Arquivo.pt image search 2020 → 2021

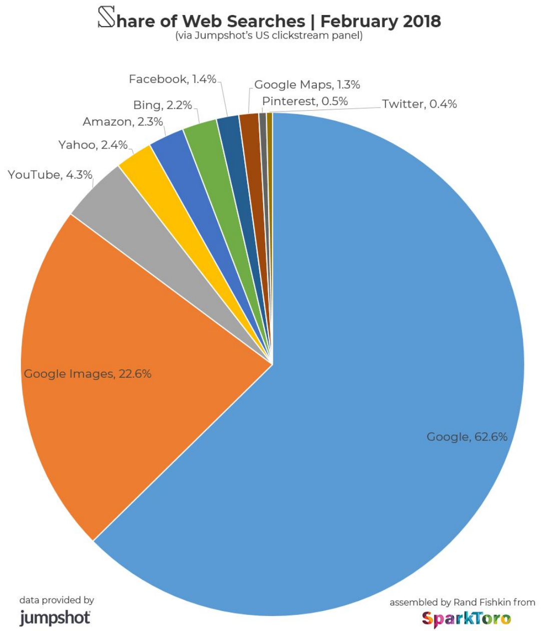
André Mourão

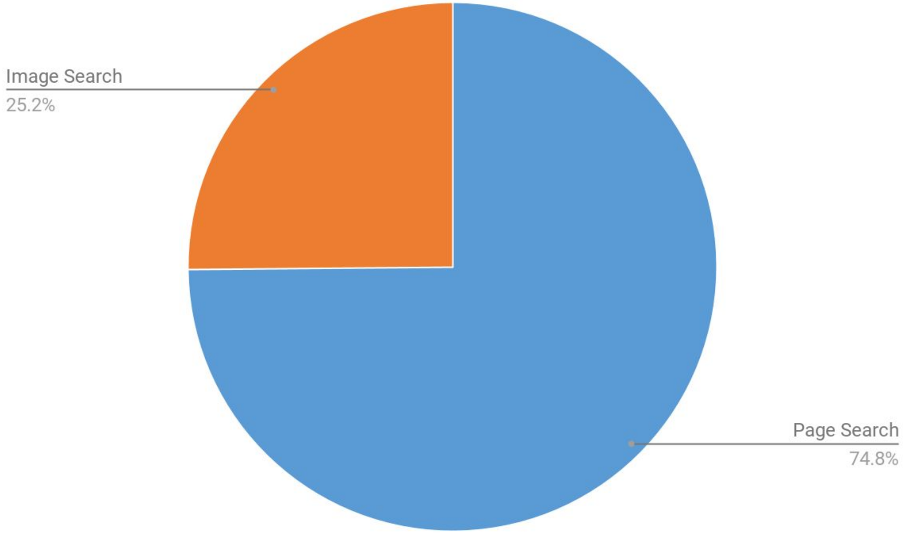
February 2nd 2021

ARQUIVO.PT

FCT Fundação para a Ciência e a Tecnologia

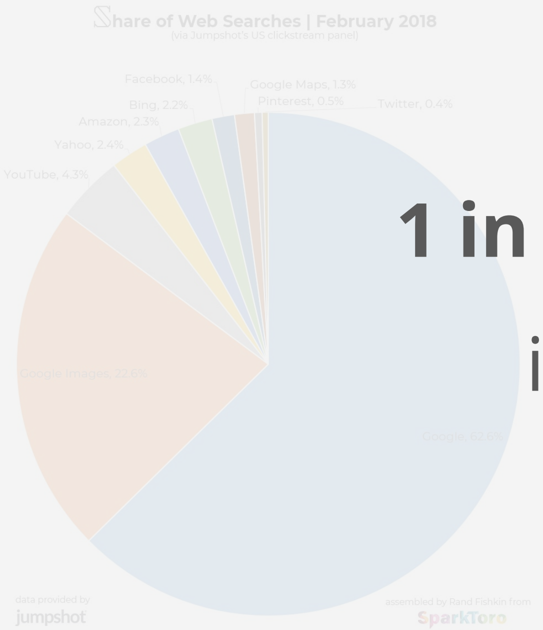
Why does image search matter?

