Pesquisando milhões de imagens no Arquivo.pt

23 de abril de 2021

André Mourão, Ph.D.
Engenheiro I&D
andre.mourao@fccn.pt

FCT
Fundação para a Ciência e a Tecnologia

Uma página da Internet como tantas outras

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

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

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

Imagens como parte significativa de uma página

Imagens

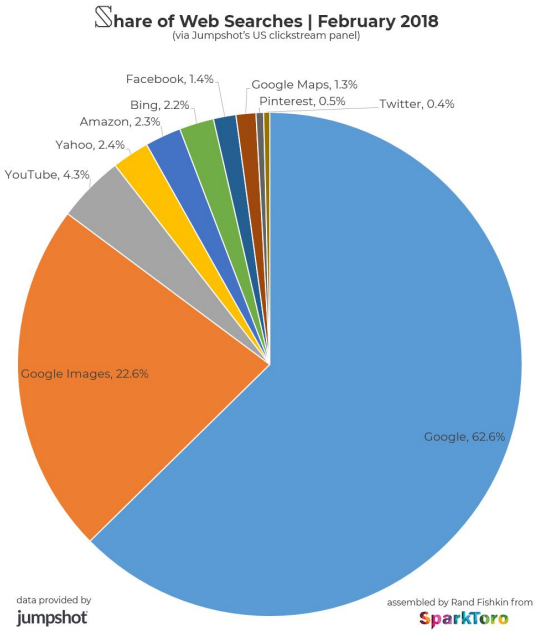
Imagem

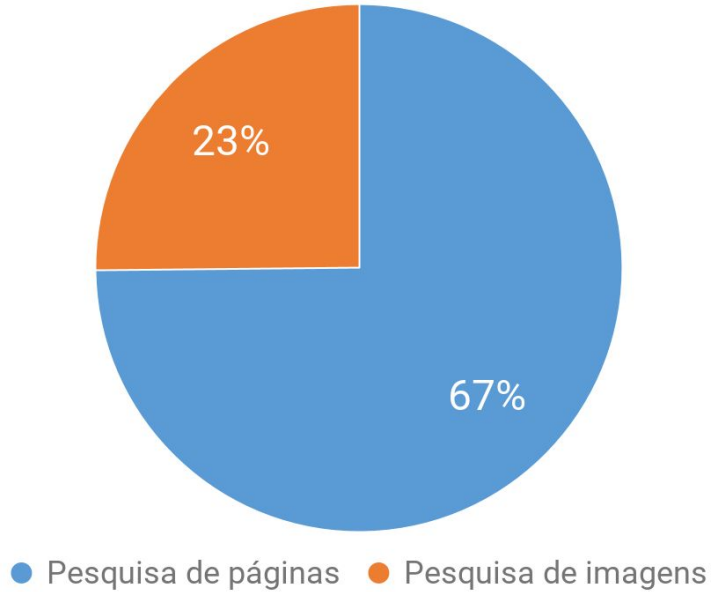
Imagens

Imagens

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

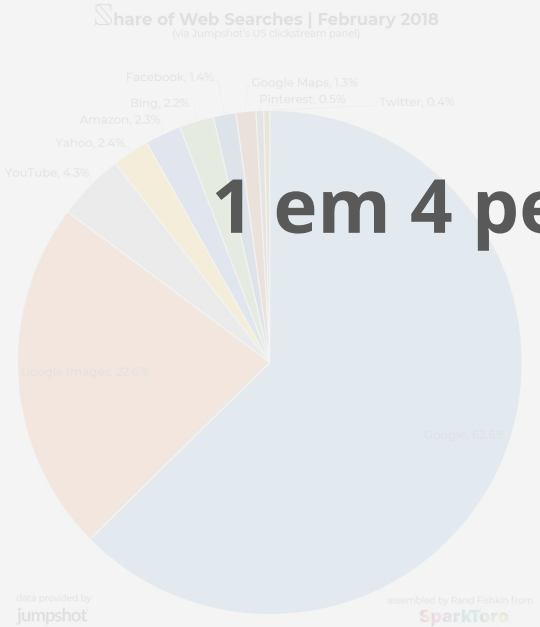
A pesquisa de imagens é importante!



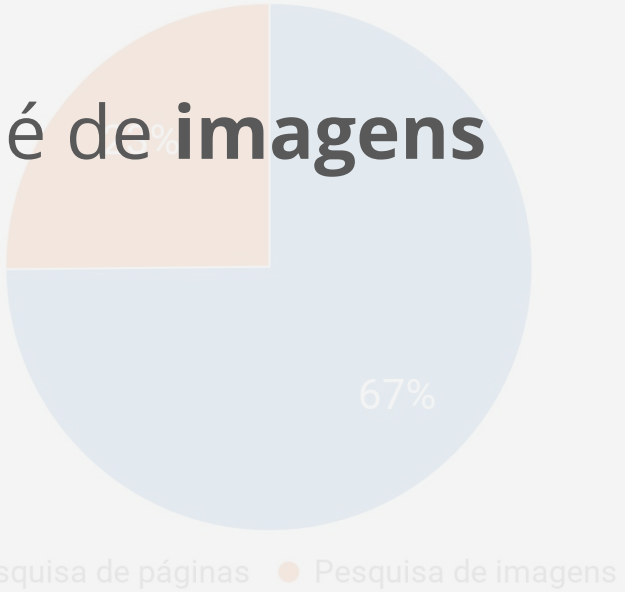


sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

A pesquisa de imagens é importante!

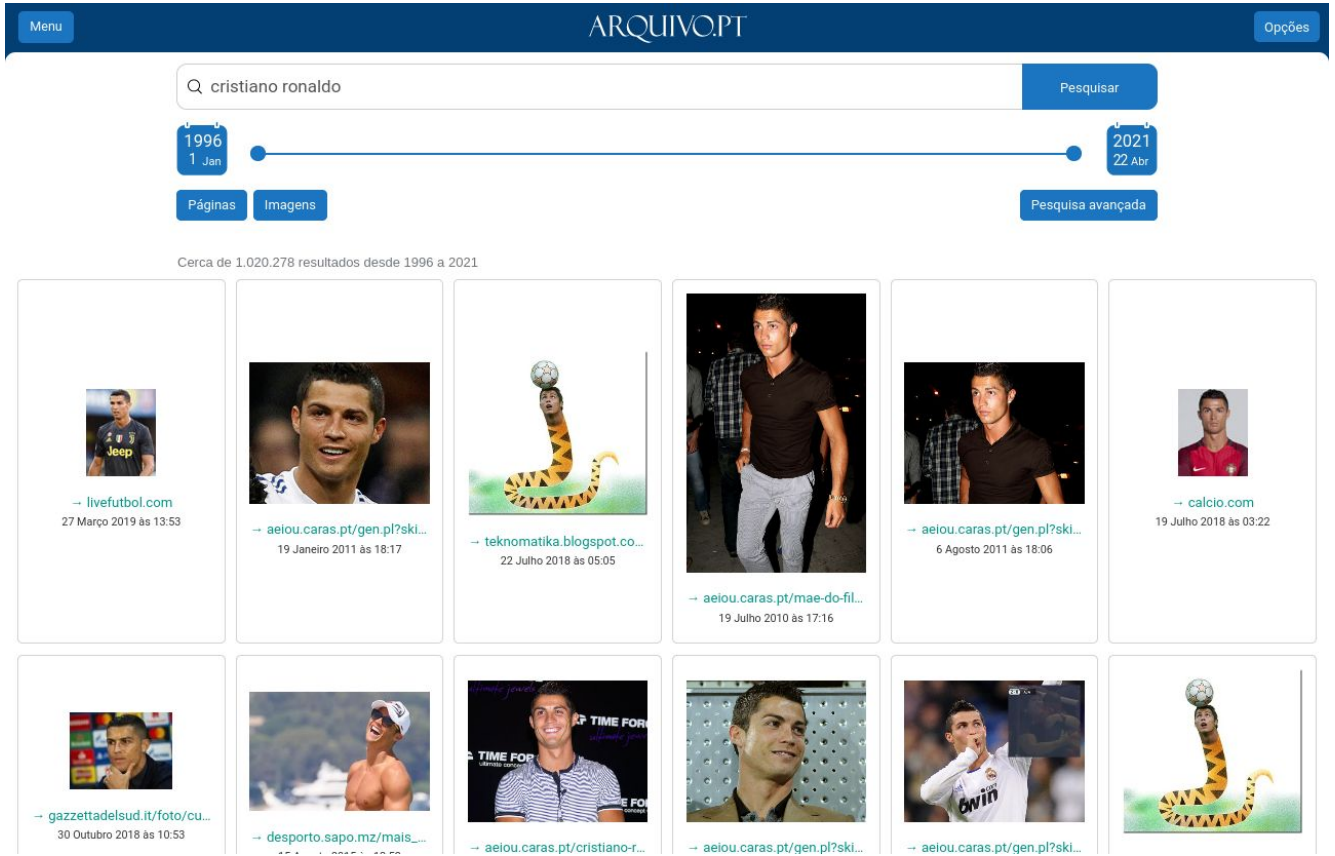


1 em 4 pesquisas é de **imagens**

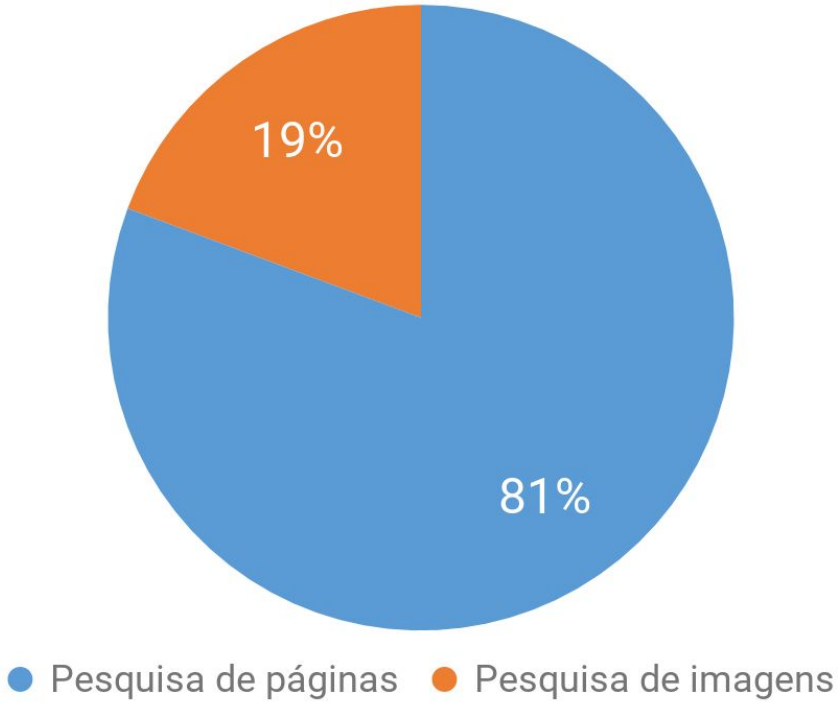


sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Pesquisa de imagens do Arquivo.pt



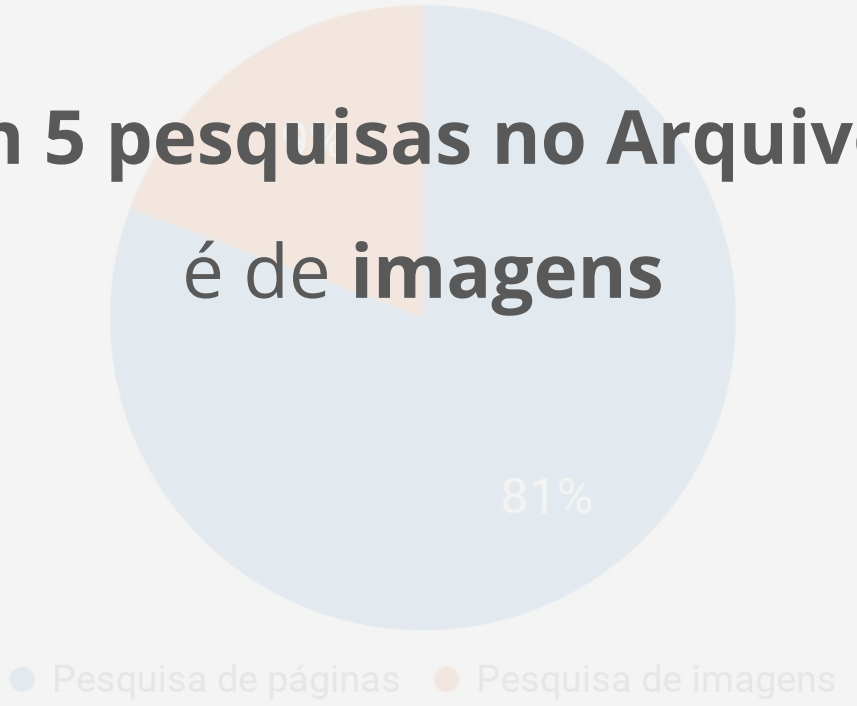
Distribuição da pesquisa de imagens no Arquivo.pt



