Pesquisando milhões de imagens no Arquivo.pt

23 de abril de 2021

André Mourão, Ph.D.
Investigação & Desenvolvimento, Arquivo.pt
andre.mourao@fccn.pt

ARQUIVO.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

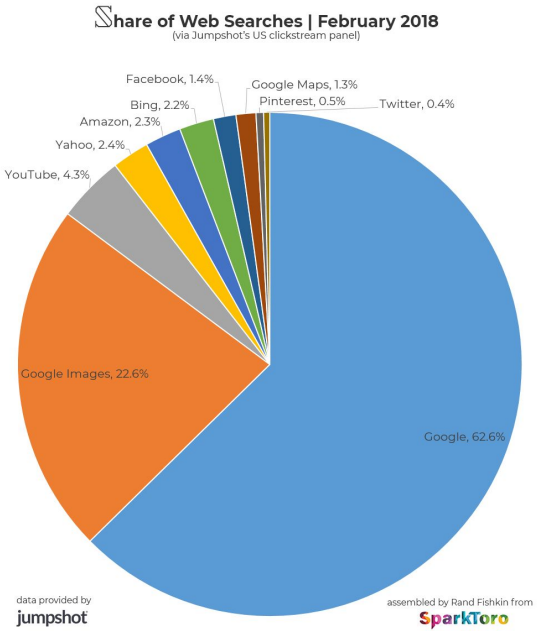
Imagem

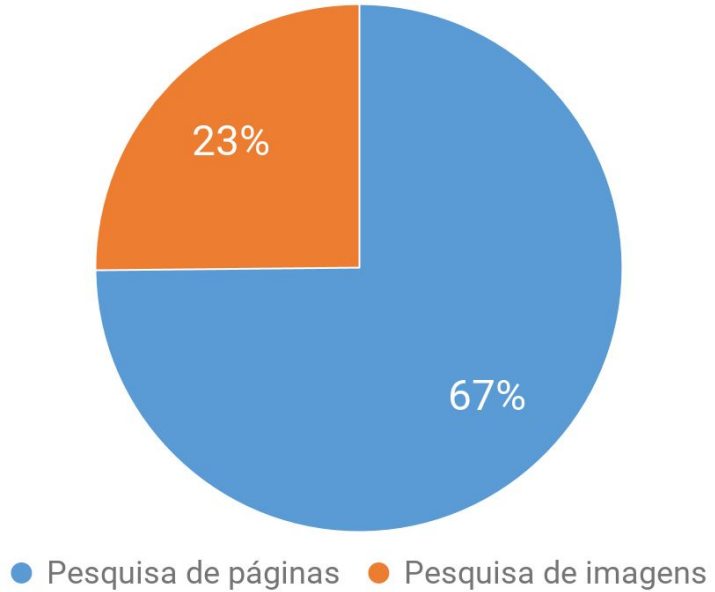
Imagens

Imagens

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

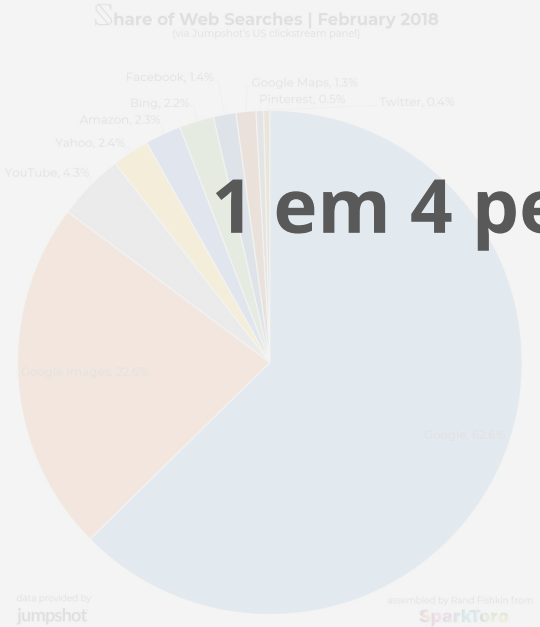
A pesquisa de imagens é importante!



