

Pesquisando milhões de imagens no Arquivo.pt

23 de abril de 2021

André Mourão, Ph.D.

Investigação & Desenvolvimento, Arquivo.pt

andre.mourao@fccn.pt

Uma página da Internet como tantas outras



The image shows a screenshot of a news article from the Portuguese website Público. The article is titled "Marcelo ficou 'muito impressionado com a personalidade política' de Modi" and is categorized under "DIPLOMACIA". The author is identified as Lusa, and the date is February 15, 2020. The main image shows Marcelo Rebelo de Sousa speaking at a podium during a press conference. The background of the podium features the text "India - Portugal Business Council" and "with Mr. Marcelo Rebelo de Sousa, President of the Republic of Portugal". Below the image, there is a caption: "Marcelo Rebelo de Sousa LUSA/ESTELA SILVA". The article text states that the President of the Republic, Marcelo Rebelo de Sousa, declared on Saturday that he was "very impressed with the political personality" of Indian Prime Minister Narendra Modi, and with his commitment to strengthening Luso-Indian relations. To the right of the article, there are social media sharing options (Facebook, Twitter, LinkedIn, etc.) and a section titled "MAIS POPULARES" (More Popular) featuring three other articles: "Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda", "FUTEBOL Tribunal aceita que se possa insultar no futebol", and "ARQUITECTURA A renovação deste apartamento é uma viagem à Lisboa do passado".

PÚBLICO 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

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

India - Portugal Business Council

with Mr. Marcelo Rebelo de Sousa, President of the Republic of Portugal

ary 2020

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.

Imagens como parte significativa de uma página

Imagens

The image shows a screenshot of a news article on the website Público. The article title is "Marcelo ficou 'muito impressionado com a personalidade política' de Modi". The main image shows Marcelo Rebelo de Sousa speaking at a podium. A large green box highlights the main image. A smaller green box highlights the social media sharing icons. Another green box highlights the 'MAIS POPULARES' section. A green line points from the word 'Imagens' on the left to the main image. Another green line points from the word 'Imagens' on the right to the social media icons. A third green line points from the word 'Imagem' on the left to the main image. A fourth green line points from the word 'Imagens' on the right to the 'MAIS POPULARES' section.

PÚBLICO

P2 ÍPSILON IMPAR 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

India - Portugal Bu
with Mr. Marcelo
ment of the
ary 2020

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

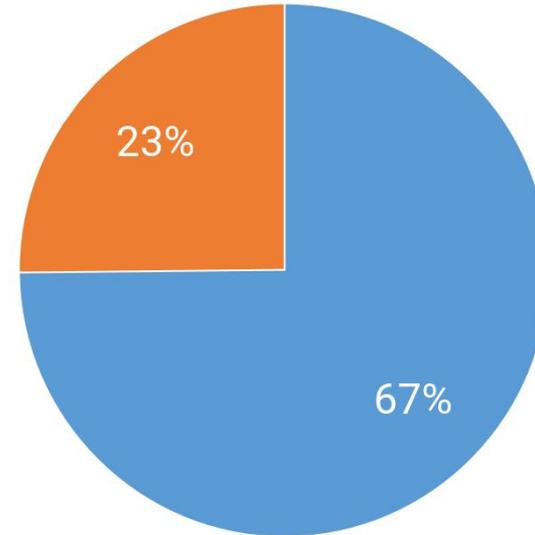
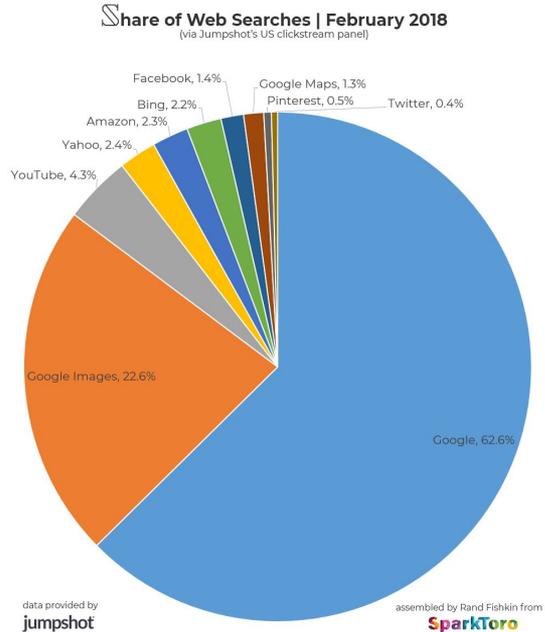
- Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda
- FUTEBOL Tribunal aceita que se possa insultar no futebol
- ARQUITECTURA A renovação deste apartamento é uma viagem à Lisboa do passado