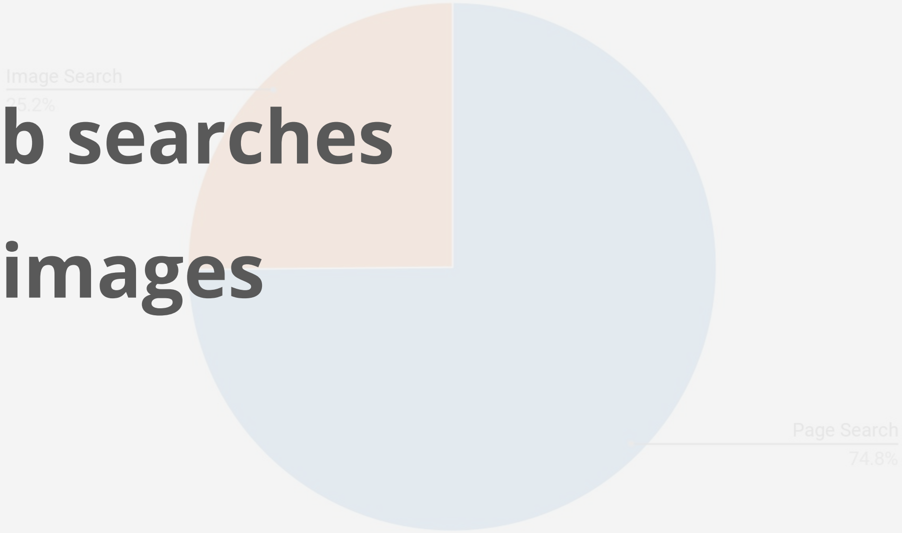




sparktoro.com/blog/new-jumpshot-2018-data-where-searches-happen-on-the-web-google-amazon-facebook-beyond/

Why does image search matter?





