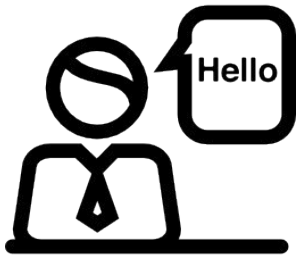
ARQUIVO.PT

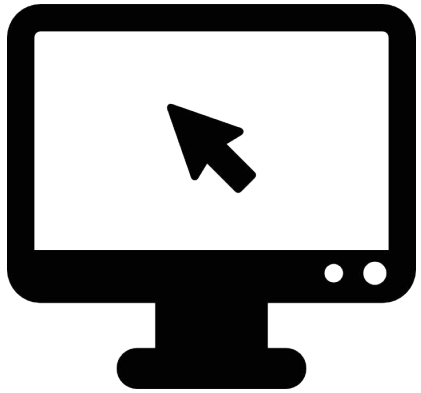
Architecture of the Portuguese Web Archive Search System

Hugo Viana
hugo.viana@fccn.pt

Presentation outline

- About Arquivo.pt;
- Architecture;
- Workflows;
- Operation;
- Ranking;



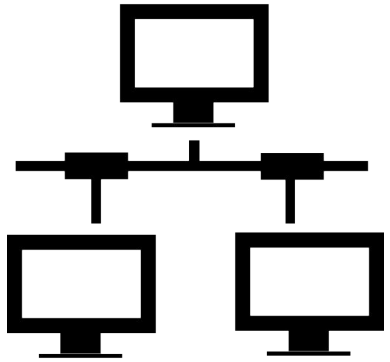


About Arquivo.pt: Searching with Arquivo.pt service

ARQUIVO.PT

Who are we?!

ARQUIVO.PT



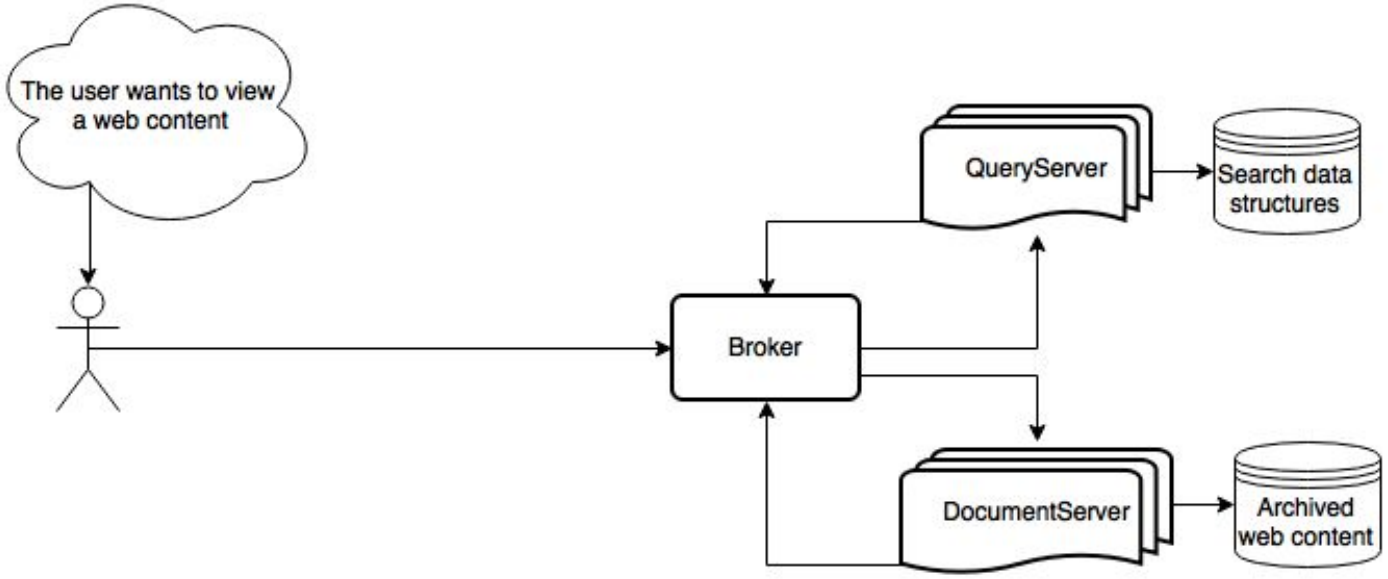
Architecture & Workflow: An overview

ARQUIVO.PT

Architecture & Workflow: *Arquivo.pt*

1. Broker
   a. Delivering the ranked search results to users;
   b. Hosting the user interfaces provided to users;
2. Cluster of QueryServers
   a. Finding relevant documents for a search query;
   b. Stores the indexes;
3. Cluster of DocumentServers
   a. Provides HTTP navigation through archived Web content;
   b. Stores the archived web content in ARC format;