Distribuição da pesquisa de imagens no Arquivo.pt

1 em 5 pesquisas no Arquivo.pt

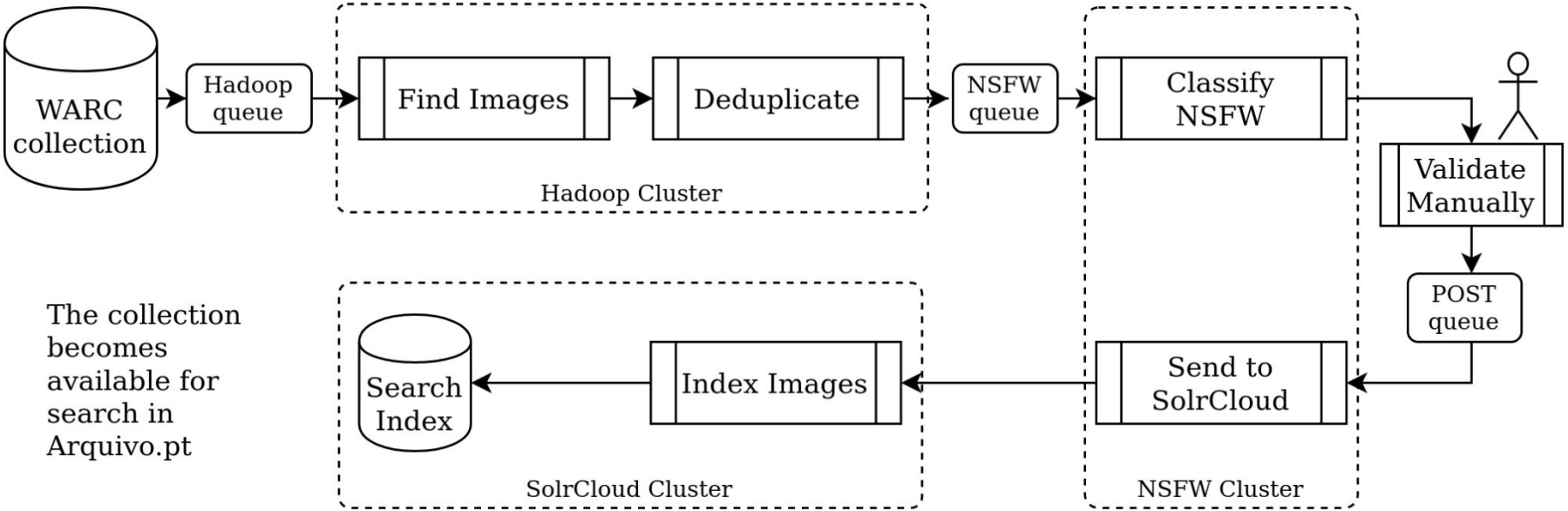
é de **imagens**



20 para 1 800 milhões de imagens... como?

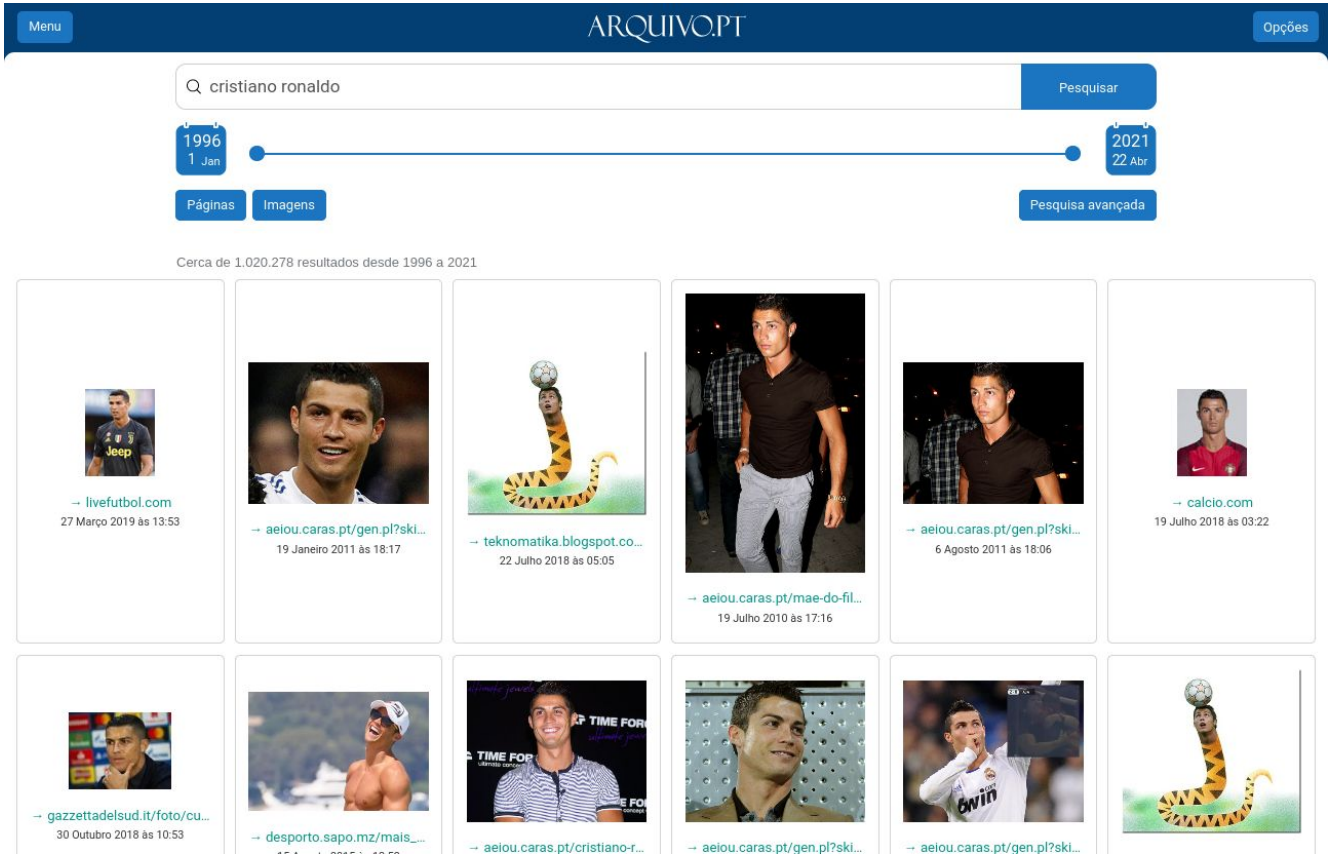
- Encontrar palavras relevantes para imagens
- Lidar com a escala dos dados recolhidos
- Tornar estas imagens pesquisáveis

Fluxo de indexação de imagens



Encontrar palavras relevantes para imagens

Palavras para pesquisar imagens



Utilizadores pesquisam através da inserção de palavras numa caixa de pesquisa

Os resultados apresentam imagens relacionadas com as palavras pesquisadas

Como associar palavras descritivas de imagens?

- “Uma imagem vale mais do que mil palavras”
- Os computadores ainda não sabem interpretar imagens como humanos
  - Embora com técnicas de *deep learning* estejam cada vez mais perto!
- As imagens apenas têm um URL e data de captura, o que não é descritivo
- Como associar palavras descritivas de imagens?

A anatomia de uma página Web

Notícia relativa à visita do Marcelo à Índia

- Imagem principal:
  - Marcelo Rebelo de Sousa discursando como parte de uma visita estatal à Índia
  - Legenda: "Marcelo Rebelo de Sousa LUSA/ESTELA SILVA"
- Imagens secundárias:
  - Ícones de partilhas em redes sociais
  - Logótipo do Público
  - Foto de um autor de uma crónica
- Outras:
  - Ligações a imagens externas
  - Imagens como fundo CSS

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

A anatomia de uma página Web

Notícia relativa à visita do Marcelo à Índia

- Imagem principal:
  - Marcelo Rebelo de Sousa discursando como parte de uma visita estatal à Índia
  - Legenda: "Marcelo Rebelo de Sousa LUSA/ESTELA SILVA"
- Imagens secundárias:
  - Ícones de partilhas em redes sociais
  - Logótipo do Público
  - Foto de um autor de uma crónica
- Outras:
  - Ligações a imagens externas
  - Imagens como fundo CSS

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

Tipos de referências a imagens em HTML

	Percentagem de referências
<img>	90%
<a>	9%
css	1%

```
<img src="image1.jpg" alt="This is an image">
<a href="image2.png">
  This is a link to an image.
</a>
<div style="background-image: url('image3.gif')">
  This div has an image background.
</div>
```

Amostra retirada do Arquivo.pt

Falem connosco se quiserem investigar mais sobre estes ou outros dados recolhidos

