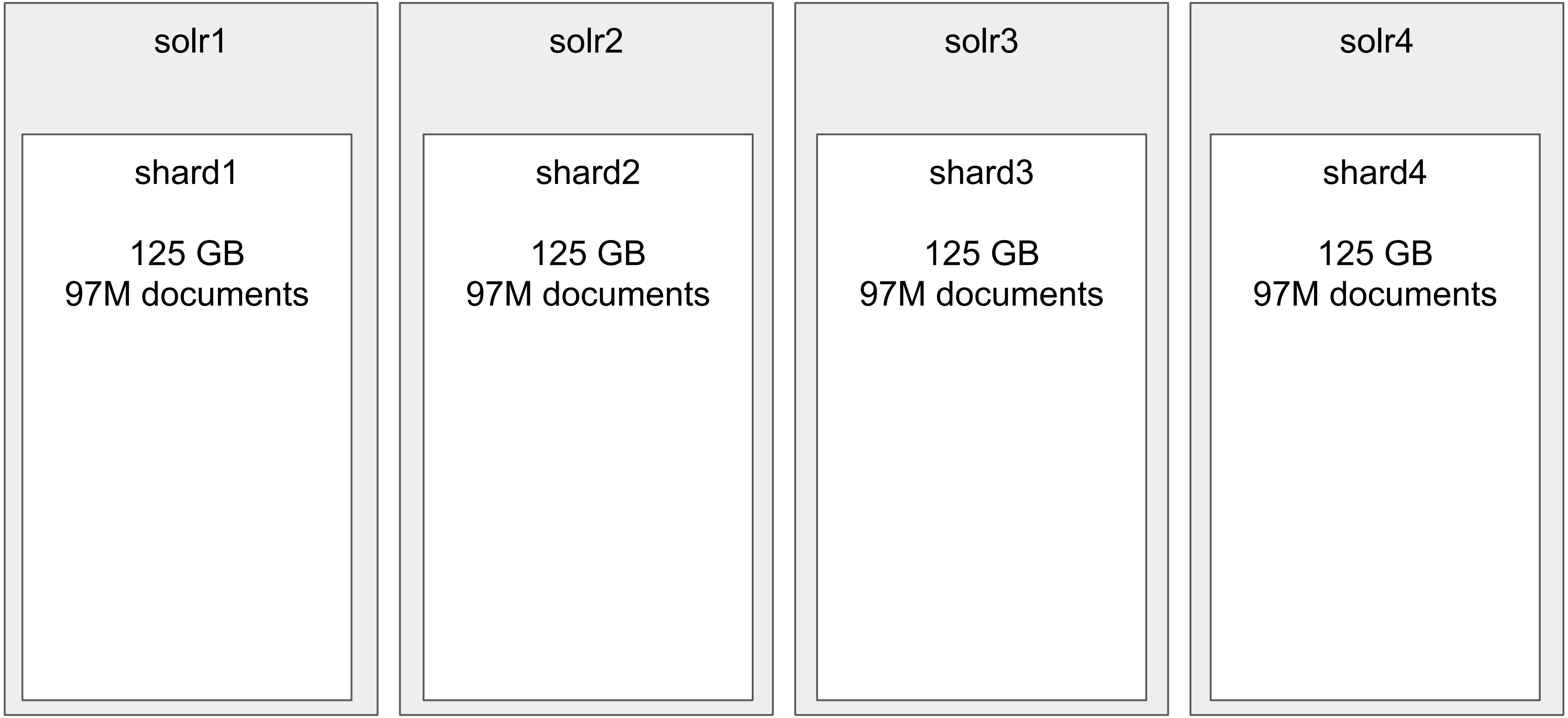
Planning SolrCloud resource allocation

- Expected index size: **~720GB**

- SolrCloud servers:
  - 8 servers, 4 per branch
    - **512GB**: p87, p91 (20/40 cores/threads)
    - **256GB**: p82, p83 (12/24 c/t), p93, p94, p98, p99 (20/40 c/t)
  - **2560GB** total, **1280GB** per branch

- No SSD, only HDD, but we have more RAM than indexed data

How we configured SolrCloud? - Try 1



solr1

shard1

125 GB
97M documents

solr2

shard2

125 GB
97M documents

solr3

shard3

125 GB
97M documents

solr4

shard4

125 GB
97M documents

Solr performance factors

- Available RAM for index file caching
  - slowdown happens when index size > RAM

....