Architecture & Workflow: *Arquivo.pt*



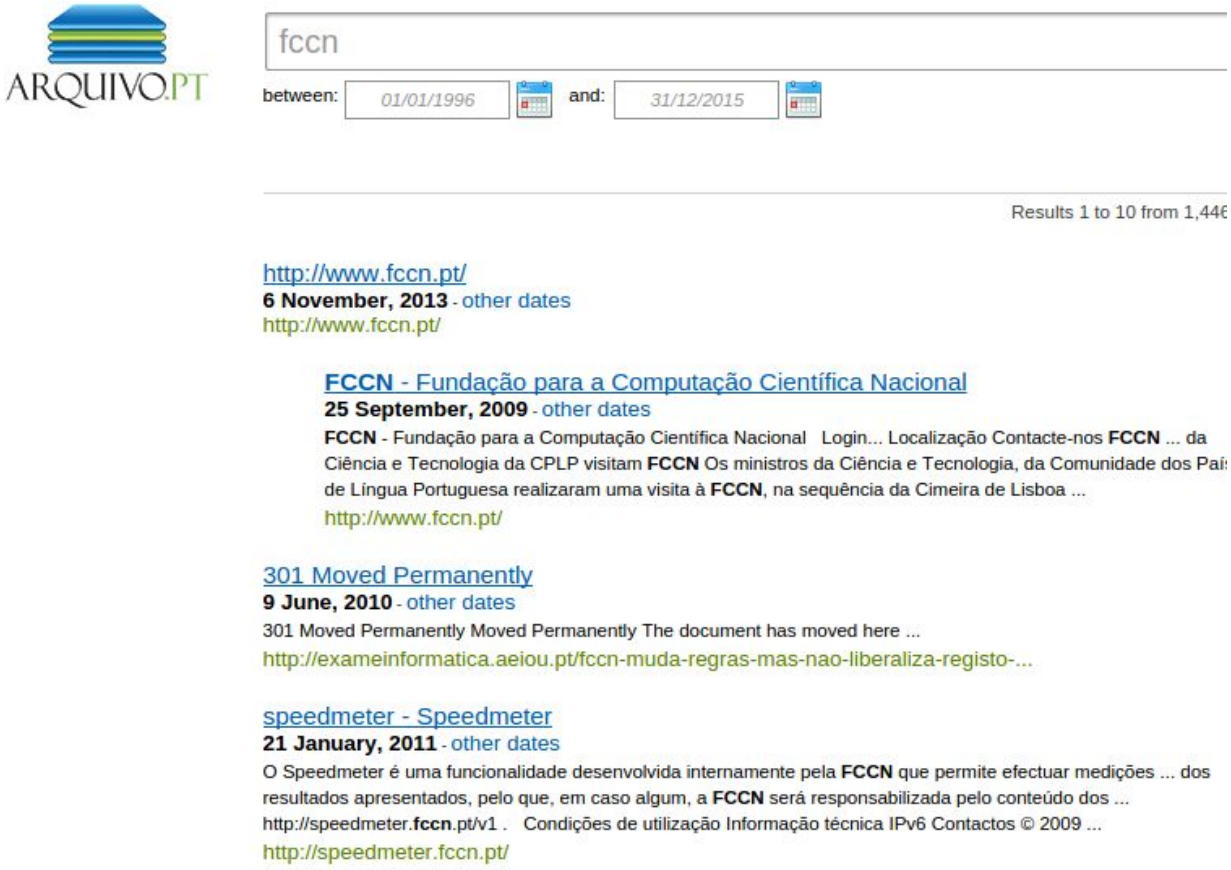
What is this all about?!

No worries ..