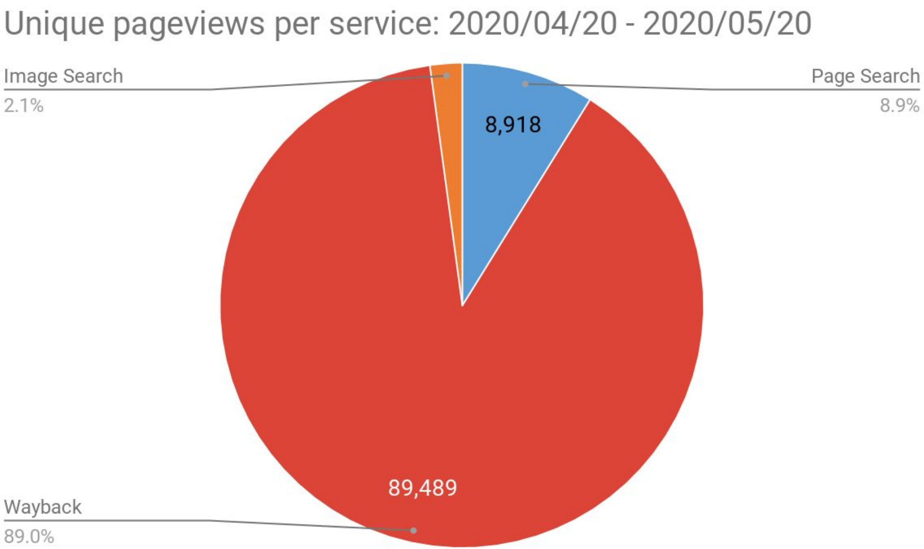
1 in 4 web searches

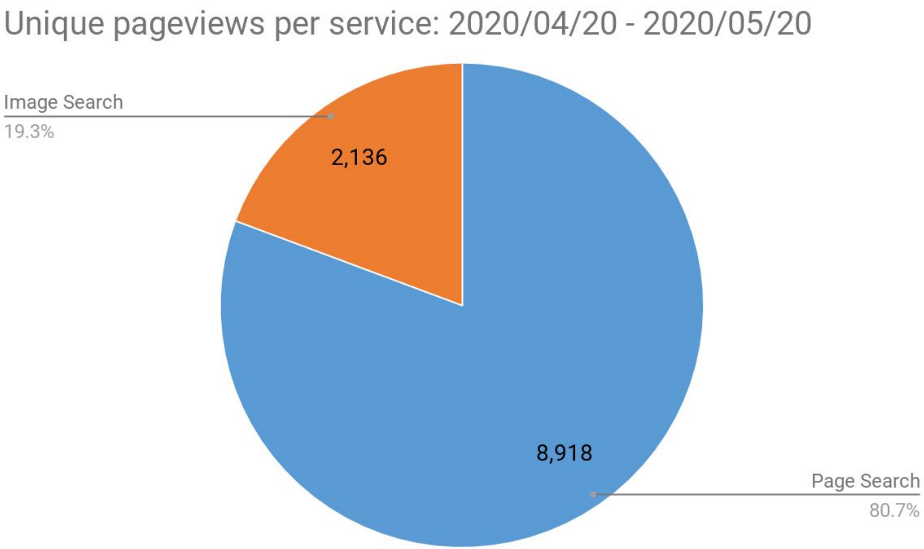
is for **images**

1 in 4 web searches is for images

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

What about Arquivo.pt?