Dentro do esqueleto

```
<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

Traduzir

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

Onde encontrar imagens no HTML?

- <img>
  - *src* (independentemente da extenção)
  - outros atributos com extensões de imagens (ex. jpg)
- <a>
  - *href* com extensões de imagens
- Fundos CSS
  - *background-url:* com extensões de imagens

Onde estão as imagens no HTML?

- <img>
- <a>
- Fundos CSS
- Imagens base64
- Imagens vindas de JS
- <figure>, <picture>

	Percentagem de referências
Normais	99.9%
base64	0.1%

Grande parte das imagens **não têm metadados** associados

- Descrições **textuais** das imagens, preenchidos aquando criação da página
  - Utilizados em casos de falha de carregamento ou para pessoas com dificuldades visuais

URL (src): https://**imagens publico**.pt/**imagens**.aspx/1440184
Texto alternativo (alt): Marcelo Rebelo de Sousa
Título (title): <vazio>

URL (src): https://**imagens publico**.pt/**imagens**.aspx/044361
Texto alternativo (alt): <vazio>
Título (title): <vazio>

- Em 1 800 milhões de imagens, **45% não têm título (*title*)** ou **texto alternativo (*alt*)**
- No caso da **notícia, só estão preenchidos** na imagem **principal**

Metadados de página **apenas** descrevem a imagem **principal**

Título da página (page title): Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO

URL da página (page url): https://www.publico.pt/2020/02/15/politica/noticia/marcelo-ficou-impressionado-personalidade-politica-modi-190427

- As imagens não podem ser indexadas se não tiverem palavras descritivas
- Como encontrar palavras **descriptivas** para **todas as imagens**?

From page metadata to image metadata

Traduzir

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

Metadata: <img> tag attributes

Traduzir

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;

Metadados HTML para imagens

Metadados da página relevantes para a imagem principal:

- **Título da página:** Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO
- **URL da página:** https www publico pt 2020 02 15 politica noticia marcelo ficou (...)

Metadados das imagens preenchidos apenas na imagem principal:

- **URL da imagem:** imagens publico pt imagens aspx 1440184
- **Texto alternativo da imagem**: Marcelo Rebelo de Sousa

- **URL da imagem**: imagens publico pt imagens aspx 1044361
- **Texto alternativo da imagem**: <vazio>

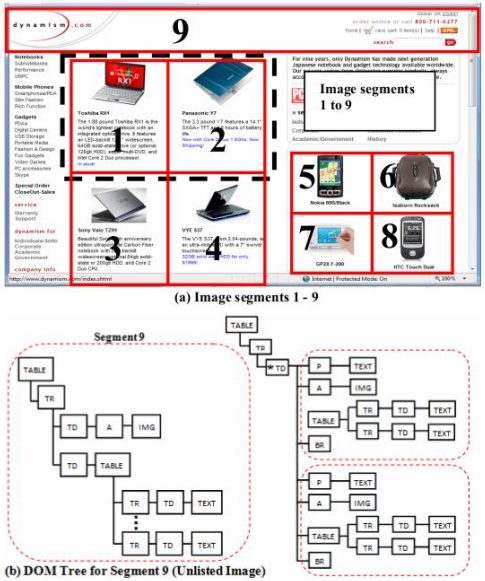
Metadata: <img> tag attributes

Traduzir

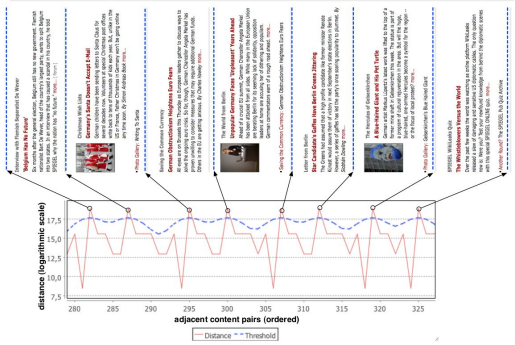
We select all <img> tags in the html and extract the following metadata:

- 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**

Descobrir palavras descritivas para imagens: Estado da Arte



Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information



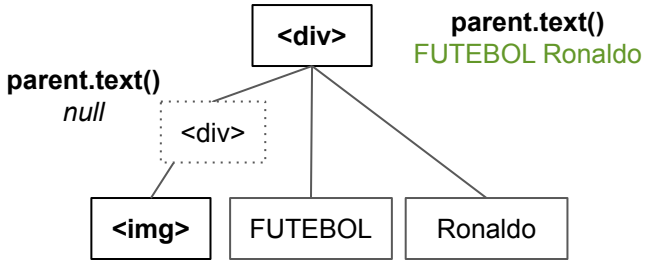
Sadet, Alcic & Conrad, Stefan. (2011). A Clustering-based Approach to Web Image Context Extraction

- Métodos existentes são demasiado complexos para aplicar à escala dos nossos dados
- Precisamos de um método escalável para milhares de milhões de páginas

Encontrando legendas para imagens no Arquivo.pt

Primeiro *parent* com *.text()*

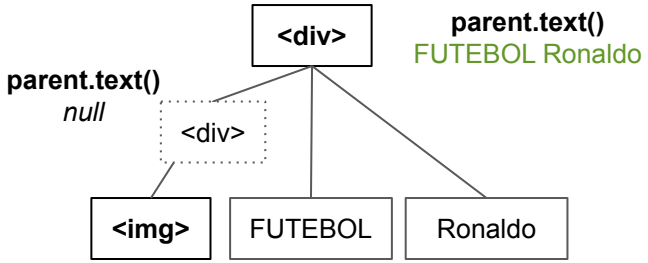
- Método principal
- Funciona bem com imagens dentro de caixas/div e páginas com estruturas *razoáveis*



Encontrando legendas para imagens no Arquivo.pt

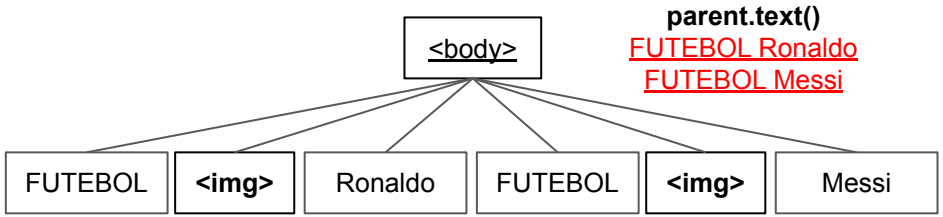
Primeiro *parent* com *.text()*

- Método principal
- Funciona bem com imagens dentro de caixas/div e páginas com estruturas *razoáveis*



.text() dos *siblings*

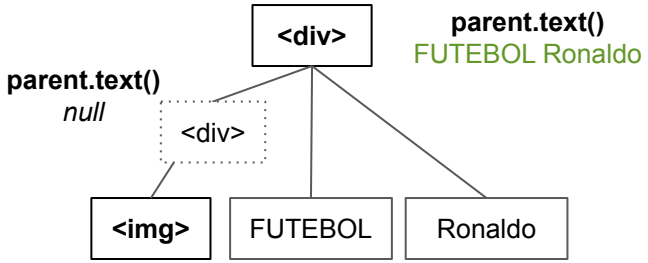
- Usado se o primeiro *parent* com texto estiver num nível com mais *siblings*
- Lista de imagens num blog



Encontrando legendas para imagens no Arquivo.pt

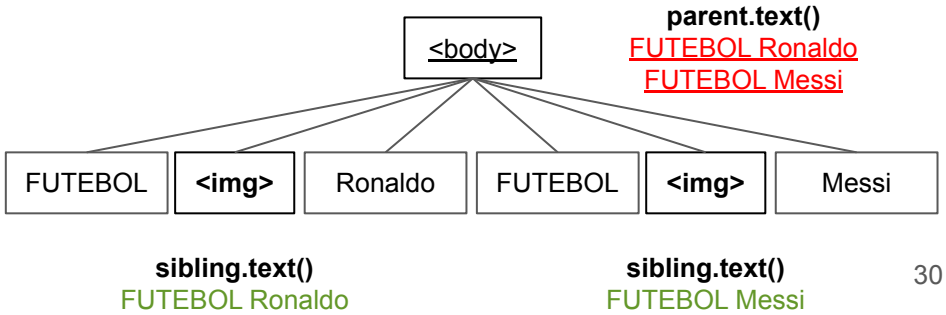
Primeiro *parent* com *.text()*

- Método principal
- Funciona bem com imagens dentro de caixas/div e páginas com estruturas *razoáveis*

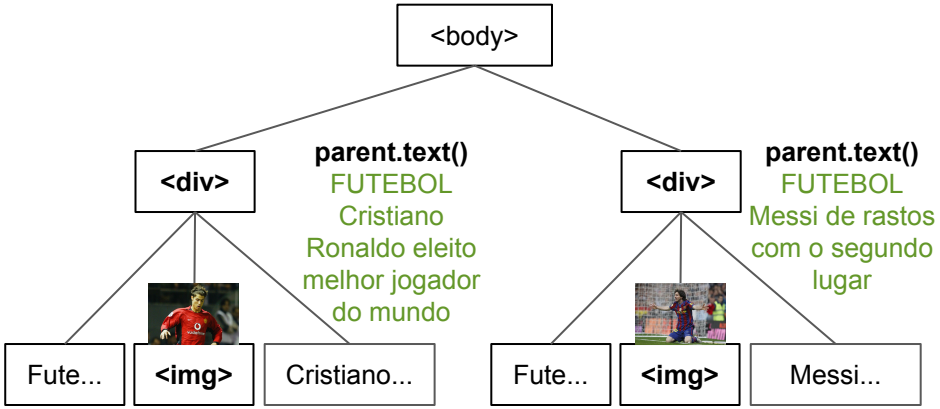


.text() dos *siblings*

- Usado se o primeiro *parent* com texto estiver num nível com mais *siblings*
- Lista de imagens num blog

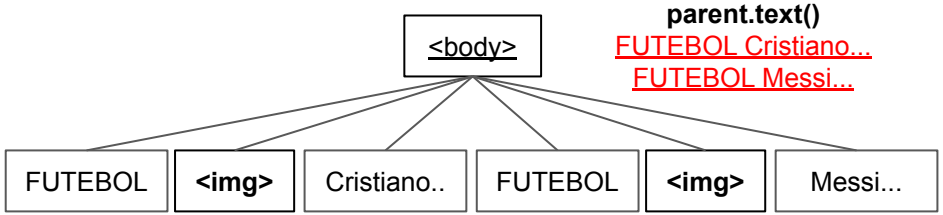


Associar palavras do elemento *parent* do HTML à imagem (legendas)



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

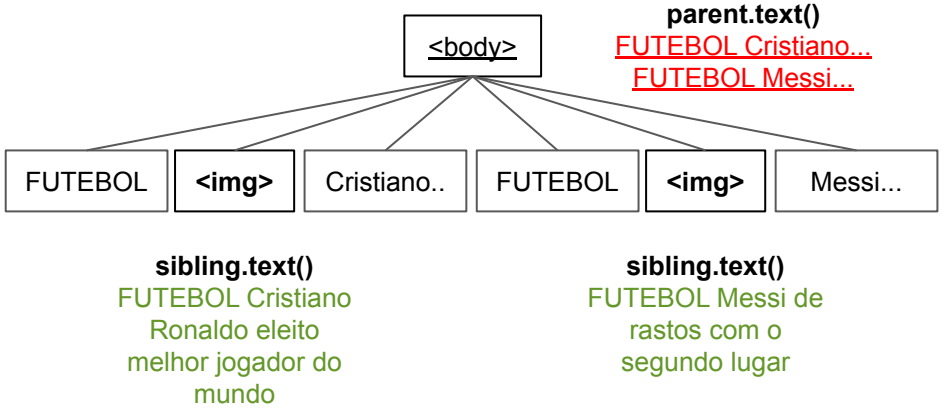
Futebol

Cristiano Ronaldo eleito melhor jogador do mundo

Futebol

Messi de rastos com o segundo lugar

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



- **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

Mais palavras descritivas para as imagens da página

URL (src): https://imagens publico.pt/imagens.aspx/1440184
Texto alternativo (alt): Marcelo Rebelo de Sousa
Título (title): <vazio>
Legenda da imagem (caption): Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