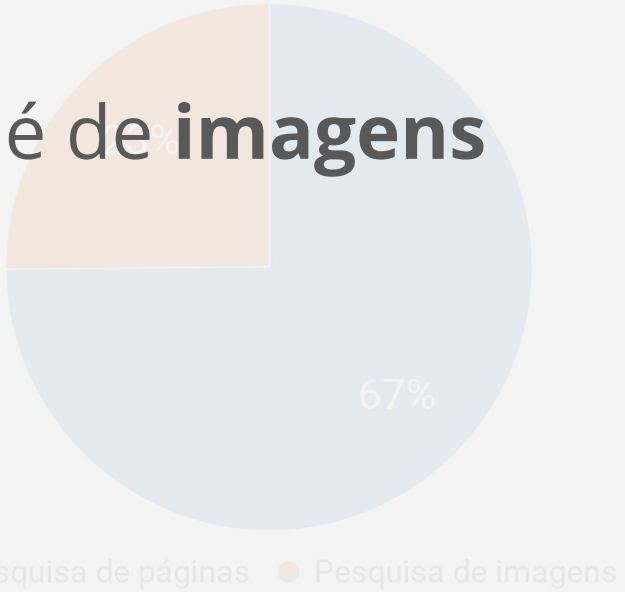


sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

A pesquisa de imagens é importante!