Imagens

Imagens

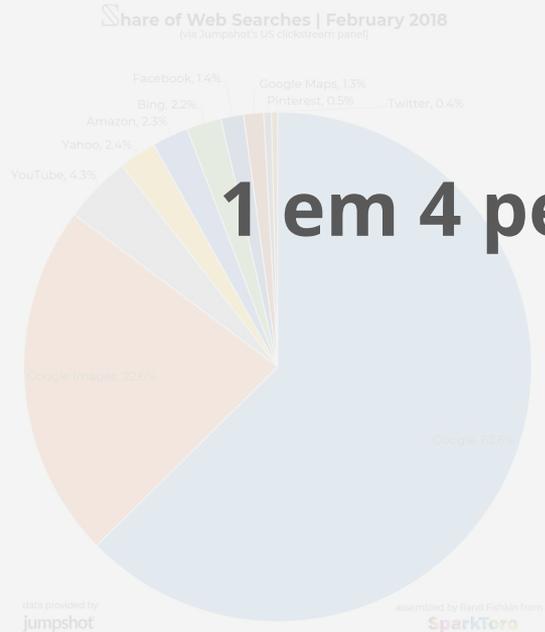
Imagem

A pesquisa de imagens é importante!

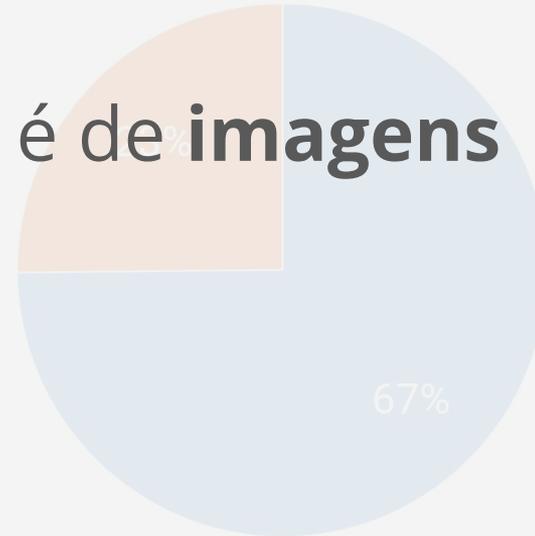


● Pesquisa de páginas ● Pesquisa de imagens

A pesquisa de imagens é importante!