Solr performance factors

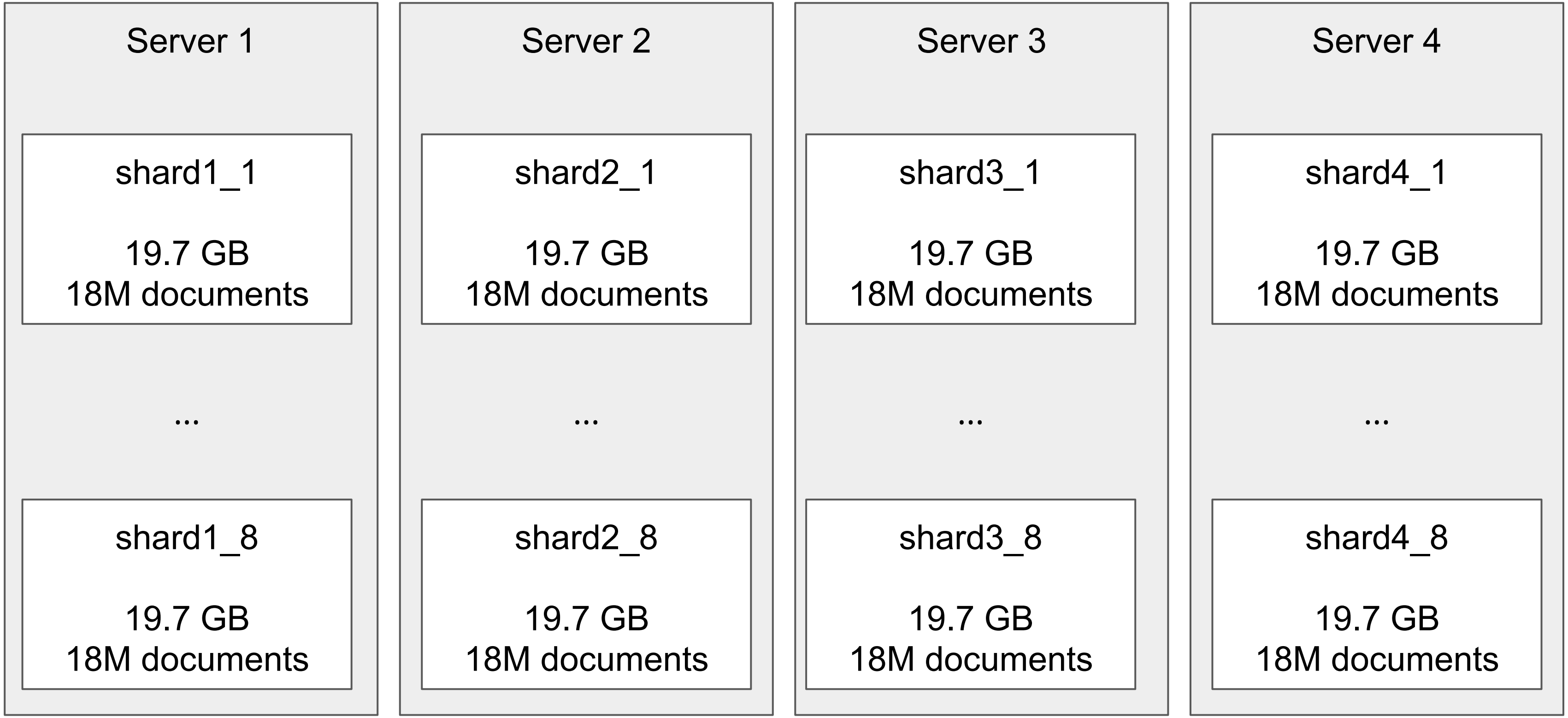
- Available RAM for index file caching
    - slowdown happens when index size > RAM or

  ....

  disk I/O skyrockets

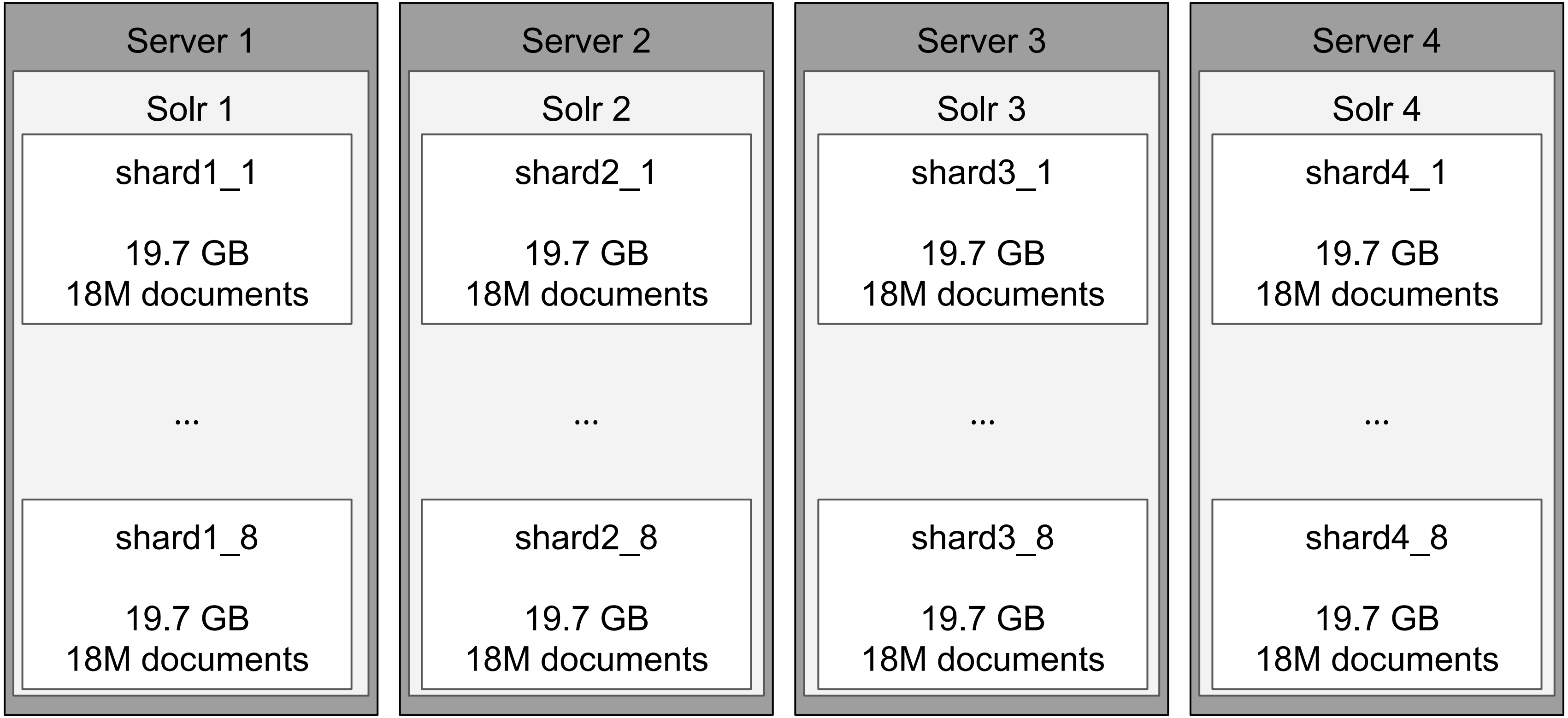
  and that is basically it
    - CPU or network are not the current bottleneck