URL (src): https://imagens publico.pt/imagens.aspx/044361
Texto alternativo (alt): <vazio>
Título (title): <vazio>
Legenda da imagem (caption): FUTEBOL Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda

Processar 1 800 milhões de images

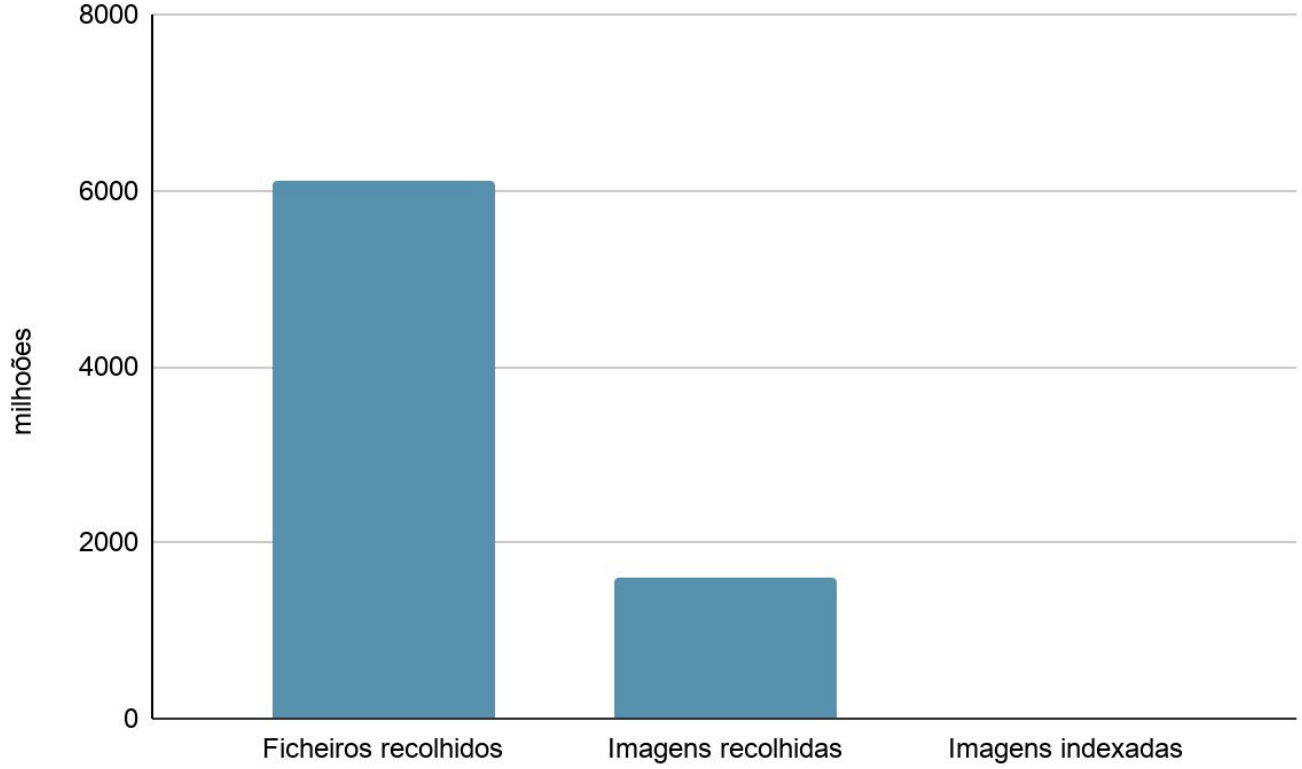
Pesquisa de imagens do Arquivo.pt (Janeiro 2020)

Imagens indexadas	**22 milhões**
Ficheiros recolhidos	6,1 mil milhões
Imagens recolhidas	1,6 mil milhões
Dados arquivados (comprimidos)	334 TB

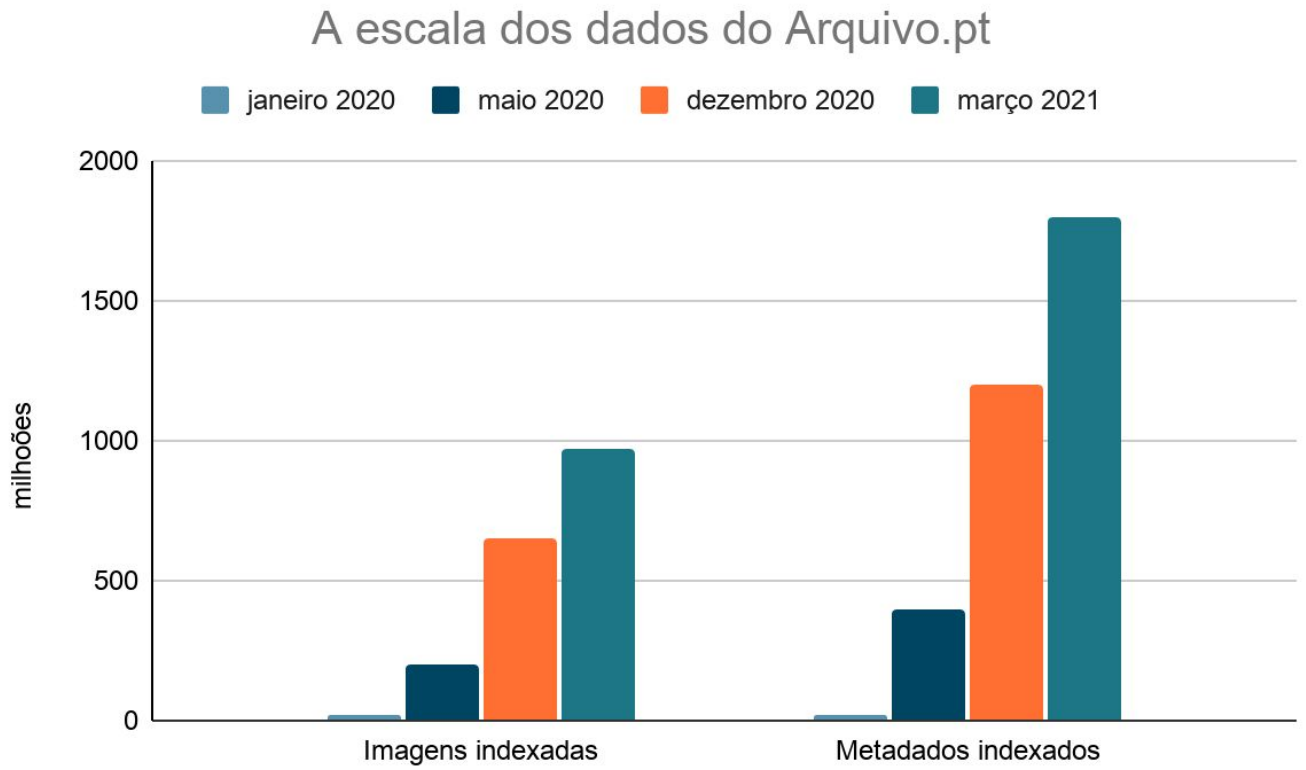
O que mudou em 2020/2021?

Imagens indexadas	**1,8 mil milhões**
Ficheiros recolhidos	8,5 mil milhões
Imagens recolhidas	2,4 mil milhões
Dados arquivados (comprimidos)	520 TB

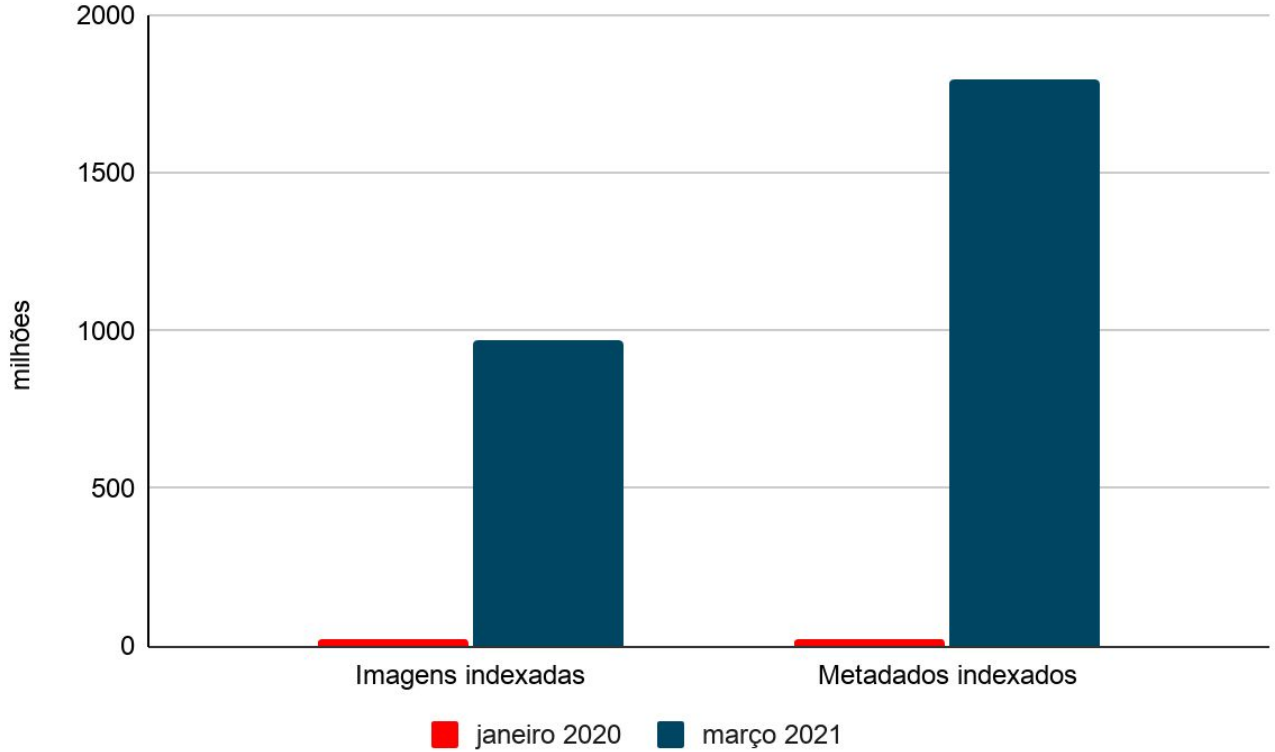
Como estávamos em janeiro de 2020?



O que mudou em 2020/2021?



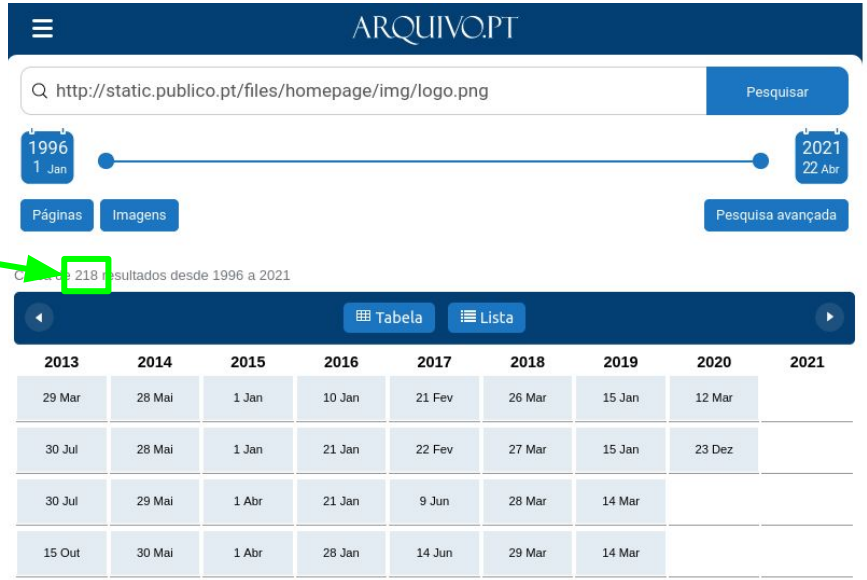
Volume de imagens indexadas: crescimento desde 2020



Como reduzir o volume de informação a indexar sem degradar a pesquisa?

- Pesquisas rápidas requerem índices em memória
- Muita informação gera grandes índices
  - Mas o nosso hardware é limitado
- É necessário reduzir volume de dados a indexar

70% das imagens arquivadas são duplicadas



- Imagens arquivadas repetidamente ao longo do tempo (ex. recolhas diárias)
- Imagens duplicadas dentro de um site (ex. logótipo de um website)
- Imagens duplicadas entre sites (ex. botões de partilha em redes sociais)

Deduplicação: reduzir indexação de imagens duplicadas

- Deduplicação de imagens no Arquivo.pt
  - Agregar imagens duplicadas em vez de indexar todas

- **Solução**: Escolher que versão indexar
  - Escolher como base os metadados da página mais antiga onde a imagem aparece
  - Adicionar todos os metadados de imagem novos das páginas restantes

- **Resultado:** 584 milhões de imagens, com informação de 1 800 milhões
  - Redução de 70% nos dados a indexar

Como estávamos em maio 2020

	Ficheiros recolhidos	Imagens recolhidas	Imagens indexadas	Metadados indexados
janeiro 2020	6 100 M	1 600 M	22 M	22 M
(est.) maio 2020			*300 M*	*600 M*

- Previsão de aumento em 13x de imagens indexadas
- Previsão de aumento em 26x de metadados indexados
  - Cada imagem aparece em média em duas páginas
  - Dobro dos metadados para indexar

Como estávamos em dezembro 2020

	Ficheiros recolhidos	Imagens recolhidas	Imagens indexadas	Metadados indexados
janeiro 2020	6 100 M	1 600 M	22 M	22 M
(est.) maio 2020			*300 M*	*600 M*
dezembro 2020			650 M	1200 M

- Aumento real de 28x em imagens indexadas
- Aumento real de 54x de metadados indexados

Como estamos em março 2021

	Ficheiros recolhidos	Imagens recolhidas	Imagens indexadas	Metadados indexados
janeiro 2020	6 100 M	1 600 M	22 M	22 M
(est.) maio 2020			*300 M*	*600 M*
dezembro 2020			650 M	1200 M
março 2021	8 500 M	2 400 M	970 M	1800 M

- Mas o Arquivo.pt continuou a crescer em 2021
- 43x mais imagens indexadas no total
- 80x mais metadados indexados no total

Opportunities for improvement

Traduzir

- 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

20 para 1800 milhões de imagens... como?

- Encontrar todas as páginas que mencionam uma imagem
- Encontrar imagens em mais atributos e *tags* HTML
- Extrair mais informação das páginas (legendas para as imagens)
- Usar apenas Hadoop: remover dependências desnecessárias (MongoDB)
- Passagem para Solr distribuído (Solr Cloud)
- Usar tipos certos de dados no Solr
- Extrair metadados das páginas e imagens
- ...

20 para 1 800 milhões de imagens... como?

- Processar 520TB de dados num *cluster* distribuído
- Encontrar metadados relevantes para imagens
- Tornar estes metadados pesquisáveis

Processar 520TB de dados

Takeways

200 milhões de imagens

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

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

9x mais ficheiros de imagens

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	[illegible]	27.65x

400 milhões de metadados

18x mais metadados de imagens

Takeways

~200-650 milhões de imagens

~9-28x mais ficheiros de imagens

~0,4-1,3 milhões de metadados

~18-56x mais metadados de imagens

Takeways

654 milhões de imagens

29x mais ficheiros de imagens

1,2 mil milhões de metadados

55x mais metadados de imagens

Mas o Arquivo.pt continuou
a crescer em 2020

Takeways

+ 317 milhões de imagens num ano (2019)

48% de crescimento

+ 610 milhões de metadados num ano

49% de crescimento

Takeways

971 milhões de imagens

42x mais imagens

1,8 mil milhões de metadados

81x mais metadados de imagens

Impacto da deduplicação

	Número de documentos
a	1,8 mil milhões de documentos imagens-páginas pares
b	584 milhões de documentos únicos (971 milhões antes de deduplicação temporal)
c	**584 milhões** de documentos, com informação de **1,8 mil milhões** de imagens-páginas

Indexar imagens no SolrCloud

Arquivos (W)ARC

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/(...)