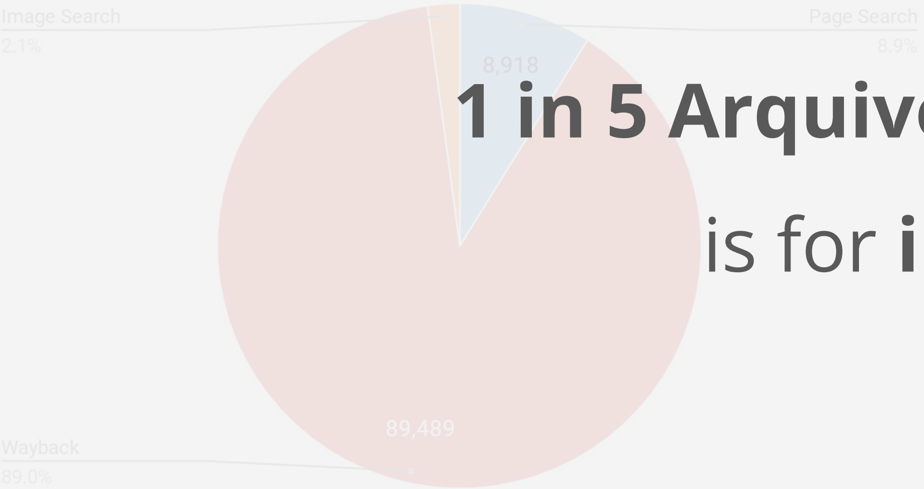




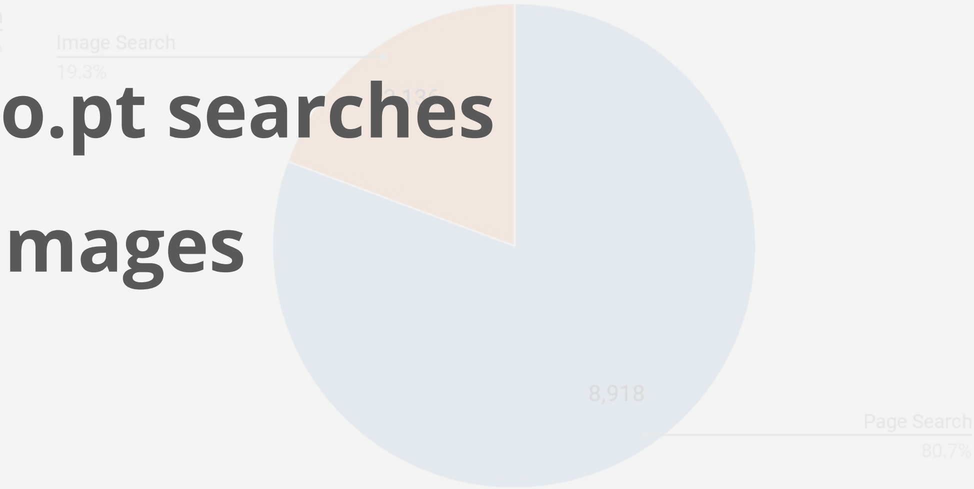
Arquivo.pt Google Analytics

What about Arquivo.pt?

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



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



1 in 5 Arquivo.pt searches

is for **images**

Arquivo.pt Google Analytics

Arquivo.pt Image Search (SCREENSHOT)

- Keyword search
- Sentence search
- Filter by time, size, file type, site
- Sort by score

Arquivo.pt Image Search API (SCREENSHOT)

- Opensource

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22,881,688**
Collection count	90
(W)ARCs	3,465,059
(W)ARC sizes	334 TB
Total collected files	5,962,498,489
Total collected images	**1,602,337,670**
Oldest image date	15/04/1994
Newest image date	14/11/2019
Daily page views	~87

Opportunities for improvement

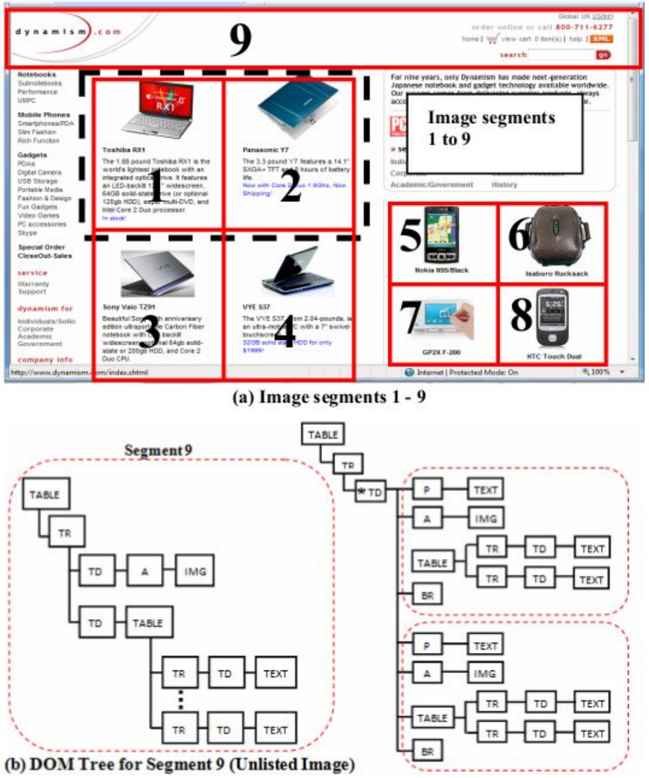
- Lack of image specific metadata
  - 43% (10,163,080 images) without imgAlt or imgTitle
- Why is the difference between collected and indexed so large?
- Only the oldest page per image is indexed
- Search result ranking does not take image popularity into account

Finding images in pages results

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

	Percentage of references
<img>	90.6%
<a>	8.7%
css	0.7%
Normal images	99.9%
base64	0.1%