How we configured SolrCloud? - Try 2



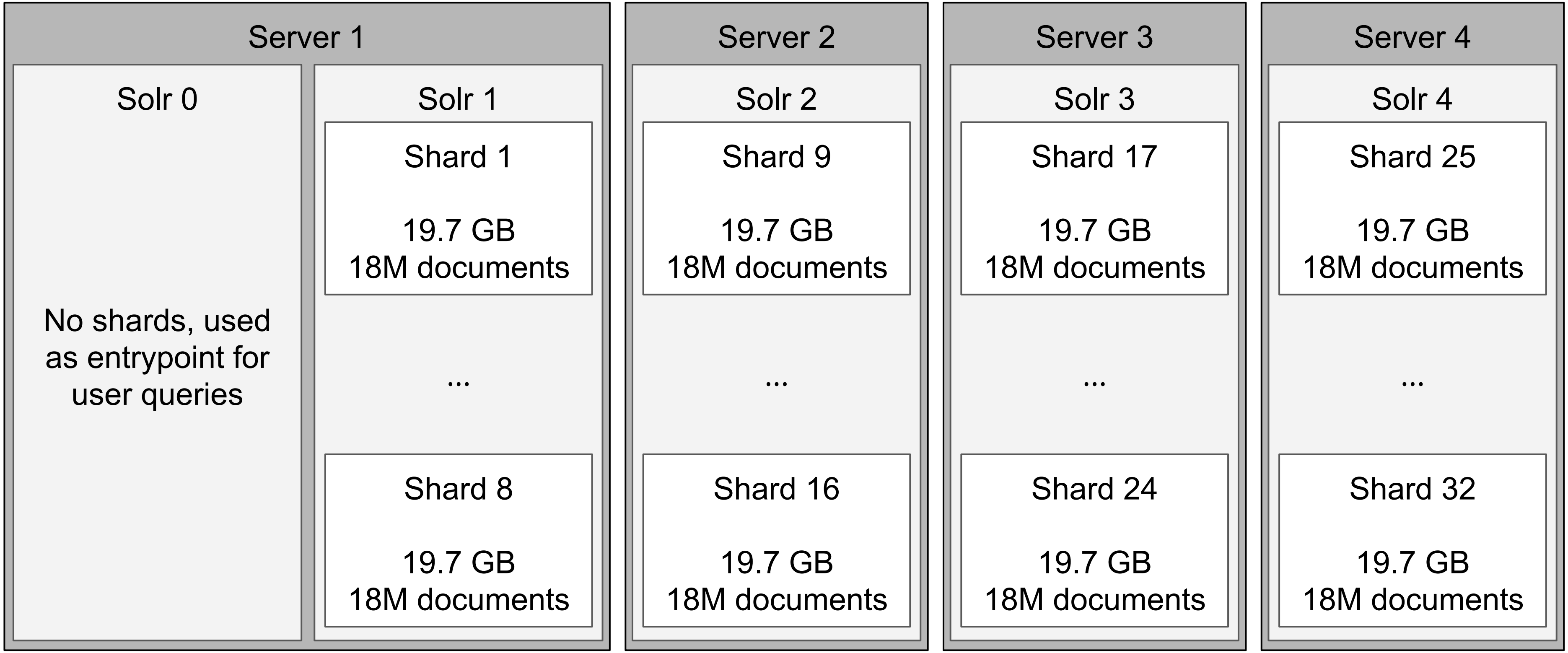
Server 1	Server 2	Server 3	Server 4
shard1_1 19.7 GB 18M documents	shard2_1 19.7 GB 18M documents	shard3_1 19.7 GB 18M documents	shard4_1 19.7 GB 18M documents
...	...	...	...
shard1_8 19.7 GB 18M documents	shard2_8 19.7 GB 18M documents	shard3_8 19.7 GB 18M documents	shard4_8 19.7 GB 18M documents