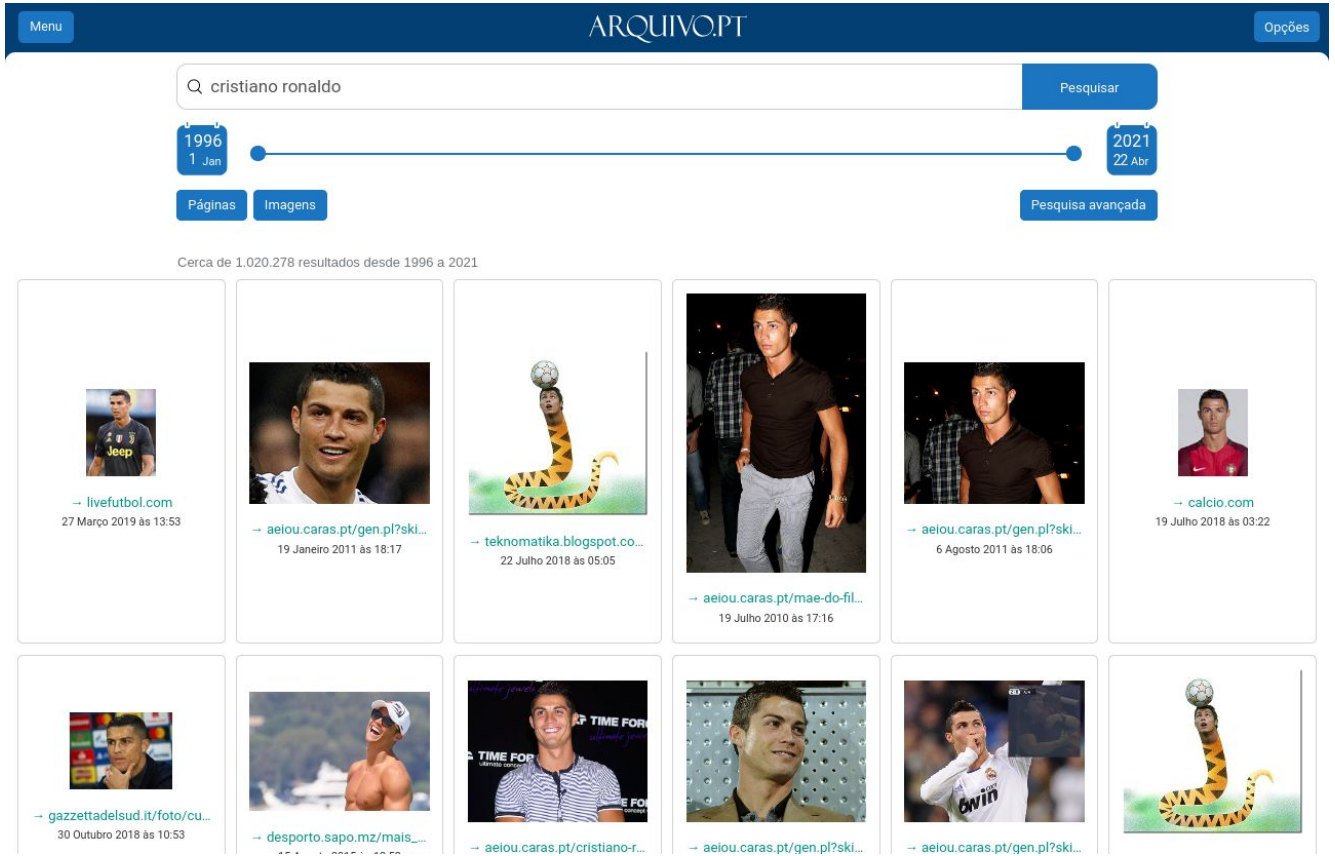


1 em 4 pesquisas é de **imagens**

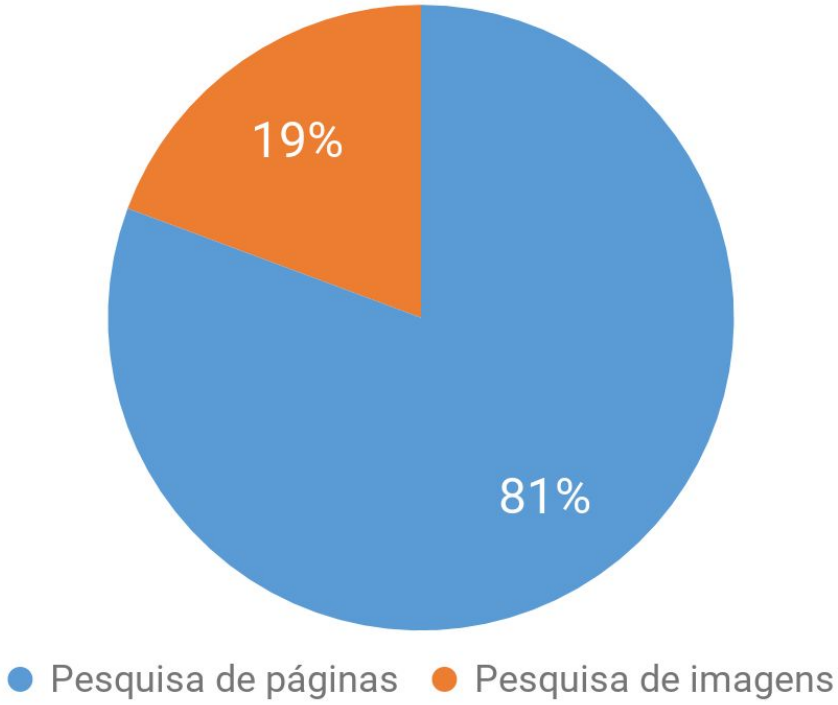


sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Pesquisa de imagens do Arquivo.pt



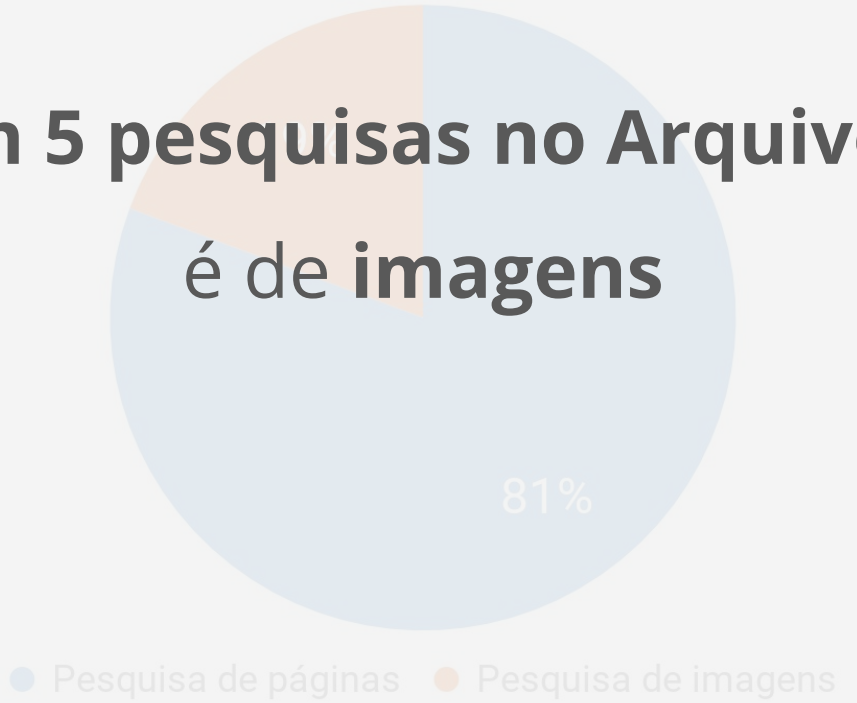
Distribuição da pesquisa de imagens no Arquivo.pt