```
<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
```

(W)ARCs, páginas HTML e imagens

Quando recolhemos imagens e páginas estas são colocadas numa fila para processamento

No melhor caso, uma página e as suas imagens seriam arquivadas juntas no mesmo ficheiro (W)ARC, mas...

- Podem estar em (W)ARCs diferentes
- Algumas imagens referenciadas nas páginas podem não ter sido capturadas
- Múltiplas versões da mesma imagem e página
- URL pode ser inconsistente entre a imagem e página (URLs relativos)
- A mesma imagem aparece em mais do que uma página
  - Mesmo URL
  - URL diferente, *digest* igual
- Imagens com o mesmo URL podem ter mudar ao longo do tempo (mesmo URL, *digest* diferente)

Fluxo de processamento

- Processamento distribuído Hadoop
  - Encontrar imagens e os seus metadados nos dados arquivados
  - Deduplicar as imagens
- Classificação NSFW
  - Identificar de conteúdo potencialmente ofensivo para utilizadores
- Indexação SolrCloud
  - Tornar 584 milhões de imagens pesquisáveis em menos de um segundo

Metadata: <a> tag attributes (new)

Traduzir

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)

Traduzir

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

Traduzir

- 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

Questões futuras para os metadados

Quantas imagens têm metadados explícitos?

- Métricas actuais mostram que **99%** das imagens têm metadados (*imgCaption*)

Mas como podemos medir a qualidade dos mesmos?

- IMG_00123.jpg não ajuda muito a pesquisa

O que fazer quando estes não existem?

- Técnicas *deep learning* de geração de legendas ou classificação

Traduzir

Examine the quality of the metadata

Encodings on the internet

Traduzir

“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”

Traduzir