How we configured SolrCloud? - Try 2



Server 1	Server 2	Server 3	Server 4
Solr 1	Solr 2	Solr 3	Solr 4
shard1_1 19.7 GB 18M documents	shard2_1 19.7 GB 18M documents	shard3_1 19.7 GB 18M documents	shard4_1 19.7 GB 18M documents
...	...	...	...
shard1_8 19.7 GB 18M documents	shard2_8 19.7 GB 18M documents	shard3_8 19.7 GB 18M documents	shard4_8 19.7 GB 18M documents

How we configured SolrCloud? - Try 2



Server 1

Solr 0

No shards, used as entrypoint for user queries

Solr 1

Shard 1

19.7 GB
18M documents

...

Shard 8

19.7 GB
18M documents

Server 2

Solr 2

Shard 9

19.7 GB
18M documents

...

Shard 16

19.7 GB
18M documents

Server 3

Solr 3

Shard 17

19.7 GB
18M documents

...

Shard 24

19.7 GB
18M documents

Server 4

Solr 4

Shard 25

19.7 GB
18M documents

...

Shard 32

19.7 GB
18M documents

How to test?

- Search with increasing concurrent users
    - 1, 3, 5, 10, 20, 50 concurrent users

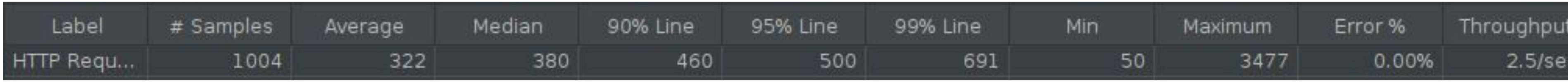
- For a set period of time
    - 5 minutes

How to select realistic queries?

- Two sets of queries:
  - User queries extracted from logs
  - Random pairs of Portuguese words
- Warmup the index using 50 queries
- Query for 5 minutes and parse the results

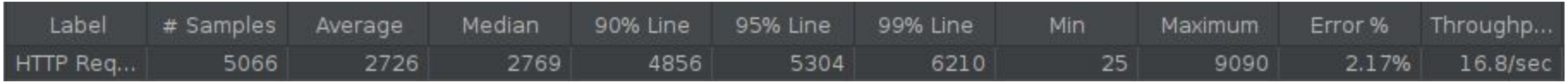
(Fresh off the press) results

Single user, random queries (pairs of portuguese words)



Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Maximum	Error %	Throughpu
HTTP Requ...	1004	322	380	460	500	691	50	3477	0.00%	2.5/se

50 users, random queries (pairs of portuguese words)