Distribuição da pesquisa de imagens no Arquivo.pt

1 em 5 pesquisas no Arquivo.pt

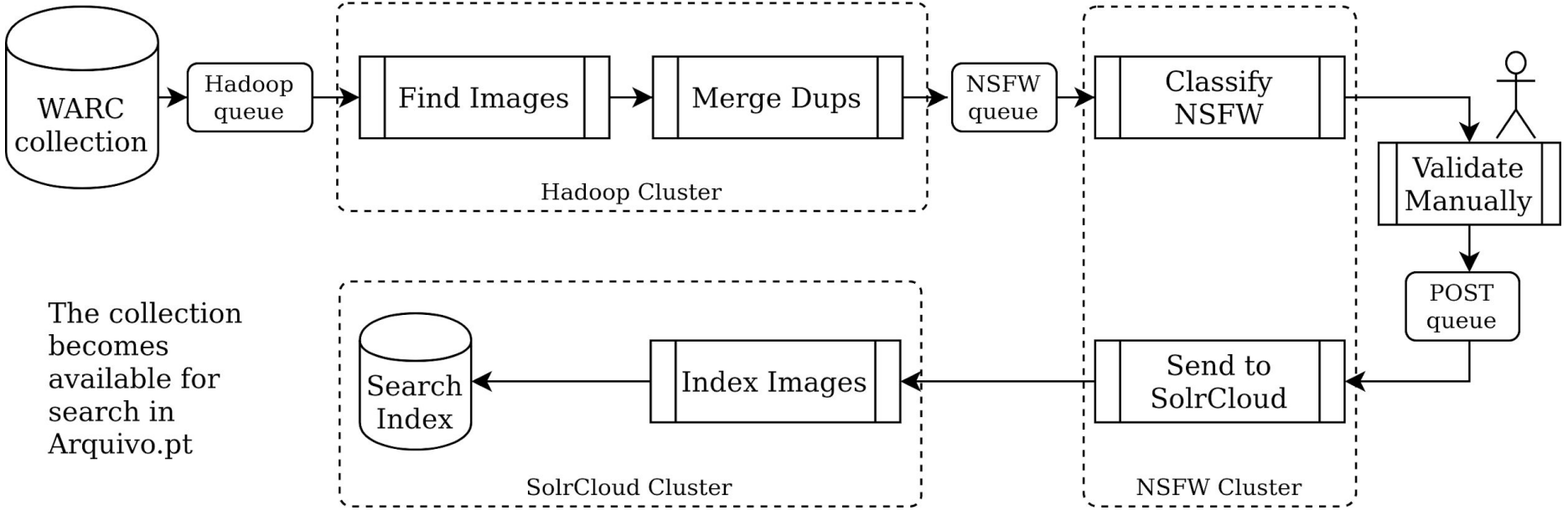
é de **imagens**



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

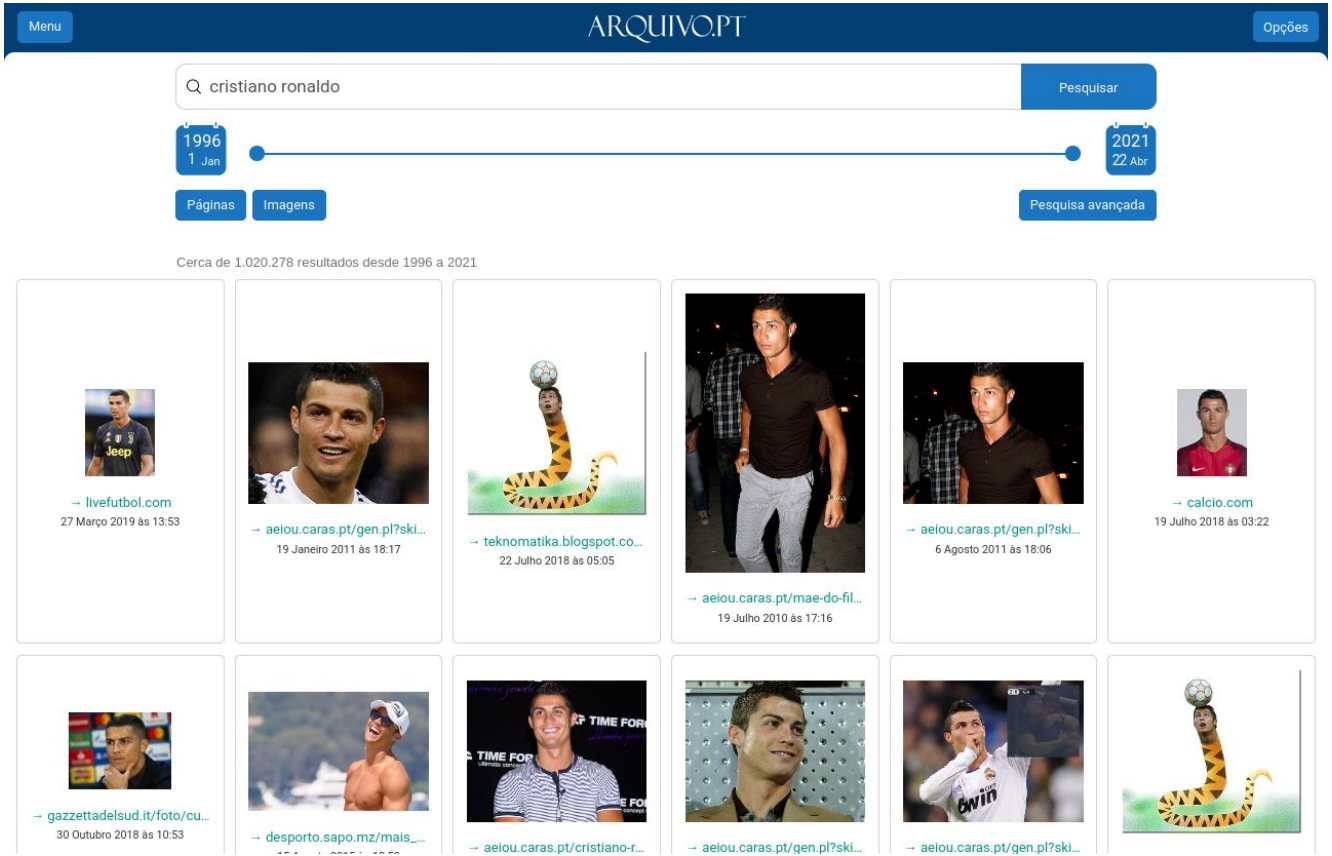
- Encontrar palavras relevantes para imagens
- Lidar com a escala dos dados recolhidos
- Tornar estas imagens pesquisáveis

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
  - Imagem de arquivo de treino de futebol
  - Renovação de um apartamento
- 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
  - Imagem de arquivo de treino de futebol
  - Renovação de um apartamento
- 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%

Amostra retirada do Arquivo.pt

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

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**

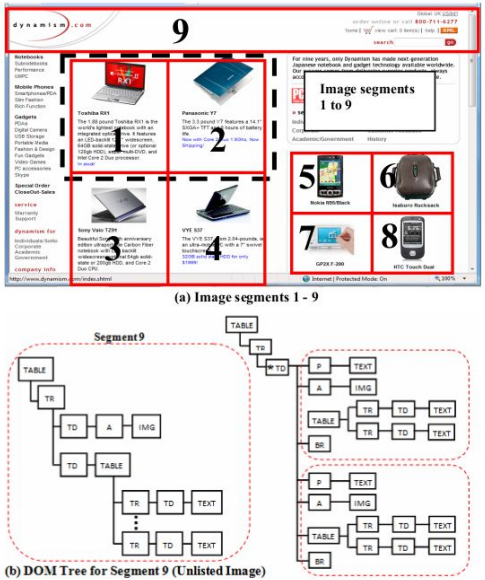
Metadados de página **apenas** descrevem a imagem **principal**

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**?