```
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;
}
```

Traduzir

```
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;
}
```

Traduzir

“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

Traduzir

Detector Mozilla: 160524

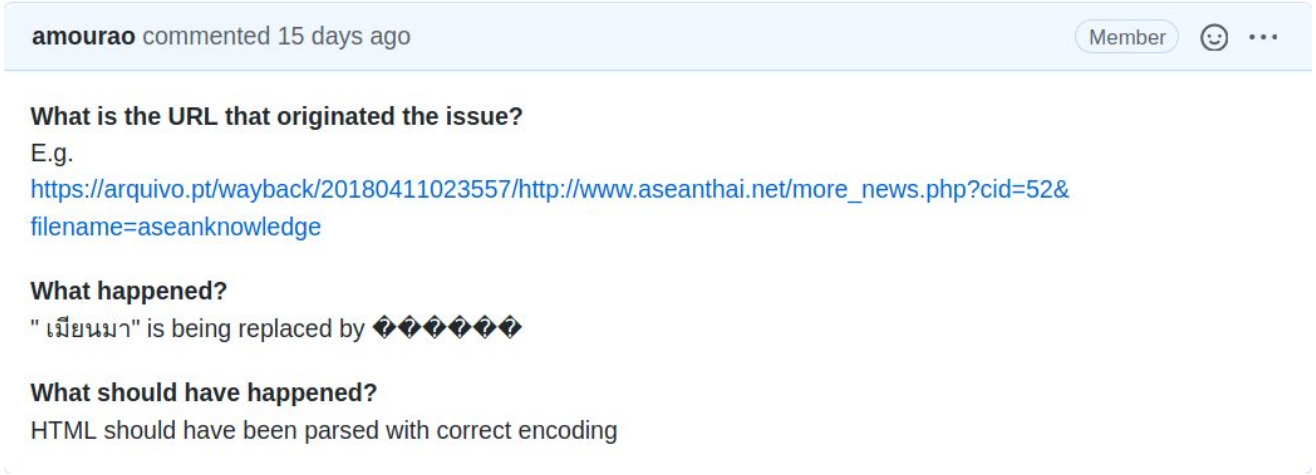
Detector Tika: 366202

Detector Mozilla + meu fix: 9665

But this does not solve all encoding issues

Traduzir

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



amourao commented 15 days ago — Member

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 ??????

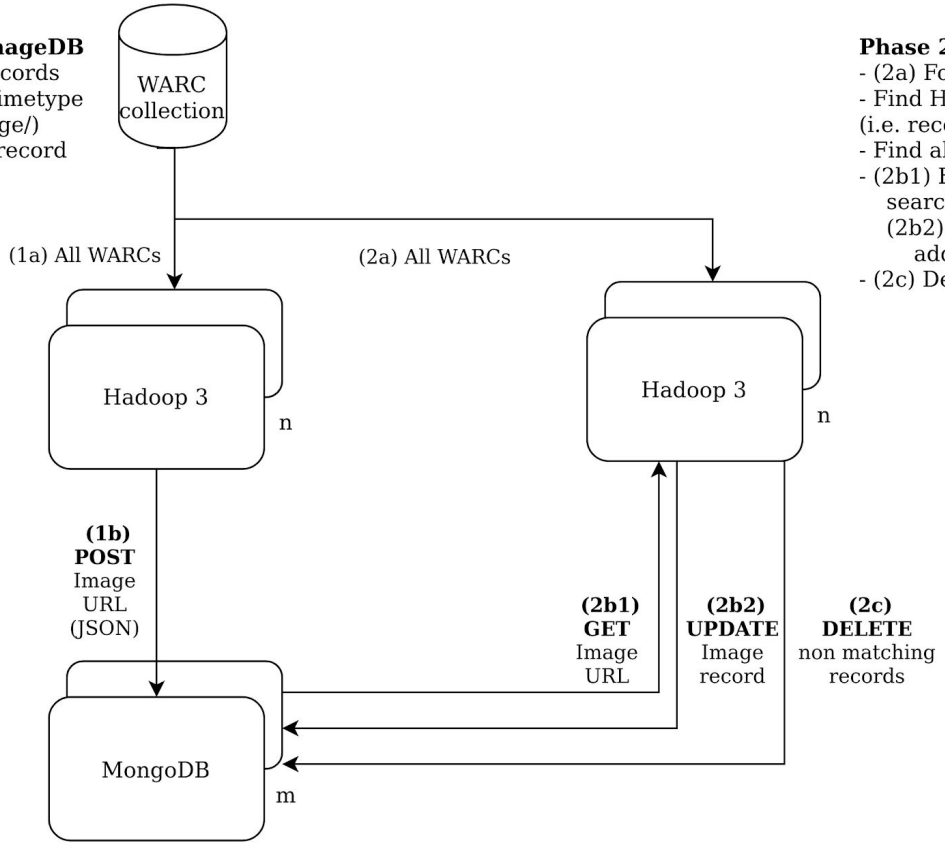
What should have happened?
HTML should have been parsed with correct encoding

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

Old Map Reduce

Traduzir

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

Traduzir

- (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

Traduzir

- 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

Extrair metadados de imagens

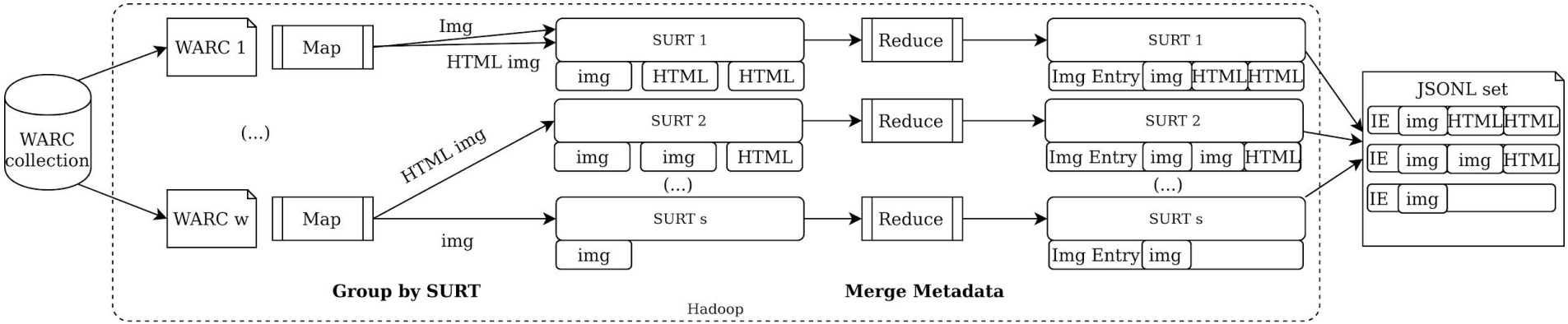
Group by SURT (*URL* normalizado):

- Para cada ficheiro de recolhido
  - Se imagem, extrair metadados (*height*, *width*, *digest*) e colocar na entrada do seu *SURT*
  - Se página, encontrar todas as entradas *<img>/<a>/css*, extrair metadados e colocar na entrada do *SURT* da imagem

Merge metadata

- Para cada entrada em cada SURT, agrupar resultados de páginas e imagens e voltar a agrupar por *digest*
  - Adicionar *img title*, *alt,* e *captions* das várias páginas onde a imagem aparece

Map Reduce: Extrair imagens e metadados



What users want in image search?

Traduzir

- 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

Imagem que aparece em duas páginas

Documento existente:

"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"],

Novo documento:

"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"],

Imagem que aparece em duas páginas

Documento existente:

"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"],

Novo documento:

"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"],

Imagem que aparece em duas páginas

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"]],

Soluções para deduplicação

- Depois de uma análise detalhada, chegámos a três cenários para dedup:
  a. a página mais antiga que referencia a imagem é o documento
  b. cada par página-imagem é um documento para pesquisa
  c. a página mais antiga juntamente com metadados de imagem de todas as páginas
     - escolher a página mais antiga onde a imagem
     - adicionar toda a informação específica da imagem (*title*, *alt*, *caption*)

Group by digest

Traduzir

Group by Digest

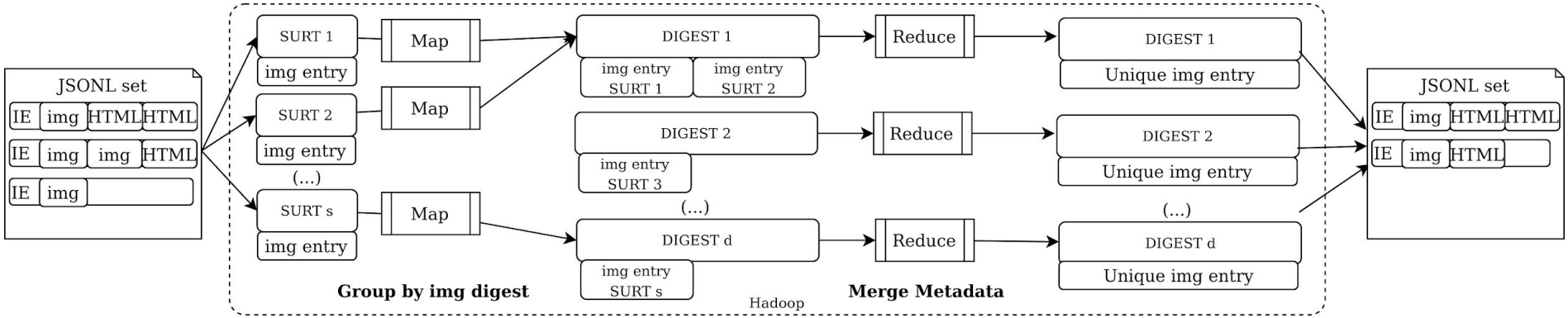
- For each image record in the JSONL, send it to the matching Digest list

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

Traduzir



Duplicates across collections

Traduzir

- 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

Traduzir

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

Traduzir

- 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

Traduzir

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

Traduzir

- **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

As previsões de Maio de 2020

Pesquisa de imagens do Arquivo.pt (Jan 2020)

Imagens indexadas	**22 milhões**
Imagens indexadas (sem duplicados)	**18 milhões**
Coleções indexadas	90
(W)ARCs	3 milhões
Tamanho dos (W)ARC	334 TB
Ficheiros recolhidos (total)	6 000 milhões
Imagens recolhidos (total)	1 602 milhões
Imagem mais antiga	15/04/1994
Imagem mais recente	14/11/2019

Teste do novo sistema

Colecção	Antigo	Novo	Diferença	Racio
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

Current Solr indexing architecture

Traduzir

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

Como indexar as imagens?

Processamento resultou em

584 milhões

de imagens para indexar