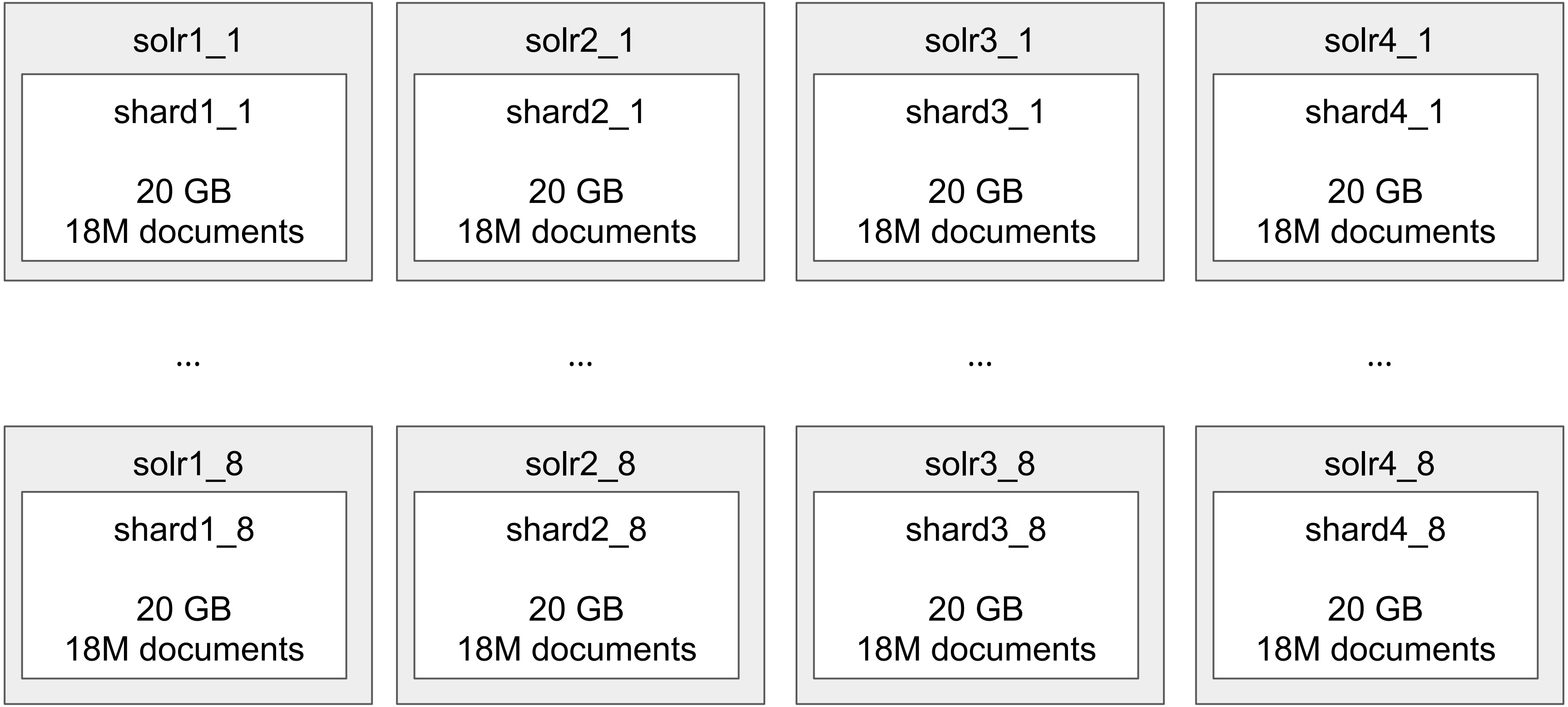


Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Maximum	Error %	Throughp...
HTTP Req...	5066	2726	2769	4856	5304	6210	25	9090	2.17%	16.8/sec

Tips and parameters

- vmtouch tool to force OS to keep index files in RAM
- Heap size: 31GB
  - Smaller sizes made Solr crash on parallel query situations
  - Larger sizes means Java can’t use compressed pointers https://lucene.apache.org/solr/guide/8_7/taking-solr-to-production.html#running-multiple-solr-nodes-per-host

How we configured SolrCloud?



solr1_1

shard1_1

20 GB
18M documents

solr2_1

shard2_1

20 GB
18M documents

solr3_1

shard3_1

20 GB
18M documents

solr4_1

shard4_1

20 GB
18M documents

...

...

...

...