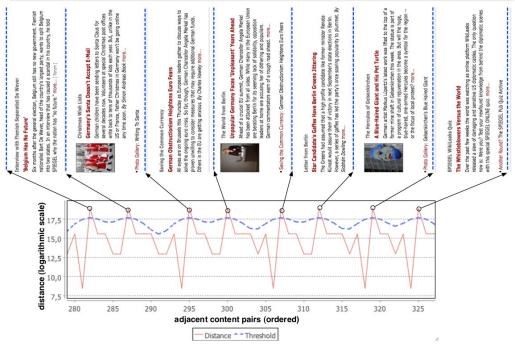
Descobrir palavras descritivas para imagens: Estado da Arte



(a) Image segments 1 - 9

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

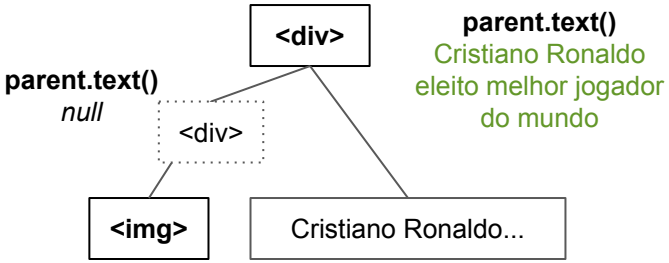
Fauzi, Fariza & Hong, Jer Lang & Belkhatir, Mohammed. (2009). Webpage segmentation for extracting images and their surrounding contextual information



Sadet, 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

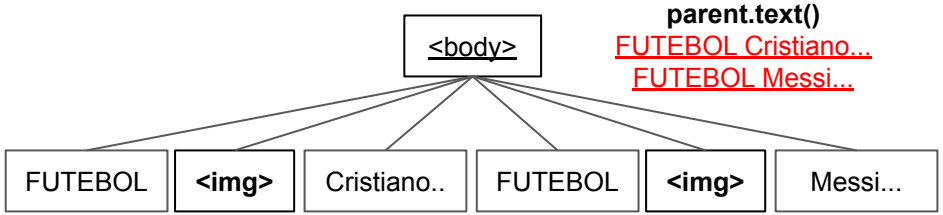
Associar palavras do elemento *parent* do HTML à imagem



Cristiano Ronaldo eleito melhor jogador do mundo

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

Hipótese falha em páginas com estrutura "*flat*"



FUTEBOL

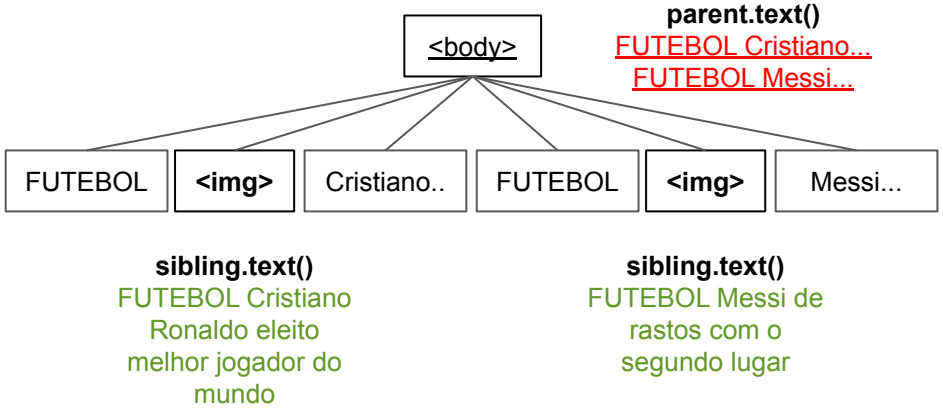
Cristiano Ronaldo eleito melhor jogador do mundo