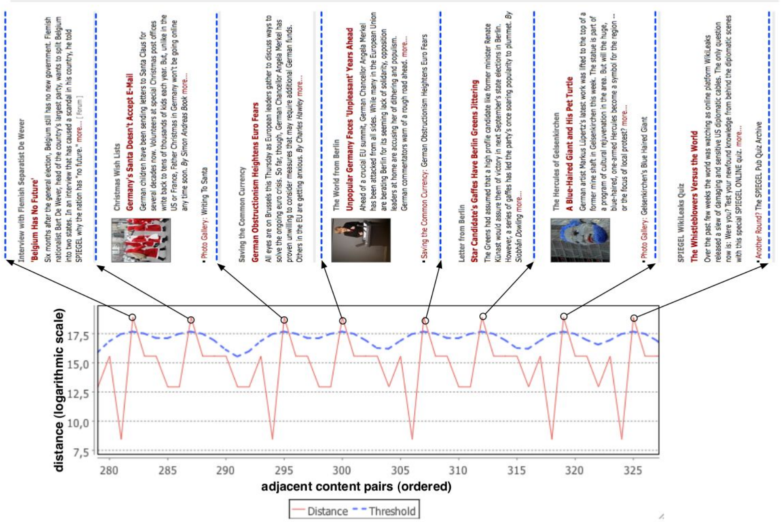
Finding an image caption



(a) Image segments 1 - 9

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

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



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

Image caption extraction

I arrived at the following method

First parent with text

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

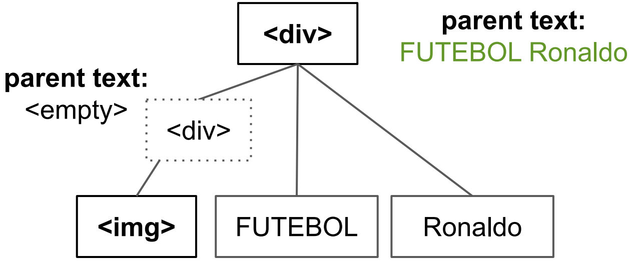
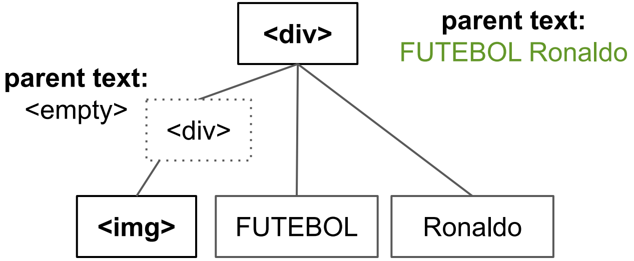


Image caption extraction

I arrived at the following method

First parent with text

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



Previous and next node text

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

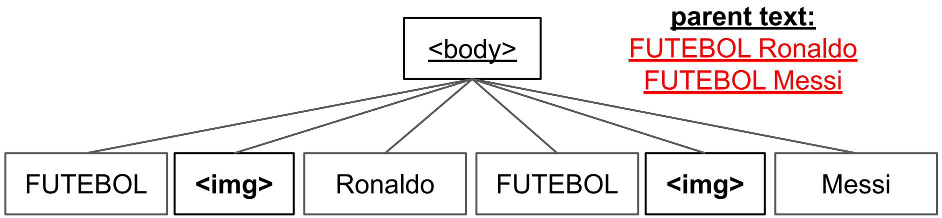
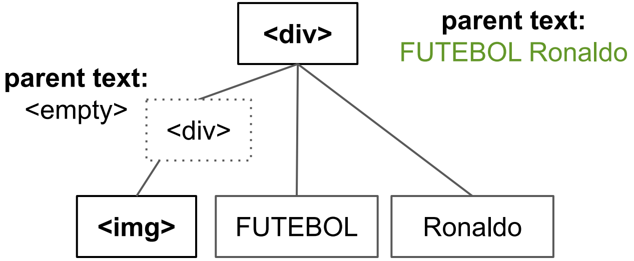