solr1_8

shard1_8

20 GB
18M documents

solr2_8

shard2_8

20 GB
18M documents

solr3_8

shard3_8

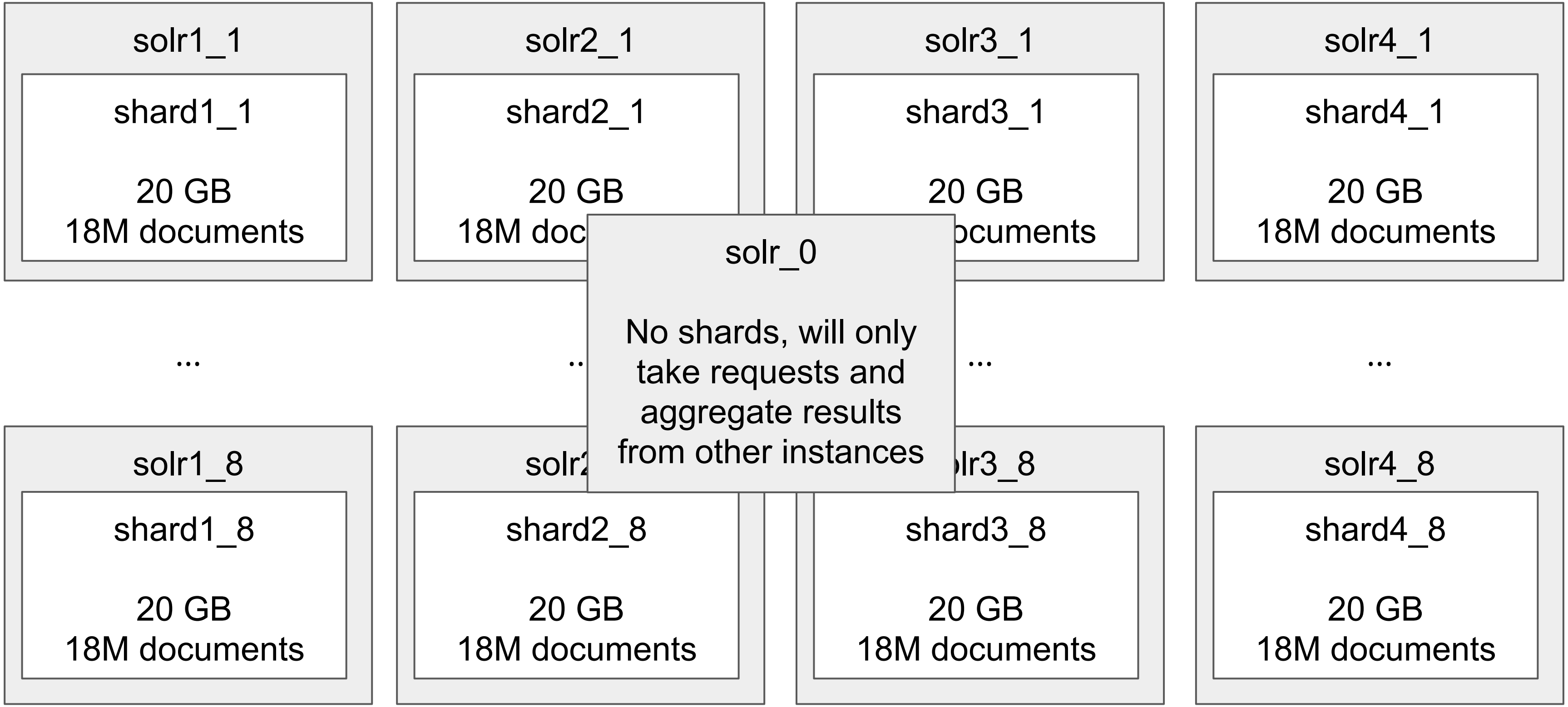
20 GB
18M documents

solr4_8

shard4_8

20 GB
18M documents

How we configured SolrCloud?



solr1_1

shard1_1

20 GB
18M documents

solr2_1

shard2_1

20 GB
18M doc

solr3_1

shard3_1

20 GB
ocuments

solr4_1

shard4_1

20 GB
18M documents

solr_0

No shards, will only take requests and aggregate results from other instances

...

..

...

...

solr1_8

shard1_8

20 GB
18M documents

solr2

shard2_8

20 GB
18M documents

lr3_8

shard3_8

20 GB
18M documents

solr4_8

shard4_8

20 GB
18M documents

Future problems: Migrate page search to SolrCloud

ARQUIVO.PT

- Currently, we have an highly customized version of Lucene optimized not to search the full posting lists

- Scale
  - 6-7,000 million documents
  - 5 servers with 4.5TB of RAM in total

Modernize NSFW libs

What about NSFW content?

- Arquivo.pt captures pages and images from **all** over the web
- This includes content that may me offensive to users
- Arquivo.pt uses an image based NSFW content classifier
- Images marked as NSFW are filtered by default from image search results

Arquivo’s NSFW classifier

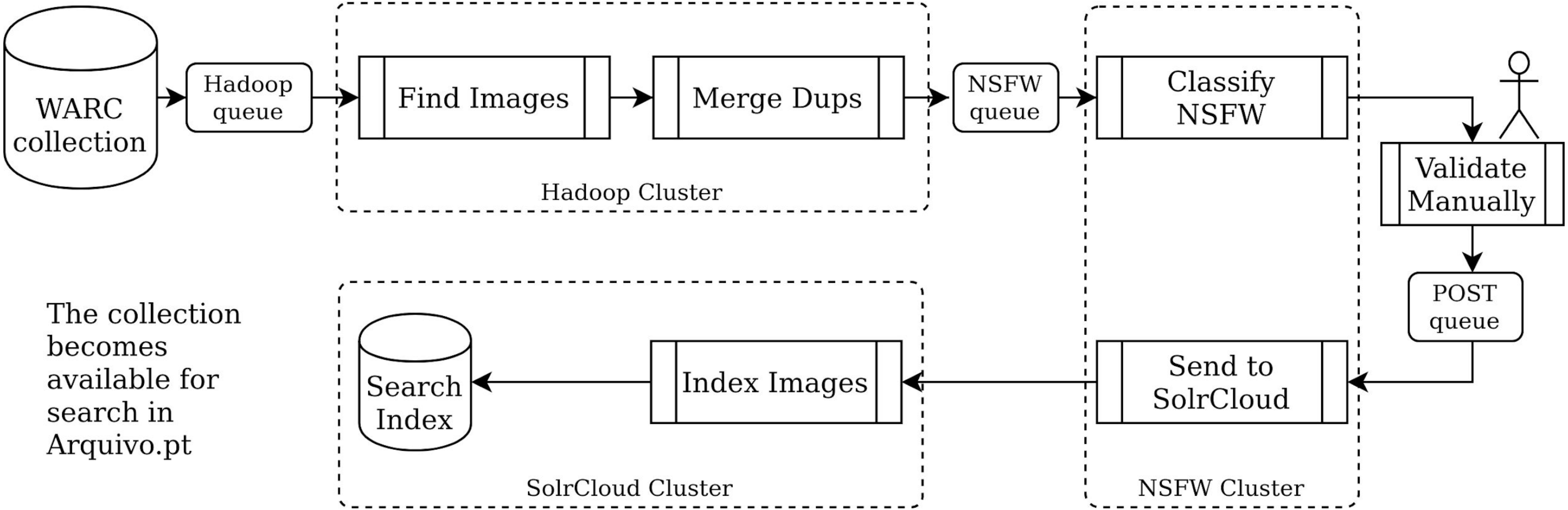
- Based in Keras ResNet
  - https://github.com/GantMan/nsfw_model
  - Reported 93% precision
  - Measurements in our test collection match it at about 90% precision
  - ~250 images per second per GPU times 2 GPUs
- Extra features
  - Detect whether the image is a drawing/vectorial or picture