ARQUIVO.PT

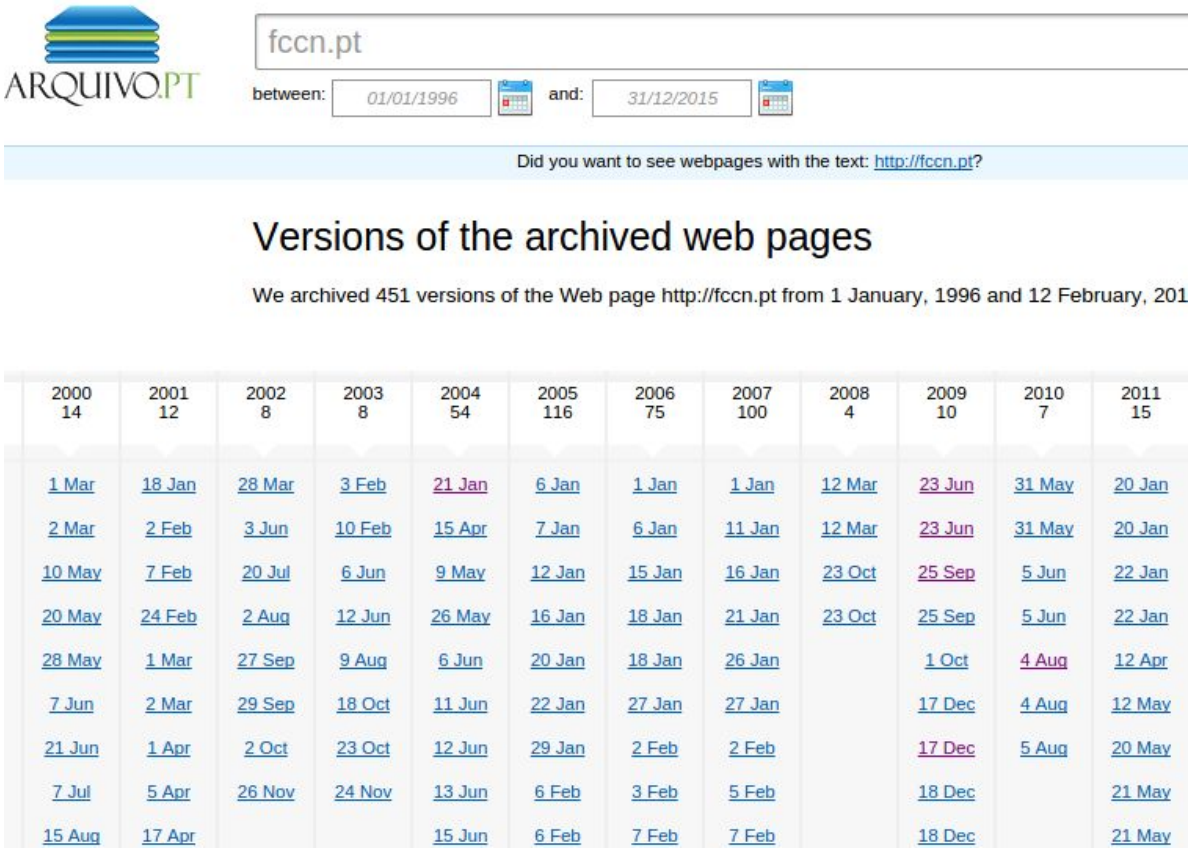
Broker: Home page

Broker: Full-text search results



ARQUIVO.PT

Broker: Url search results



2000 14	2001 12	2002 8	2003 8	2004 54	2005 116	2006 75	2007 100	2008 4	2009 10	2010 7	2011 15
1 Mar	18 Jan	28 Mar	3 Feb	21 Jan	6 Jan	1 Jan	1 Jan	12 Mar	23 Jun	31 May	20 Jan
2 Mar	2 Feb	3 Jun	10 Feb	15 Apr	7 Jan	6 Jan	11 Jan	12 Mar	23 Jun	31 May	20 Jan
10 May	7 Feb	20 Jul	6 Jun	9 May	12 Jan	15 Jan	16 Jan	23 Oct	25 Sep	5 Jun	22 Jan
20 May	24 Feb	2 Aug	12 Jun	26 May	16 Jan	18 Jan	21 Jan	23 Oct	25 Sep	5 Jun	22 Jan
28 May	1 Mar	27 Sep	9 Aug	6 Jun	20 Jan	18 Jan	26 Jan		1 Oct	4 Aug	12 Apr
7 Jun	2 Mar	29 Sep	18 Oct	11 Jun	22 Jan	27 Jan	27 Jan		17 Dec	4 Aug	12 May
21 Jun	1 Apr	2 Oct	23 Oct	12 Jun	29 Jan	2 Feb	2 Feb		17 Dec	5 Aug	20 May
7 Jul	5 Apr	26 Nov	24 Nov	13 Jun	6 Feb	3 Feb	5 Feb		18 Dec		21 May
15 Aug	17 Apr			15 Jun	6 Feb	7 Feb	7 Feb		18 Dec		21 May