FUTEBOL

Messi de rastos com o segundo lugar

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

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

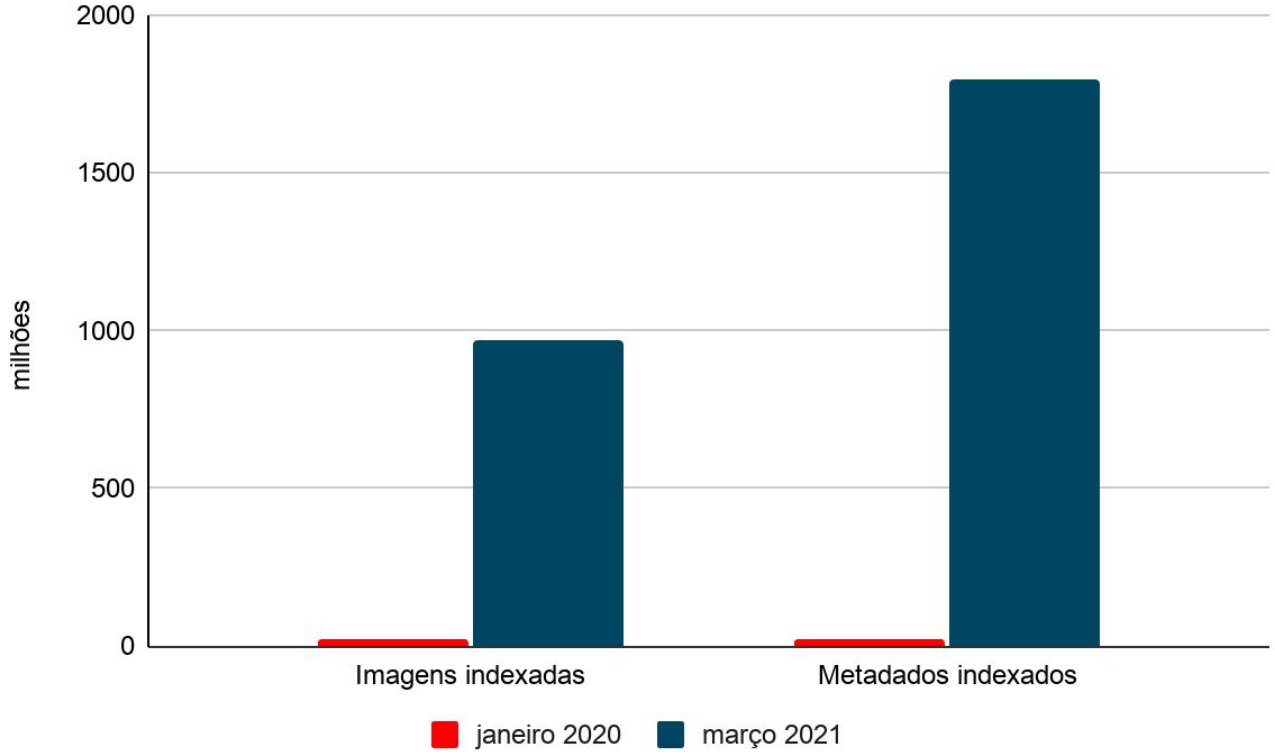
Mais palavras descritivas para as imagens da página

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

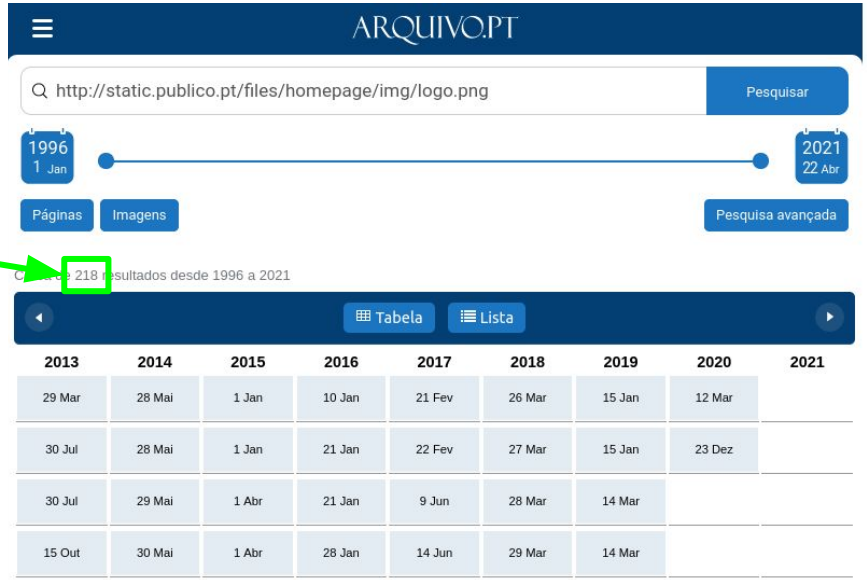
Volume de imagens indexadas: crescimento desde 2020