NSFW

- Antigo: ~40 imagens por segundo
  - estimativa para processamento dos novos dados (assumindo 2 GPU): ~150 dias
- Novo: ~250 imagens por segundo
  - estimativa para processamento dos novos dados (assumindo 2 GPU): ~30 dias
-
- Antigo:
  - Precision   Recall   F1
  - 0.92   0.94   0.93
- Novo:
  - Precision   Recall   F1
  - 0.94   0.88   0.91

Architecture/Pipeline

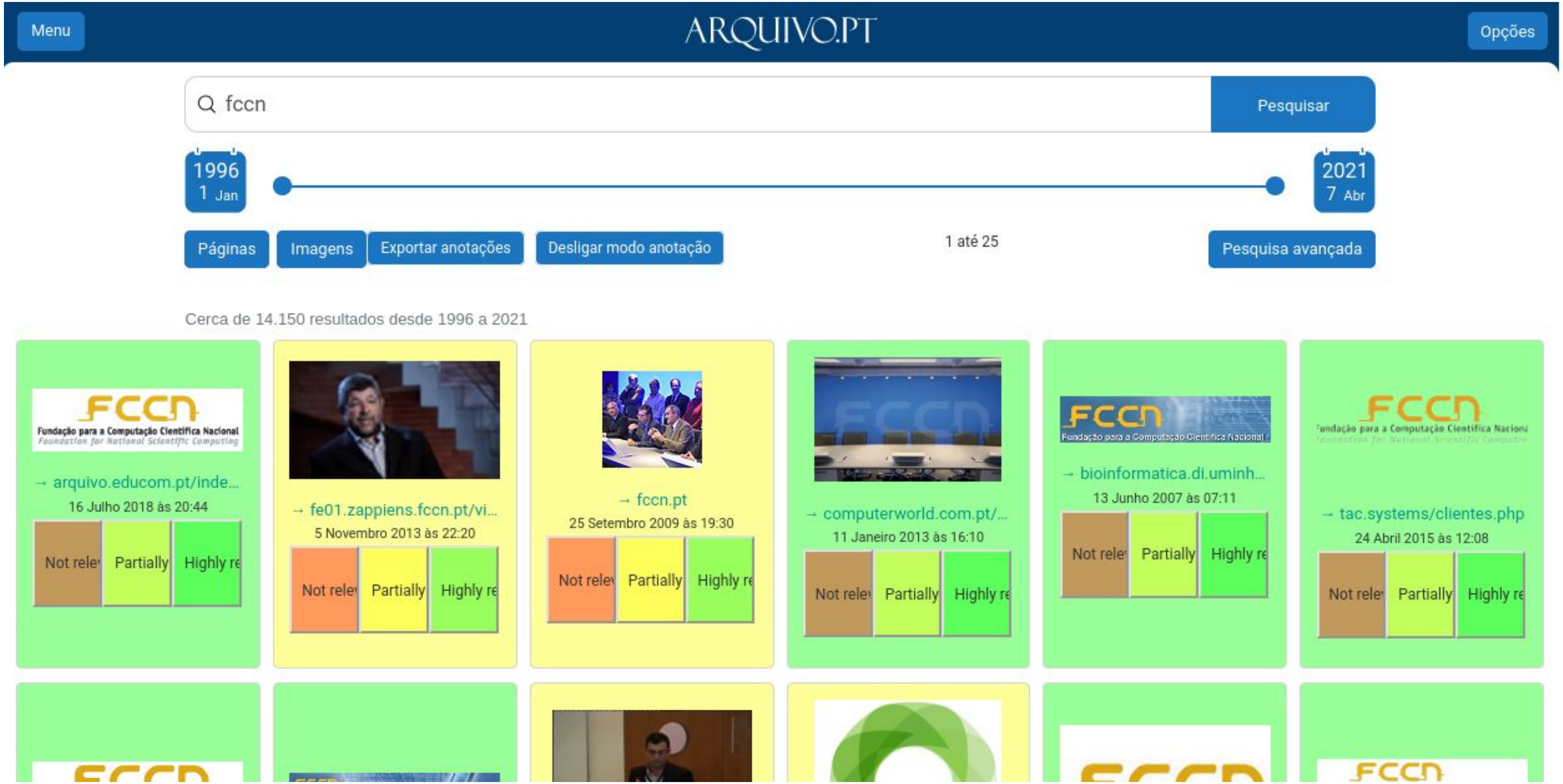
Queue



The collection becomes available for search in Arquivo.pt

Relevance assessment

Annotator



fccn

1 até 25

Cerca de 14.150 resultados desde 1996 a 2021

→ arquivo.educom.pt/inde...
16 Julho 2018 às 20:44

→ fe01.zappiens.fccn.pt/vi...
5 Novembro 2013 às 22:20

→ fccn.pt
25 Setembro 2009 às 19:30

→ computerworld.com.pt/...
11 Janeiro 2013 às 16:10

→ bioinformatica.di.uminh...
13 Junho 2007 às 07:11

→ tac.systems/clientes.php
24 Abril 2015 às 12:08

Results on TestCollection (2020)