Onde colocar para permitir uma pesquisa rápida?

Arquivo.pt response time guidelines

Traduzir

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

Alocação de recursos para SolrCloud

- Tamanho de índice esperado: **~720GB**

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

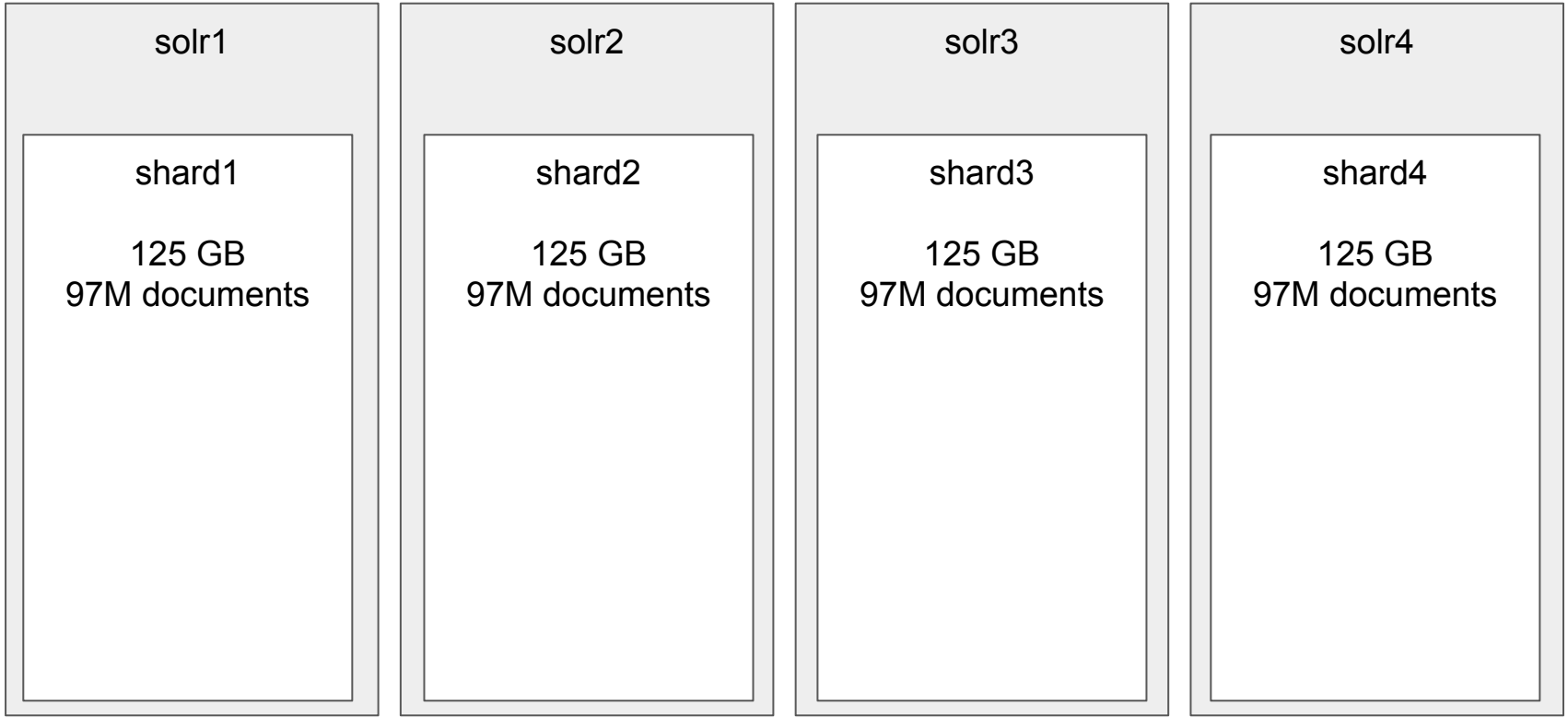
- Sem *SSDs*, só *HDD*, mas felizmente temos RAM suficiente para o índice

Solr e SolrCloud

- Plataforma de pesquisa de texto e metadados
  - “Irmão” do ElasticSearch, ambos baseados no Apache Lucene
  - Processa e indexa documentos estruturados para pesquisas eficiente
  - Encontra e ordena documentos que contenham os termos das *queries*
- Desenhado para ser rápido e distribuído em vários servidores (SolrCloud)
  - 4 servidores (256-512GB), com 146M imagens cada
  - Tempo de resposta médio inferior a 500 ms com 4 utilizadore

How we configured SolrCloud? - Try 1

Traduzir



Solr performance factors

Traduzir

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

....

Solr performance factors

Traduzir

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

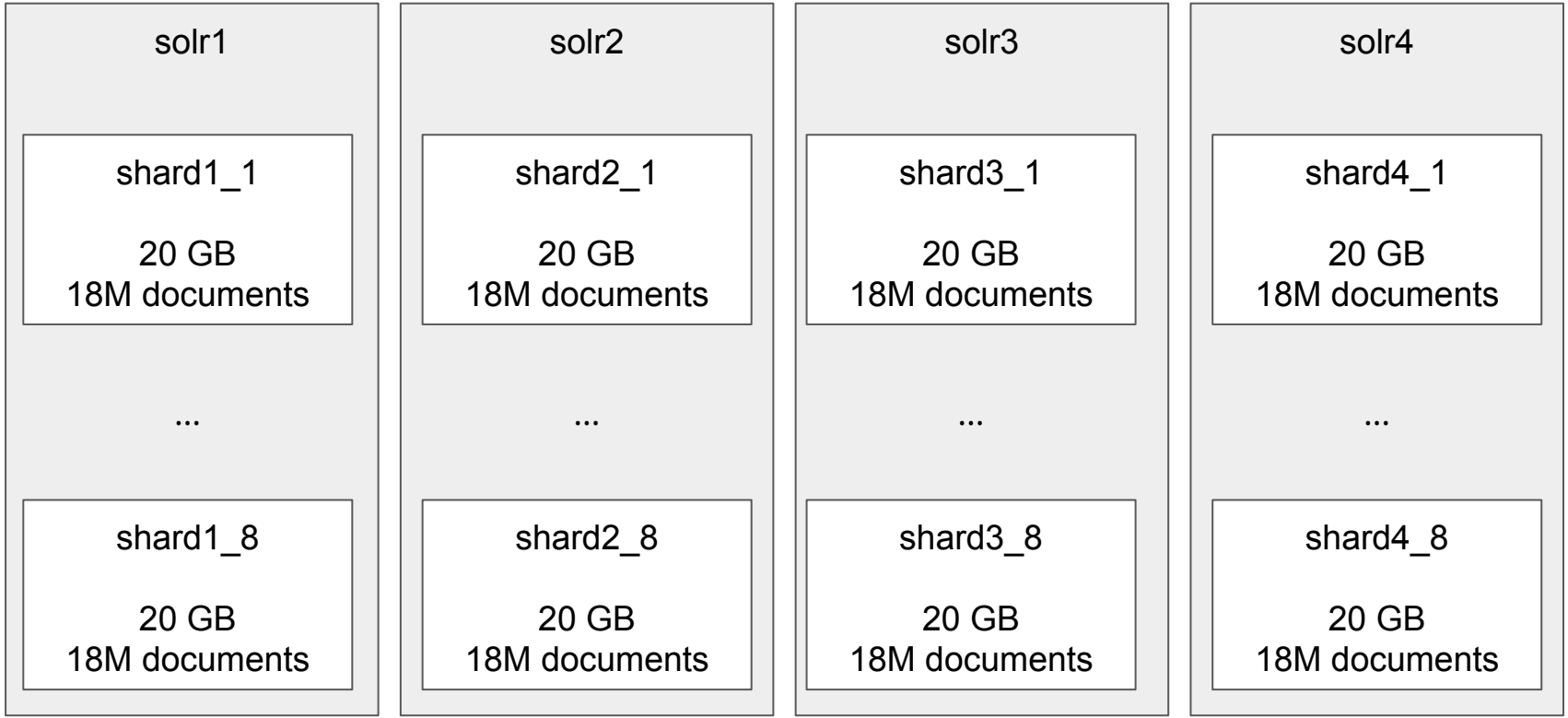
  ....

  disk I/O skyrockets

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

How we configured SolrCloud? - Try 2

Traduzir



How to test?

Traduzir

- 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?

Traduzir

- 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

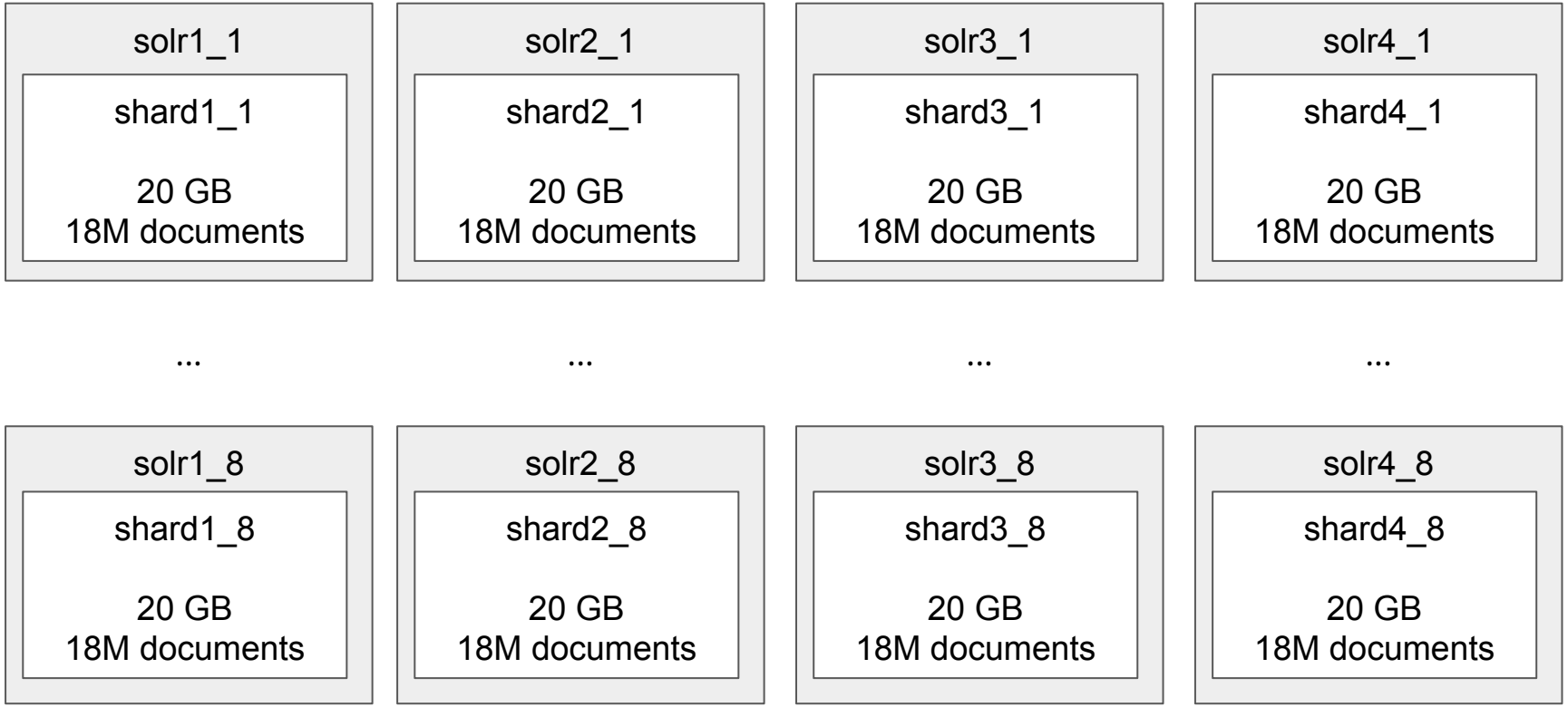
Tips and parameters

Traduzir

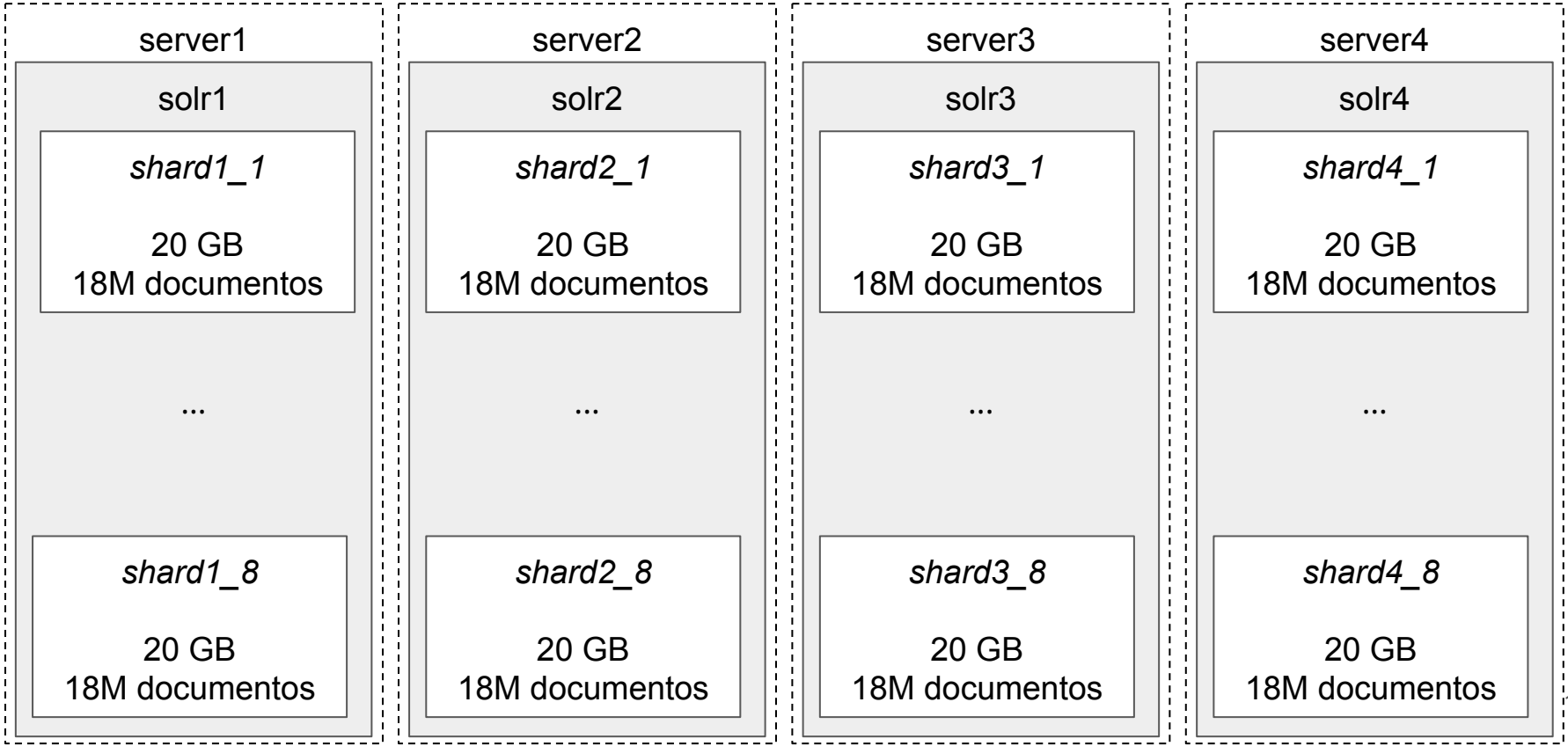
- vmtouch tool to force OS to keep index files in RAM
- Heap size: 31GB
  - Smaller sizes made Solr crash on parallel query situations
  - Larger sizes means Java can't use compressed pointers https://lucene.apache.org/solr/guide/8_7/taking-solr-to-production.html#running-multiple-solr-nodes-per-host

How we configured SolrCloud?

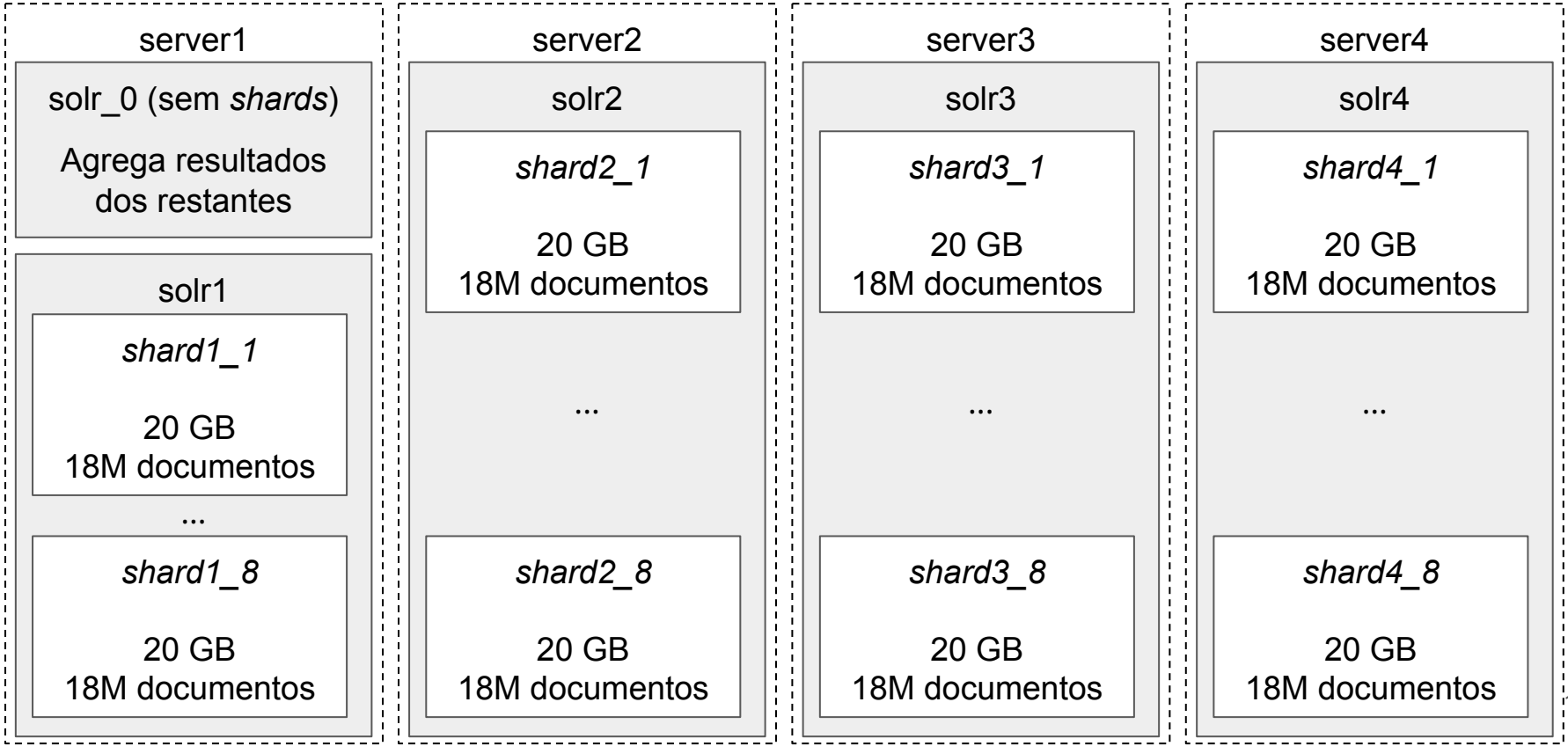
Traduzir



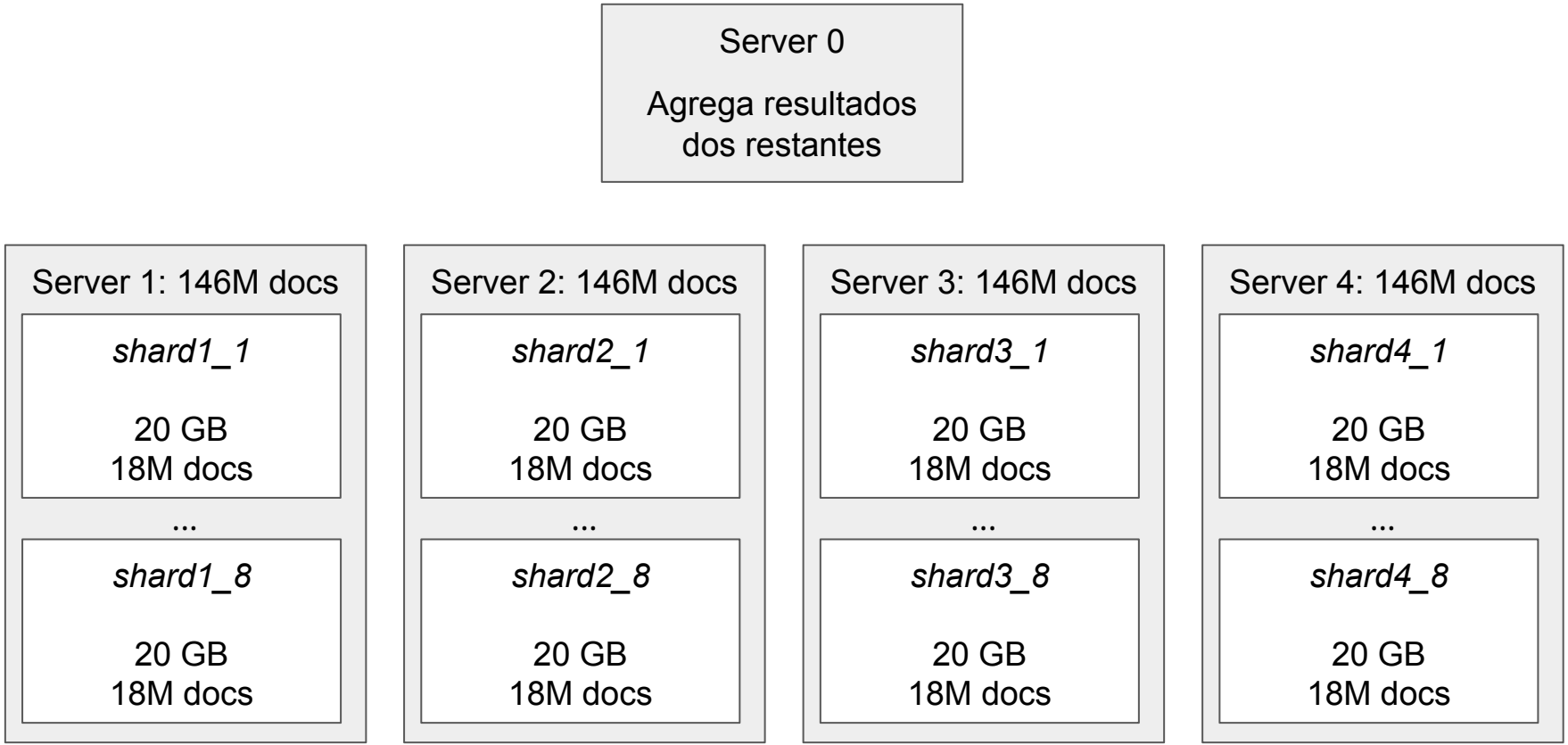
Como configurámos o SolrCloud?



Como configurámos o SolrCloud?



Como configurámos o SolrCloud?



Tempo de resposta/testes de carga

Pedidos em paralelo	Média	Mediana	Percentil 95%	Percentil 99%	*Throughput*
1	115 ms	74 ms	235 ms	769 ms	8 queries/seg
3	120 ms	76 ms	259 ms	872 ms	24 queries/seg
5	136 ms	85 ms	304 ms	1059 ms	36 queries/seg
10	211 ms	128 ms	501 ms	1718 ms	46 queries/seg
25	489 ms	266 ms	1297 ms	4334 ms	50 queries/seg
50	970 ms	593 ms	2694 ms	6699 ms	50 queries/seg

Pesquisas recolhidas de *logs* de utilização

Tempo de resposta pesquisa de imagens

Pedidos em paralelo	Média	Mediana	Percentil 95%	Percentil 99%	*Throughput*
1	115 ms	74 ms	235 ms	769 ms	8 queries/seg
5	136 ms	85 ms	304 ms	1059 ms	36 queries/seg
10	211 ms	128 ms	501 ms	1718 ms	46 queries/seg
50	970 ms	593 ms	2694 ms	6699 ms	50 queries/seg

Pesquisas recolhidas de *logs* de utilização

Future problems: Migrate page search to SolrC

Traduzir

- 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

Modernização de bibliotecas para classificação NSFW

Conteúdo *Not Safe for Work* (*NSFW*)?

- Arquivo.pt captura páginas e imagens de **toda** a internet
- Isto pode conter conteúdo ofensivo para os utilizadores
- No Arquivo.pt utilizamos um classificador para detectar este conteúdo em imagens
- Imagens marcadas como NSFW são filtradas dos resultados de pesquisa

Classificador NSFW do Arquivo.pt's

- Baseado em *ResNet* e implementado no *Tensorflow/Keras*
  - https://github.com/GantMan/nsfw_model
- Reportam **93% de precisão** de classificação
  - Nos nossos testes, medimos uma **precisão de 90%**
- ~500 imagens por segundo utilizado **duas NVIDIA Tesla P4**

NSFW

Traduzir

- 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

Traduzir

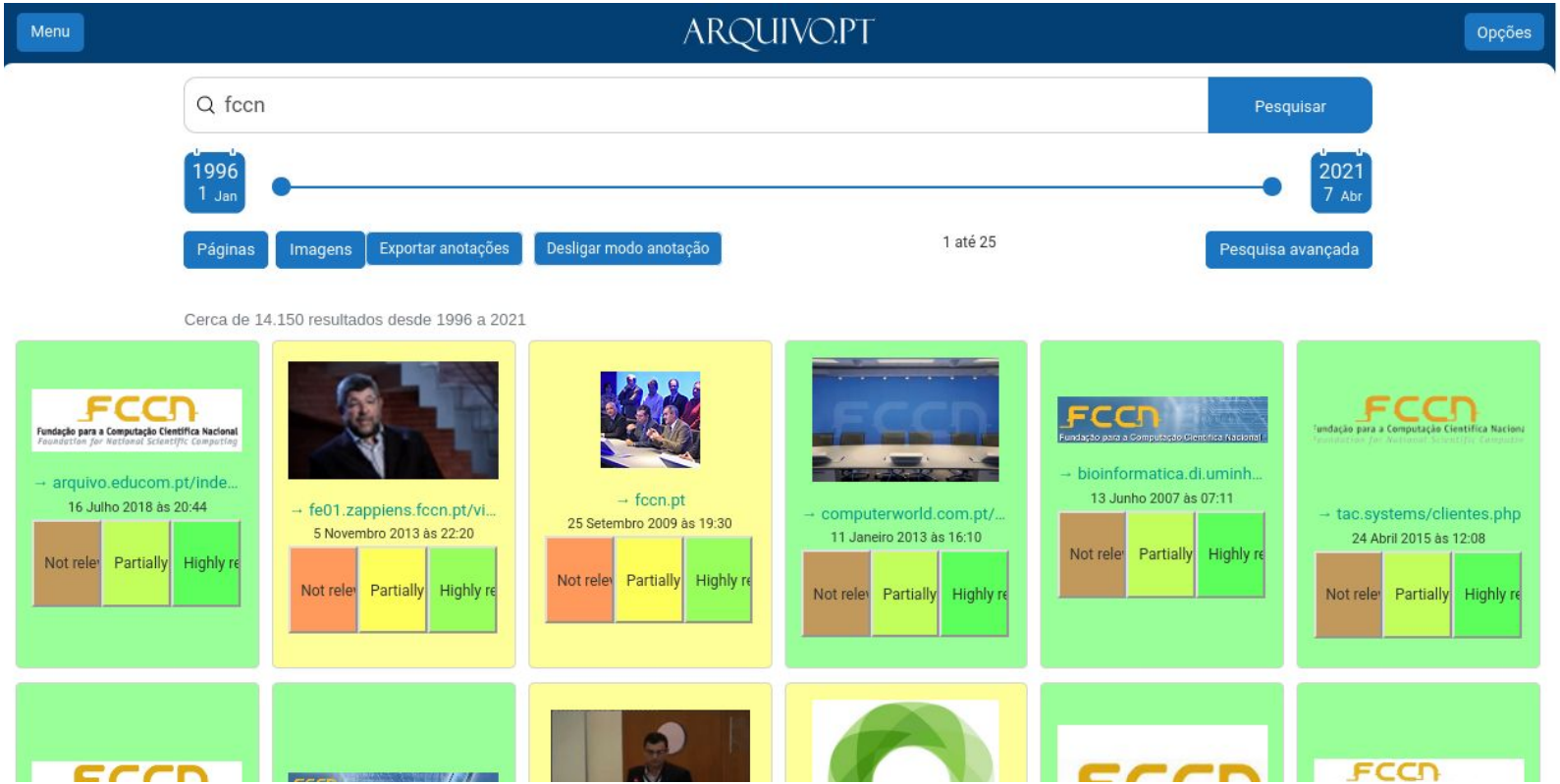
Architecture/Pipeline

Traduzir

Relevance assessment

Annotator

Traduzir



Results on TestCollection (2020)

Metric	Arquiv
mAP	0.5