ARQUIVO.PT

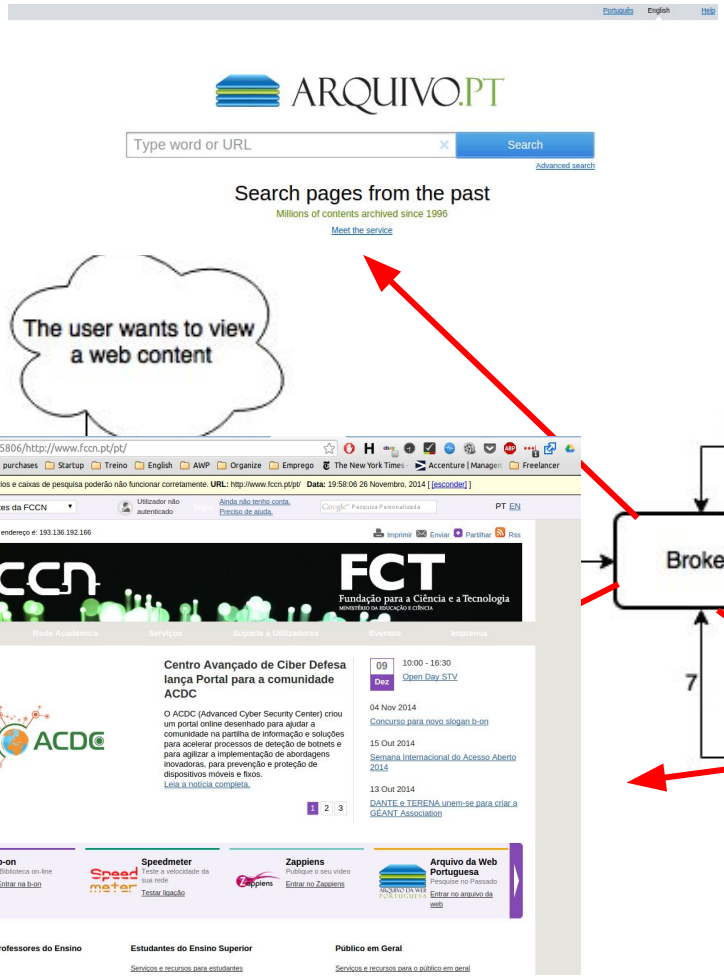
QueryServer: Produces the results lists.

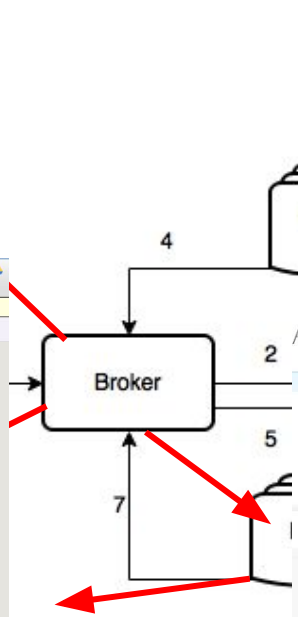
```
<description>PWA search results for query: fccn</description>
<link>http://archive.pt</link>
<opensearch:totalResults>1446620</opensearch:totalResults>
<opensearch:startIndex>0</opensearch:startIndex>
<opensearch:itemsPerPage>10</opensearch:itemsPerPage>
<opensearch:Query role="request" searchTerms="fccn" startPage="1"/>
<item>
<title>http://www.fccn.pt/</title>
<source url="http://www.fccn.pt/">Original URL of http://www.fccn.pt/</source>
<link>http://arquivo.pt/wayback/20131107104120/http://www.fccn.pt/</link>
<pwa:id>832</pwa:id>
<pwa:index>14</pwa:index>
<pwa:arcname>IAH-20131106221644-08421-p13.arquivo.pt</pwa:arcname>
<pwa:arcoffset>86890984</pwa:arcoffset>
<pwa:digest>d41d8cd98f00b204e9800998ecf8427e</pwa:digest>
<pwa:tstamp>20131106222844000</pwa:tstamp>
<pwa:contentLength>417</pwa:contentLength>
<pwa:primaryType>text</pwa:primaryType>
<pwa:subType>html</pwa:subType>
</item>
```