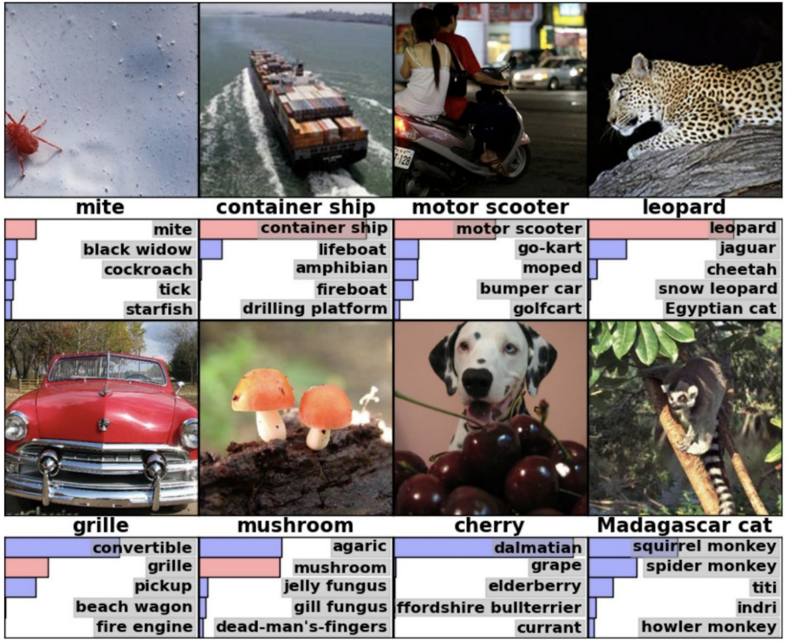
Metric	**Arquivo.pt**
mAP	0.5471
nDCG@1	0.6800
nDCG@5	0.5480
nDCG@10	0.4800
nDCG@20	0.4270
P@1	0.6800
P@5	0.5930
P@10	0.5703
P@20	0.5834
S@1	0.6800
S@5	0.8200
S@10	0.8600
S@20	0.9000

Summary of what changed in 2020?

- More metadata per image
  - All pages that mention the image are parsed
  - Heuristic extraction of image captions from the HTML page structure
  - Additional features extracted from the HTML and images
- Improved NSFW image processing
  - 7x faster processing (40 -> 280 images per second)
  - Returns more image information for ranking (e.g. drawing vs. photo)
- Improved indexing architecture and processing
  - Removed MongoDB dependency
  - Ensure all archived images and pages are parsed
  - Find images in <a> links, CSS and JS code
- Distributed search index
  - Transition from single node Solr to distributed SolrCloud architecture
  - Improved schema so that the index only grows by 32% when covering 81x more images

Plan for the future

- Deal with images that have **no metadata**
  - Cannot find pages for 300+ million images
  - Deep Image classification, **tag extraction**
- Content based hashes
  - Similar images show up all over the place (different resolutions and formats)
  - Find and deduplicate **near duplicates**
- Improve Solr **ranking**
  - Use the newly extracted popularity features



2020 vs 2021

January 2020	January 2021	Improvement
22 million images in pages (only one version of the image is indexed per collection)	1,862 million images in pages	**81x** more images in pages parsed
	967 million image files	**42x** more image files parsed
17 million deduplicated documents	584 million deduplicated documents	**33x** more unique images, removing duplicates across collections
49% have image metadata (imgAlt, imgTitle)	99%+ have image metadata (imgAlt, imgTitle, imgCaption)	**+51 p.p.** images can be found with relevant contextual information
~570 GB search index	~750 GB search index	**Only 32% larger** after a **813% increase** in information parsed
1 Solr server	4 SolrCloud servers	**Only 3 more nodes** after a **813% increase** in information parsed

2020 vs 2021

January 2020	**January 2021**	**Improvement**
22,881,688 images in pages (only one version of the image is indexed per collection)	1,862,311,456 images in pages	**81.39x** more images in pages parsed
	967,184,126 image files	**42.26x** more image files parsed
17,643,047 deduplicated documents	584,242,176 deduplicated documents	**33.11x** more unique images, removing duplicates across collections
48.7% have image metadata (imgAlt, imgTitle)	99.6% have image metadata (imgAlt, imgTitle, imgCaption)	**+50.9 p.p.** images can be found with relevant contextual information
~570 GB search index	~750 GB search index	**Only 32% larger** after a **813% increase** in information parsed
1 Solr server	4 SolrCloud servers	**Only 3 more nodes** after a **813% increase** in information parsed

Ranking features for 2021

imgCaption

- portion of the HTML page text that is closest to the image

matchingImages

- number of times the image was crawled (by image content digest)

matchingPages

- number of times the image was referenced on <*img*> tags, css or JS

imagesInOriginalPage

- number of images in the oldest page

imageMetadataChanges

- number of times that the image metadata (alt, title or caption) changes

pageMetadataChanges

- number of times that the page metadata (title) changes

drawing/photo

- whether the image is a drawing or a photo