Image caption extraction

I arrived at the following method

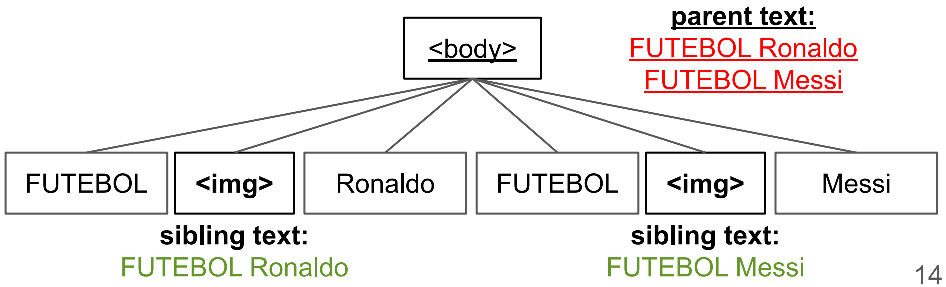
First parent with text

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



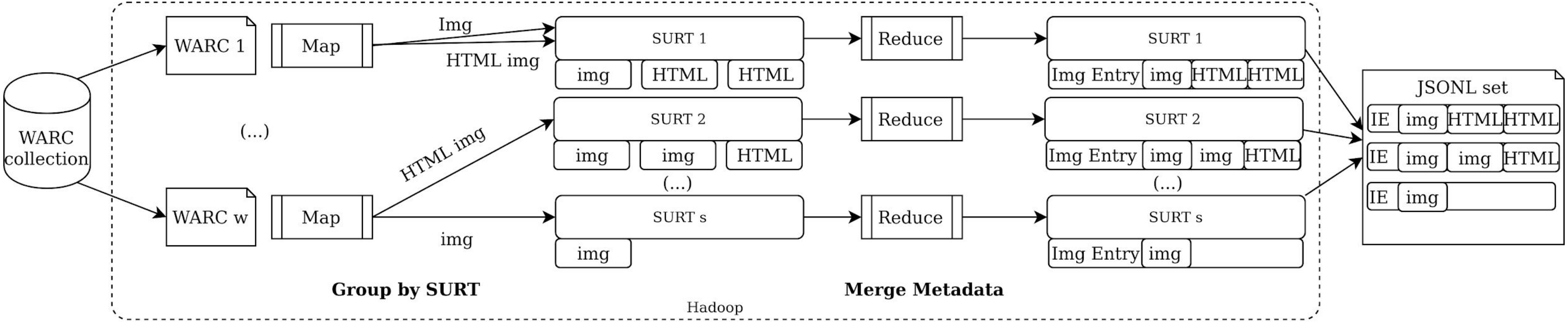
Previous and next node text

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



Indexing Architecture

Map Reduce: Extract images and metadata



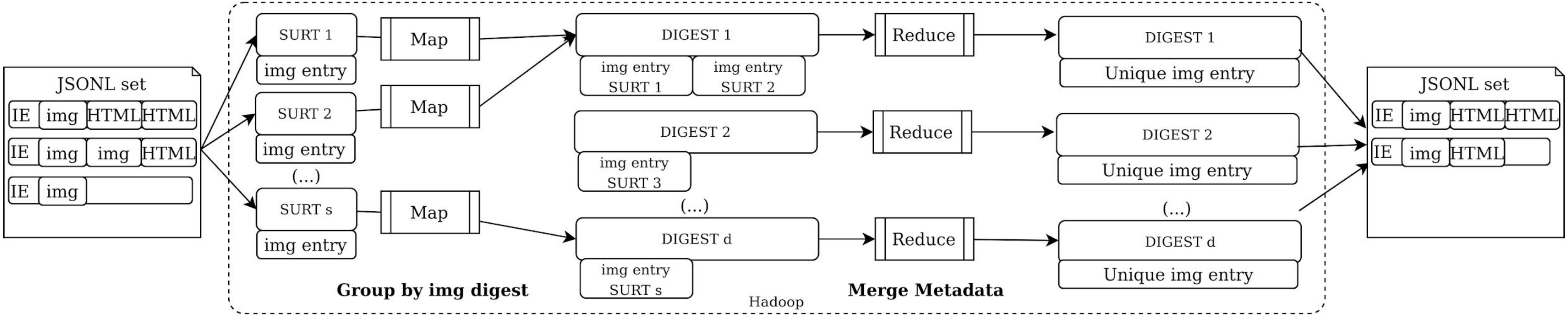
How to deal with duplicate information?