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



- 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

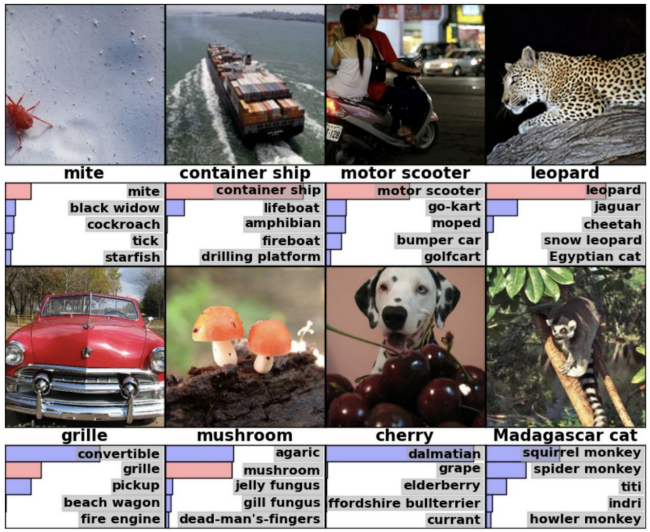
- **Resultado:** 584 milhões de imagens, com informação de 1 800 milhões
  - Redução de 70% nos dados a indexar

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 <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

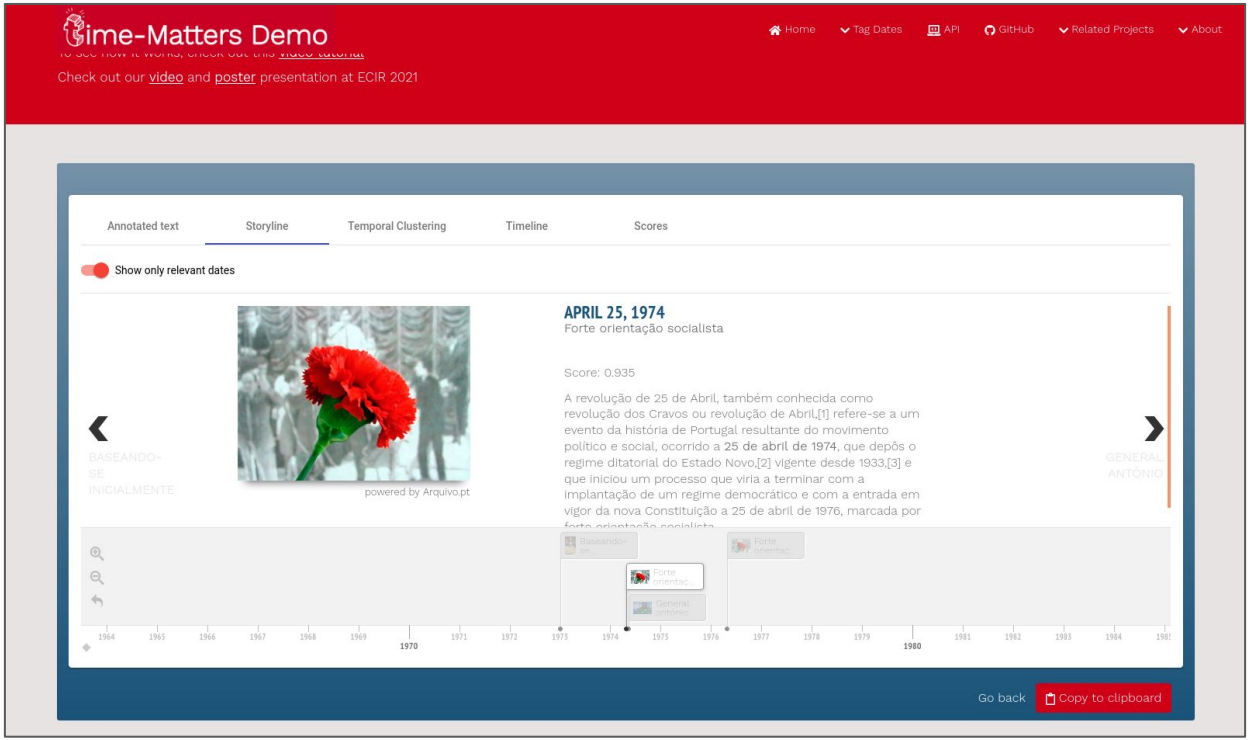
- Imagens **sem metadata**
  - **300+ milhões** de imagens sem texto associado
  - Legendagem baseada em redes neuronais
- Imagens **semelhantes**
  - Mesma imagem, resoluções e/ou formatos diferentes
  - Fazer deduplicação de ***near duplicates***
- Melhorar **ordem dos resultados**
  - Construir coleção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



Prémio Arquivo.pt 2021

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

Casos de uso (API pesquisa de imagens)



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

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