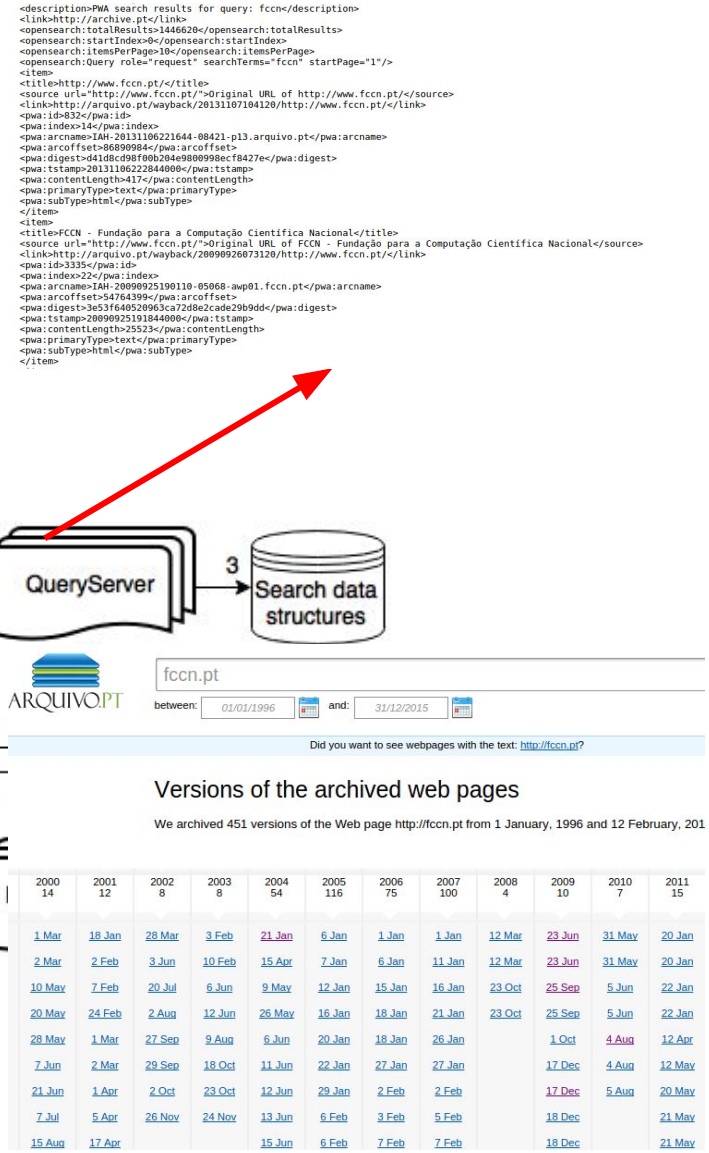
DocumentServer: Web archived content.

ARQUIVO.PT

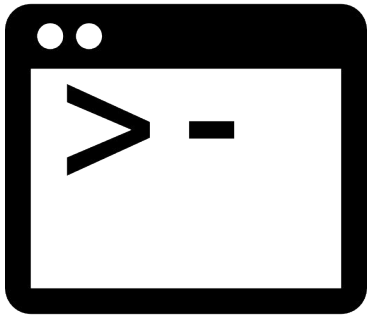
How all this flows?