1 em 4 pesquisas é de imagens



● Pesquisa de páginas ● Pesquisa de imagens

Pesquisa de imagens do Arquivo.pt

Menu Opções

ARQUIVO.PT

Q cristiano ronaldo Pesquisar

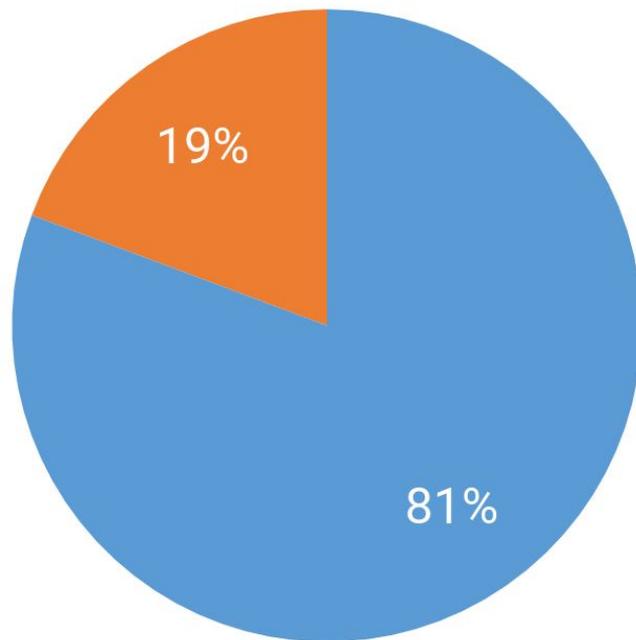
1996 1 Jan 2021 22 Abr

Páginas Imagens Pesquisa avançada

Cerca de 1.020.278 resultados desde 1996 a 2021

 <p>→ livefutbol.com 27 Março 2019 às 13:53</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Janeiro 2011 às 18:17</p>	 <p>→ teknomatika.blogspot.co... 22 Julho 2018 às 05:05</p>	 <p>→ aeiou.caras.pt/mae-do-fil... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 6 Agosto 2011 às 18:06</p>	 <p>→ calcio.com 19 Julho 2018 às 03:22</p>
 <p>→ gazzettadelsud.it/foto/cu... 30 Outubro 2018 às 10:53</p>	 <p>→ desporto.sapo.mz/mais... 22 Julho 2018 às 05:05</p>	 <p>→ aeiou.caras.pt/cristiano-r... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Julho 2010 às 17:16</p>	