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

Deduplication selected solution

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

Map Reduce: Group by digest



Duplicates across collections

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

My predictions in May 2020

Arquivo.pt Image Search (as of Jan 2020)

Indexed images	**22,881,688**
Collection count	90
(W)ARCs	3,465,059
(W)ARC sizes	334 TB
Total collected files	5,962,498,489
Total collected images	**1,602,337,670**
Oldest image date	15/04/1994
Newest image date	14/11/2019
Daily page views	~87

Tested collections - number of images

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

Takeways

~200,000,000-650,000,000 images

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

~9-28x more images

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

~400,000,000-1,300,000,000 pages (2/image)

~18-56x more pages

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

Takeways

~200,000,000-650,000,000 images

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

[illegible]

~9-28x more images

~400,000,000-1,300,000,000 pages (2/image)

~18-56x more pages

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

Takeways

654,061,494 images

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

28.58x more images

23,589,395 -> [illegible]

1,252,612,982 pages (1.91/image)

54.74x more pages

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

Takeways

+ 316,943,978 images in one year (2019)

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

48% growth

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

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

+ 609,698,474 pages in one year (2019)

49% growth

Takeways

971,005,472 images

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

42.44x more images

[illegible] -> [illegible]

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

1,862,311,456 pages (1.91/image)

81.34x more pages

Impact of deduplication

	Number of documents
a	1,862,311,456 image-page pair documents
b	584,242,176 unique documents (971,005,472 before deduplication across collections)
c	**584,242,176** documents, containing information from all 1,862,311,456 image-page pairs

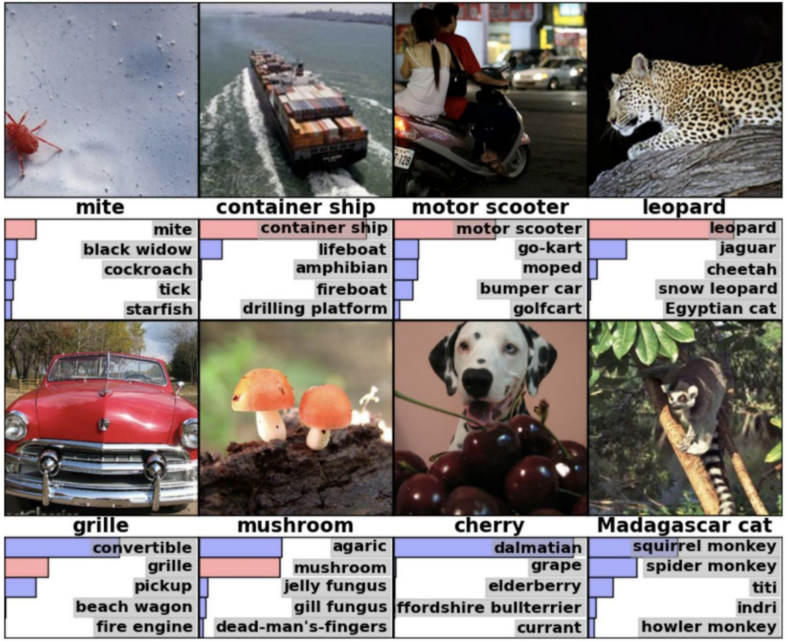
Where to index these **584,242,176** documents?

Summary of what changed in 2020?

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

Plan for the future

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



Ranking features for 2021

ARQUIVO.PT

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