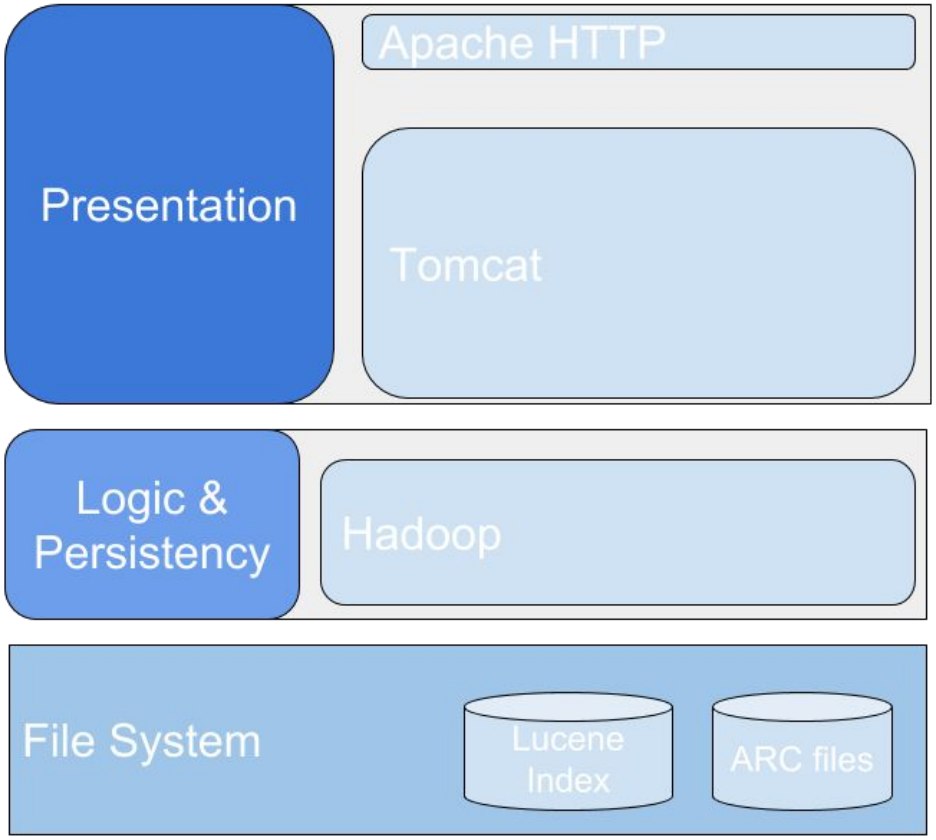
ARQUIVO.PT



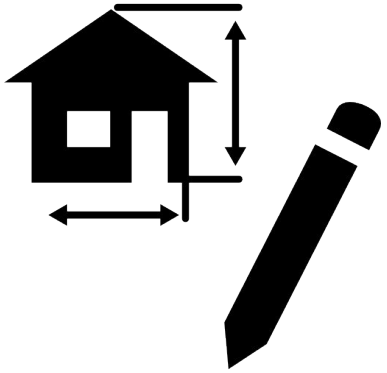
Workflows:
Software components

ARQUIVO.PT

Software components for the search system: workflow



ARQUIVO.PT

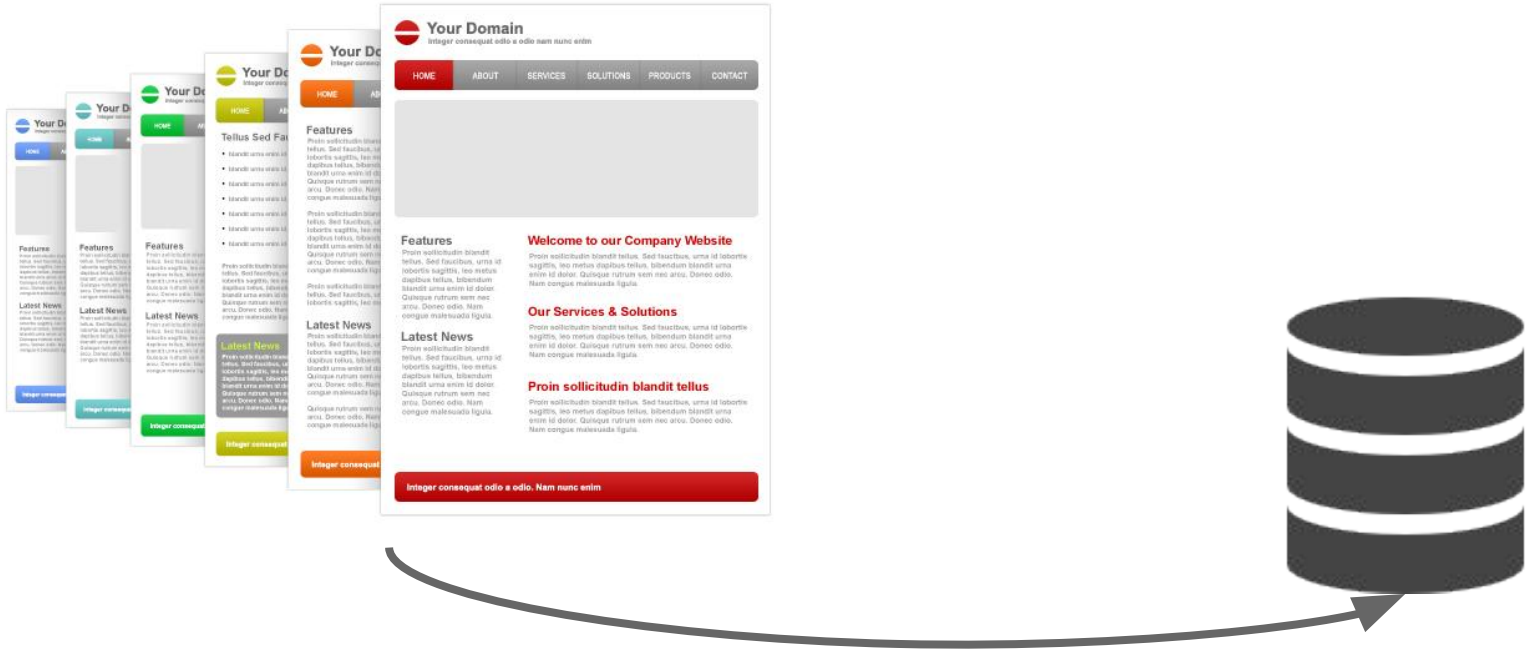


Operations Procedures

Indexing a collection

1. **Parsing the ARC files;**
2. **Invert index;**
3. Create index;
4. **Pruning index;**
5. **Remove stopwords.**