Distribuição da pesquisa de imagens no Arquivo.pt



● Pesquisa de páginas ● Pesquisa de imagens

Distribuição da pesquisa de imagens no Arquivo.pt

**1 em 5 pesquisas no Arquivo.pt
é de **imagens****

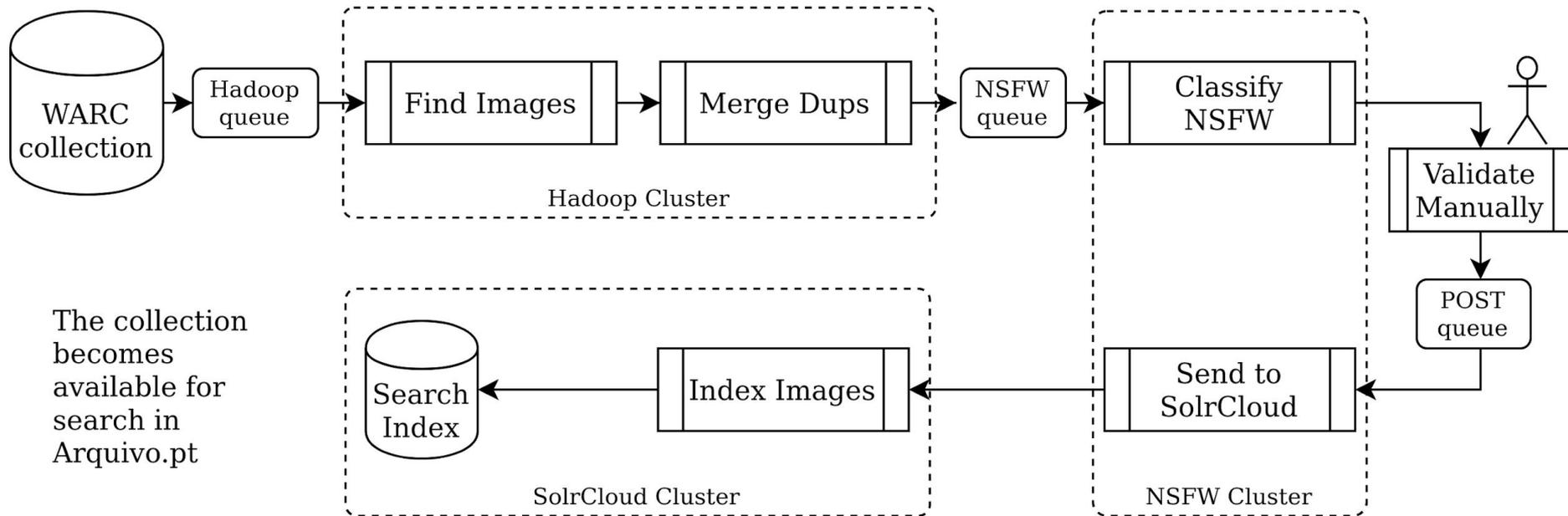


● Pesquisa de páginas ● Pesquisa de imagens

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

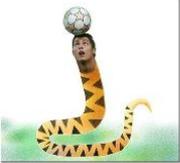
Menu ARQUIVO.PT Opções

Q cristiano ronaldo Pesquisar

1996 1 Jan 2021 22 Abr

Páginas Imagens Pesquisa avançada

Cerca de 1.020.278 resultados desde 1996 a 2021

 <p>→ livefutbol.com 27 Março 2019 às 13:53</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Janeiro 2011 às 18:17</p>	 <p>→ teknomatika.blogspot.co... 22 Julho 2018 às 05:05</p>	 <p>→ aeiou.caras.pt/mae-do-fil... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 6 Agosto 2011 às 18:06</p>	 <p>→ calcio.com 19 Julho 2018 às 03:22</p>
 <p>→ gazzettadelsud.it/foto/cu... 30 Outubro 2018 às 10:53</p>	 <p>→ desporto.sapo.mz/mais... 12 Maio 2011 às 18:28</p>	 <p>→ aeiou.caras.pt/cristiano-r... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Julho 2010 às 17:16</p>	 <p>→ aeiou.caras.pt/gen.pl?ski... 19 Julho 2010 às 17:16</p>	

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

36 PARTILHAS

MAIS POPULARES

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

FUTEBLO Tribunal aceita que se possa insultar no futebol

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

Marcelo Rebelo de Sousa LUSA/ESTELA SILVA

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

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.

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
 - Imagem de arquivo de treino de futebol
 - Renovação de um apartamento
- Outras:
 - Ligações a imagens externas
 - Imagens como fundo CSS

A anatomia de uma página Web

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

India - Portugal Bu
with Mr. Marcelo
President of the
ary 2020

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.

36 PARTILHAS

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

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
 - Imagem de arquivo de treino de futebol
 - Renovação de um apartamento
- Outras:
 - Ligações a imagens externas
 - Imagens como fundo CSS

Maioria 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 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>

- Em 1 800 milhões de imagens, **~50% não têm título** ou **texto alternativo**
- No caso da **notícia**, **só estão preenchidos** na imagem **principal**

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

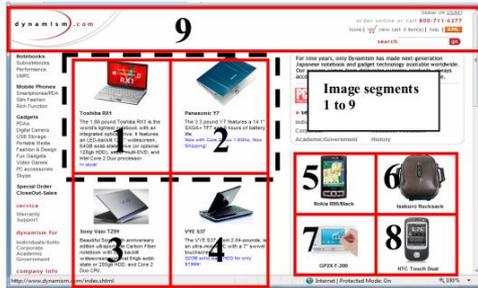


Título da página: Marcelo ficou “muito impressionado com a personalidade política” de Modi | Diplomacia | PÚBLICO

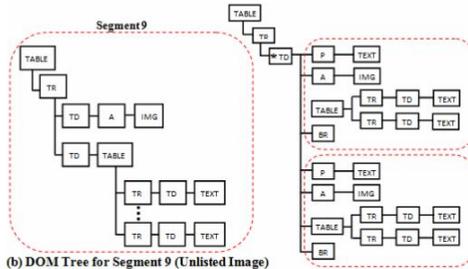
URL da página: <https://www.publico.pt/2020/02/15/politica/noticia/marcelo-ficou-impressionado-personalidade-politica-modi-1904277>