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

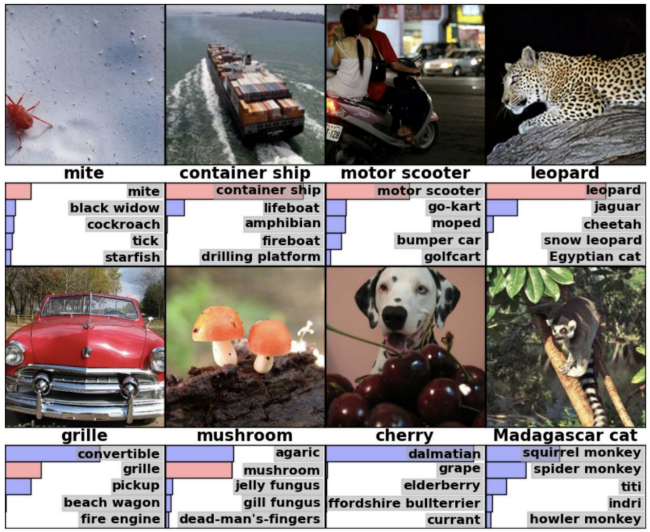
Traduzir

Melhorias da nova pesquisa de imagens

- **Mais** imagens e metadados
  - Todas as páginas onde a imagem aparece são processadas
  - Extração heurística de legendas de imagens a partir da estrutura do HTML
- **Melhorada arquitectura** de indexação
  - Indexadas imagens de <img>, links <a> e CSS
- Melhorado processamento de sistema de **classificação NSFW**
  - 7x mais rápido (80 -> 500 imagens por segundo)
- Pesquisa **distribuída**
  - Transição para uma arquitectura SolrCloud distribuída
  - 4 servidores (com 512GB de RAM cada), com 146M imagens cada
  - Tempo de resposta médio inferior a 500 ms com 4 utilizadores em paralelo

Planos para o futuro

- Imagens **sem metadata**
  - **300+ milhões** de imagens sem texto associado
  - Legendagem baseada em redes neuronais
- Imagens **semelhantes**
  - Mesma imagem, resoluções e/ou formatos diferentes
  - Fazer deduplicação de ***near duplicates***
- Melhorar **ordenação dos resultados**
  - Construir coleção de teste para avaliar sistema actual



2020 vs 2021

Janeiro 2020	**Janeiro 2021**	**Melhoramento**
22 milhões de imagens em páginas (apenas uma versão de cada coleção é indexada)	1,862 mil milhões de imagens	**81x** mais imagens em páginas analisadas
	967 milhões de imagens	**42x** mais imagens analisadas
17 milhões de documentos de-duplicados	584 milhões de documentos de-duplicados	**33x** mais imagens únicas, removendo duplicados entre coleções
49% têm metadados (*imgAlt*, *imgTitle*)	99%+ têm metadados (*imgAlt*, *imgTitle*, *imgCaption*)	**+51 p.p.** imagens com informação contextual relevante
~570 GB índice de pesquisa	~750 GB índice de pesquisa	**Apenas 32% mais** após um **aumento de 813%** em informação analisada
1 servidor Solr	4 servidores SolrCloud	**Apenas mais 3 servidores necessários** após **aumento de 81x** em info analisada

2020 vs 2021

Janeiro 2020	**Janeiro 2021**	**Melhoramento**
22,881,688 de imagens em páginas (apenas uma versão de cada coleção é indexada)	1,862,311,456 milhões de imagens	**81.39x** mais imagens em páginas analisadas
	967,184,126 milhões de imagens	**42.26x** mais imagens analisadas
17,643,047 de documentos de-duplicados	584,242,176 milhões de documentos de-duplicados	**33.11x** mais imagens únicas, removendo duplicados entre coleções
48.7% têm metadados (*imgAlt*, *imgTitle*)	99.6% têm metadados (*imgAlt*, *imgTitle*, *imgCaption*)	**+50.9 p.p.** imagens com informação contextual relevante
~570 GB índice de pesquisa	~750 GB índice de pesquisa	**Apenas 32% mais** após um **aumento de 813%** em informação analisada
1 servidor Solr	4 servidores SolrCloud	**Apenas mais 3 servidores necessários** após **aumento de 81x** em info analisada

Ranking features for 2021

Traduzir

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

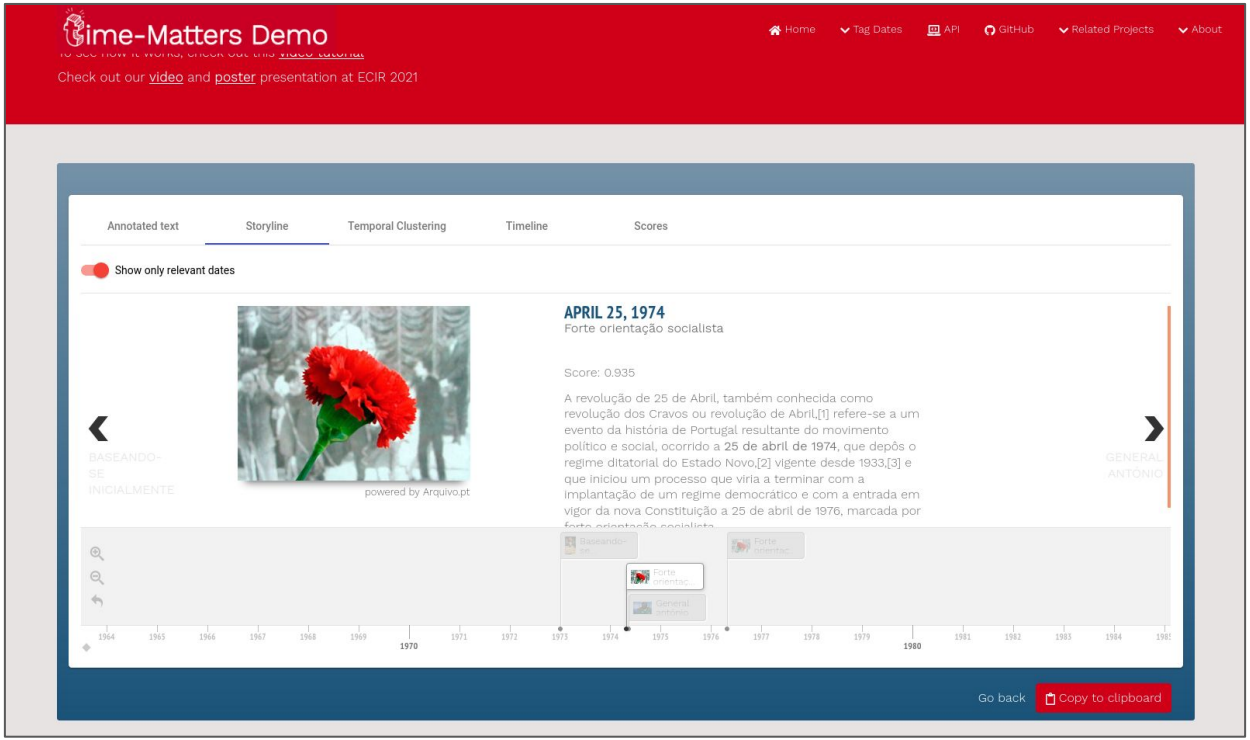
drawing/photo
- whether the image is a drawing or a photo

Arquivo.pt em acesso aberto!

- **Pesquise** 1 800 milhões de imagens e 8 000 milhões de páginas
- **Código Aberto**: github.com/arquivo/
- **APIs**: arquivo.pt/api
  - **arquivo.pt/api/imagesearch**
- Contactos
  - contacto@arquivo.pt
  - github.com/arquivo/pwa-technologies/issues

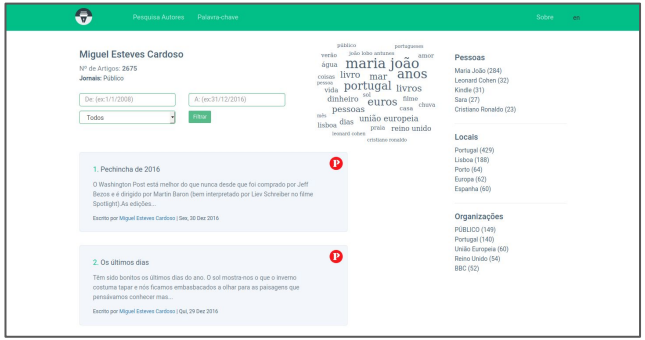
Fundação para a Ciência e a Tecnologia

Casos de uso (API pesquisa de imagens)

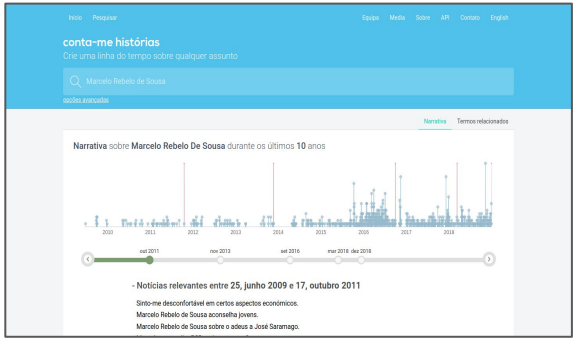


Time Matters: http://time-matters.inesctec.pt/

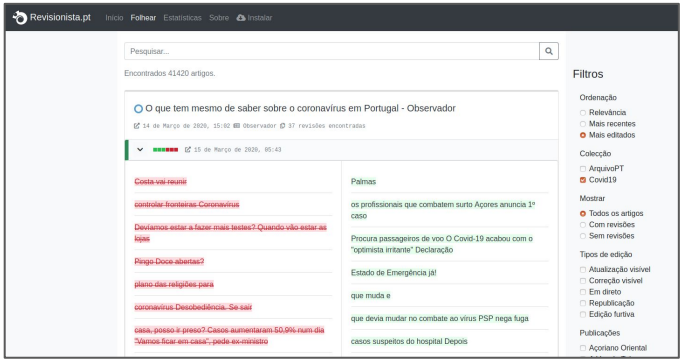
Casos de uso (outras APIs)



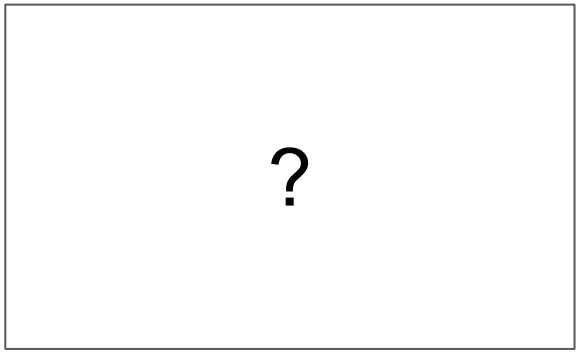
Arquivo de opinião: http://arquivodeopiniao.pt/pt/



Conta-me Histórias: http://contamehistorias.pt/arquivopt/



Revisionista.PT: https://revisionista.pt



<insira o seu projecto aqui>

Prémio Arquivo.pt 2021

- **Desafio**: Utilize a API de pesquisa de imagens para ganhar 10 000 euros