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

Descobrir palavras descritivas para imagens: Estado da Arte

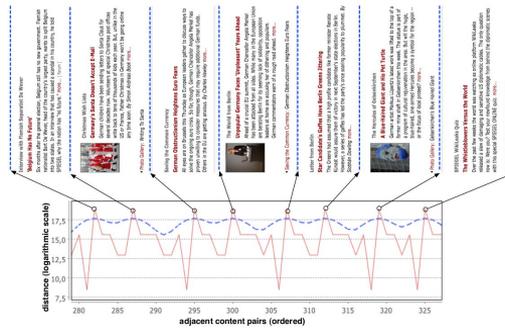


(a) Image segments 1 - 9



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

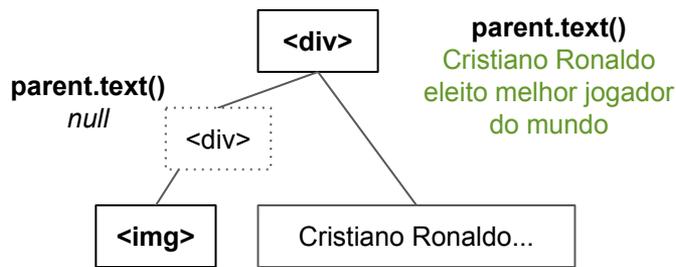
Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information



Sadet, Alci & 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

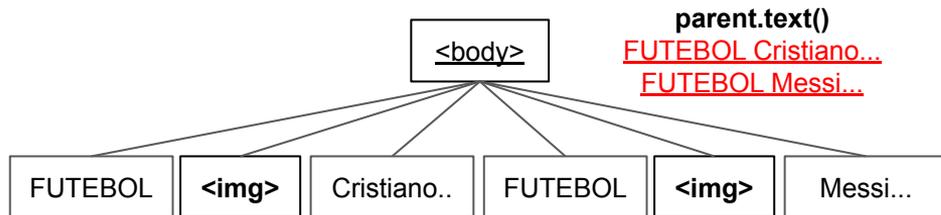
Associar palavras do elemento *parent* do HTML à imagem



Cristiano
Ronaldo eleito
melhor jogador
do mundo

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)

FUTEVOL



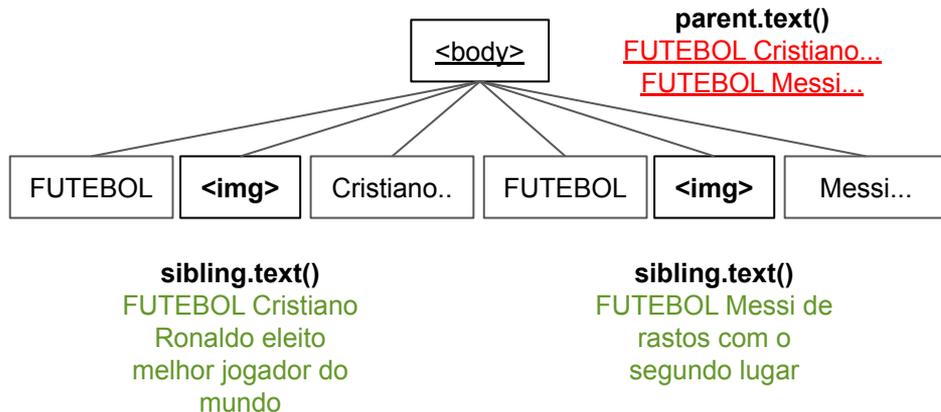
Cristiano Ronaldo eleito melhor jogador do mundo

FUTEVOL



Messi de rastros com o segundo lugar

Solução adoptada: método híbrido



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

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 da imagem: [imagens publico pt imagens aspx 1440184](#)

Texto alternativo da imagem: [Marcelo Rebelo de Sousa](#)

Legenda da imagem: [Marcelo Rebelo de Sousa LUSA/ESTELA SILVA](#)

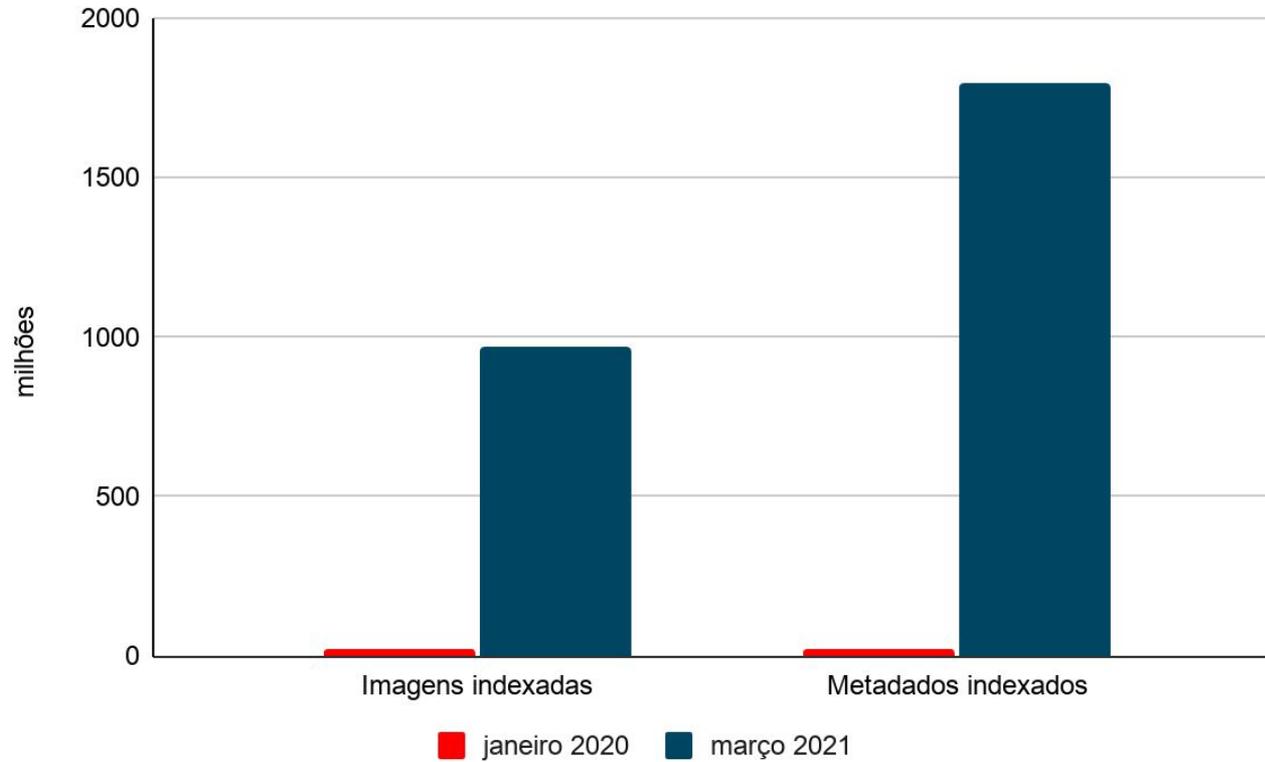


URL da imagem: [imagens publico pt imagens aspx 1044361](#)

Legenda da imagem: [FUTEBOL Moussa Marega, deixa-me dizer-te uma coisa - Opinião de Adriano Miranda](#)

Processar 1 800 milhões de images

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



P PÚBLICO | ÍPSILON | ÍMPAR | FUGAS | P3 | CINECARTAZ | CLUBE P

POLÍTICA | PS | CDS-PP | BE

DIPLOMACIA

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

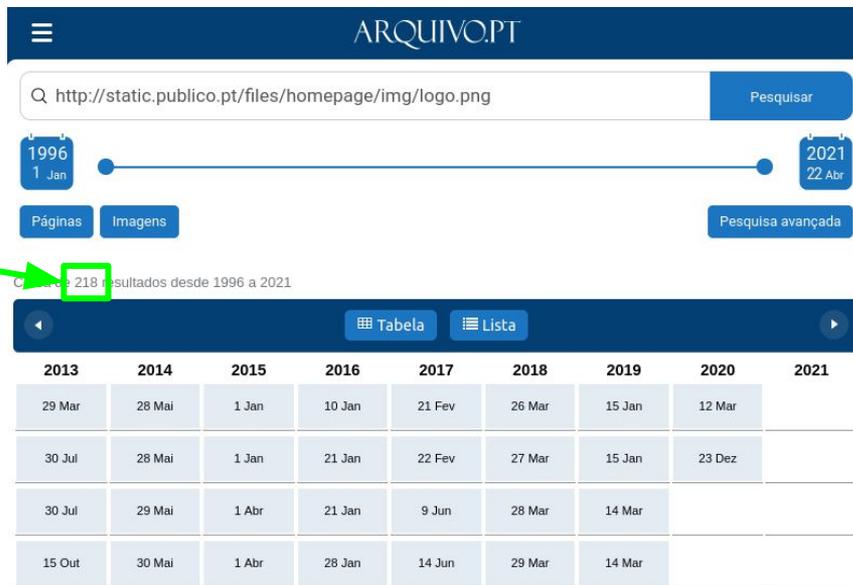
O Presidente da República está de visita de estado à Índia.

Lusa - 15 de Fevereiro de 2020, 11:43

36 PARTILHAS

MAIS POPULARES

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



ARQUIVO.PT

Q http://static.publico.pt/files/homepage/img/logo.png

Pesquisar

1996 1 Jan 2021 22 Abr

Páginas Imagens Pesquisa avançada

218 resultados desde 1996 a 2021

Tabela Lista

2013	2014	2015	2016	2017	2018	2019	2020	2021
29 Mar	28 Mai	1 Jan	10 Jan	21 Fev	26 Mar	15 Jan	12 Mar	
30 Jul	28 Mai	1 Jan	21 Jan	22 Fev	27 Mar	15 Jan	23 Dez	
30 Jul	29 Mai	1 Abr	21 Jan	9 Jun	28 Mar	14 Mar		
15 Out	30 Mai	1 Abr	28 Jan	14 Jun	29 Mar	14 Mar		

- 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

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 , 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 **ordem dos resultados**
 - Construir coleção de teste para avaliar sistema actual



Prémio Arquivo.pt 2021

Prémio Arquivo.pt
2021

Concorra até 4 de maio

VIAJE NO TEMPO

e ganhe 10.000€

Crie um trabalho individual ou em grupo que use o Arquivo.pt

1º Classificado 10.000€
2º Classificado 3.000€
3º Classificado 2.000€

Menção honrosa

Saiba mais em:
arquivo.pt/premio2021

FCT Fundação para a Ciência e a Tecnologia

COMISSÃO DE AVALIAÇÃO DO PRÉMIO ARQUIVO.PT 2021

 O Presidente da República

P

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

Casos de uso (API pesquisa de imagens)

The screenshot displays the 'Time-Matters Demo' interface. At the top, there is a navigation bar with links for Home, Tag Dates, API, GitHub, Related Projects, and About. Below the navigation bar, a red banner contains the text 'Check out our [video](#) and [poster](#) presentation at ECIR 2021'. The main content area features a horizontal timeline from 1964 to 1981. A specific date, 'APRIL 25, 1974', is highlighted, with a corresponding image of a red carnation flower. The text next to the image reads 'Forte orientação socialista' and 'Score: 0.935'. Below the image, there is a detailed description of the 1974 revolution in Portugal. The interface also includes a search bar, a 'Show only relevant dates' toggle, and a 'Copy to clipboard' button.

Time-Matters Demo
to see how it works, check out the [video](#) or [poster](#)
Check out our [video](#) and [poster](#) presentation at ECIR 2021

Home Tag Dates API GitHub Related Projects About

Annotated text Storyline Temporal Clustering Timeline Scores

Show only relevant dates

APRIL 25, 1974
Forte orientação socialista

Score: 0.935

A revolução de 25 de Abril, também conhecida como revolução dos Cravos ou revolução de Abril,[1] refere-se a um evento da história de Portugal resultante do movimento político e social, ocorrido a 25 de abril de 1974, que após o regime ditatorial do Estado Novo,[2] vigente desde 1933,[3] e que iniciou um processo que viria a terminar com a implantação de um regime democrático e com a entrada em vigor da nova Constituição a 25 de abril de 1976, marcada por forte orientação socialista.

BASEANDO-SE INICIALMENTE
GENERAL ANTÓNIO

Go back Copy to clipboard

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

Arquivo.pt em acesso aberto!



- **Pesquisa** 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