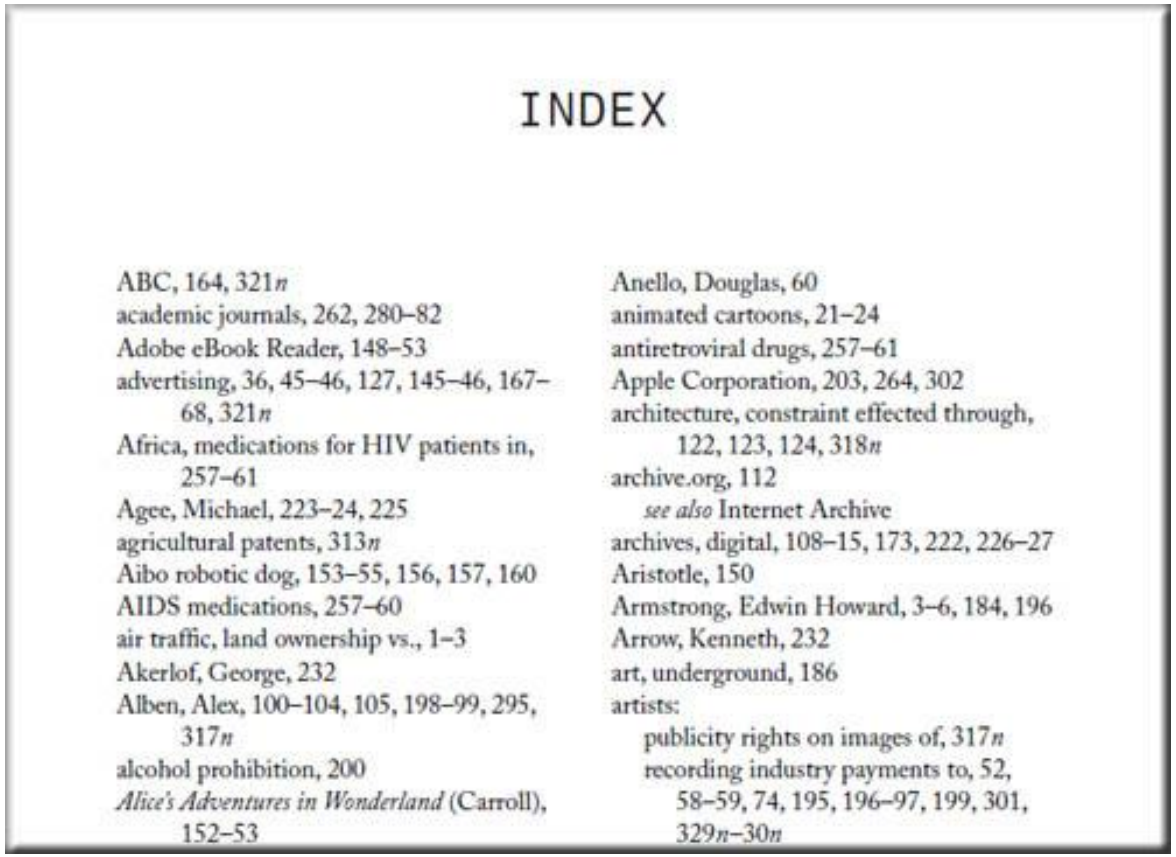
Parsing the ARC files



Parse the archived web content to the Hadoop file system

ARQUIVO.PT

Invert index: What is this?



INDEX

ARQUIVO.PT

Invert index: An example

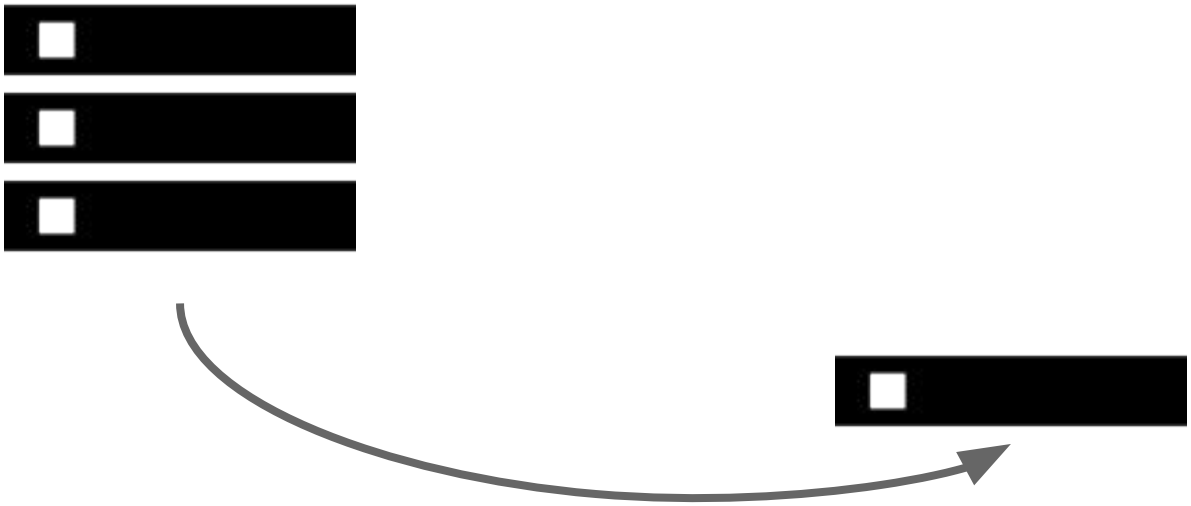
ID	Text
1	Baseball is played during summer months.
2	Summer is the time for picnics here.
3	Months later we found out why.
4	Why is summer so hot here

↑ Sample document data

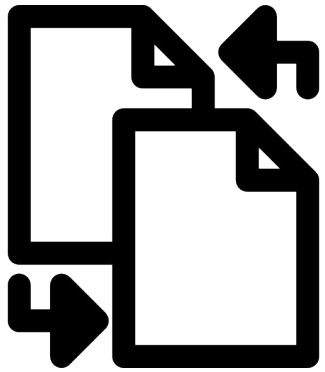
Dictionary and posting lists →

Term	Freq	Document ids
baseball	1	[1]
during	1	[1]
found	1	[3]
here	2	[2], [4]
hot	1	[4]
is	3	[1], [2], [4]
months	2	[1], [3]
summer	3	[1], [2], [4]
the	1	[2]
why	2	[3], [4]

Pruning index and stopwords.



Reducing the indexes



Ranking features used in the Arquivo.pt

ARQUIVO.PT

Ranking features used in the Arquivo.pt

Name of the feature	Field in which the the feature is applied
Nutch	content + url + host + anchor + title
MinSpanCovUnord	title
MinSpanCovUnord	content
MinSpanCovOrd	anchor

Nutch

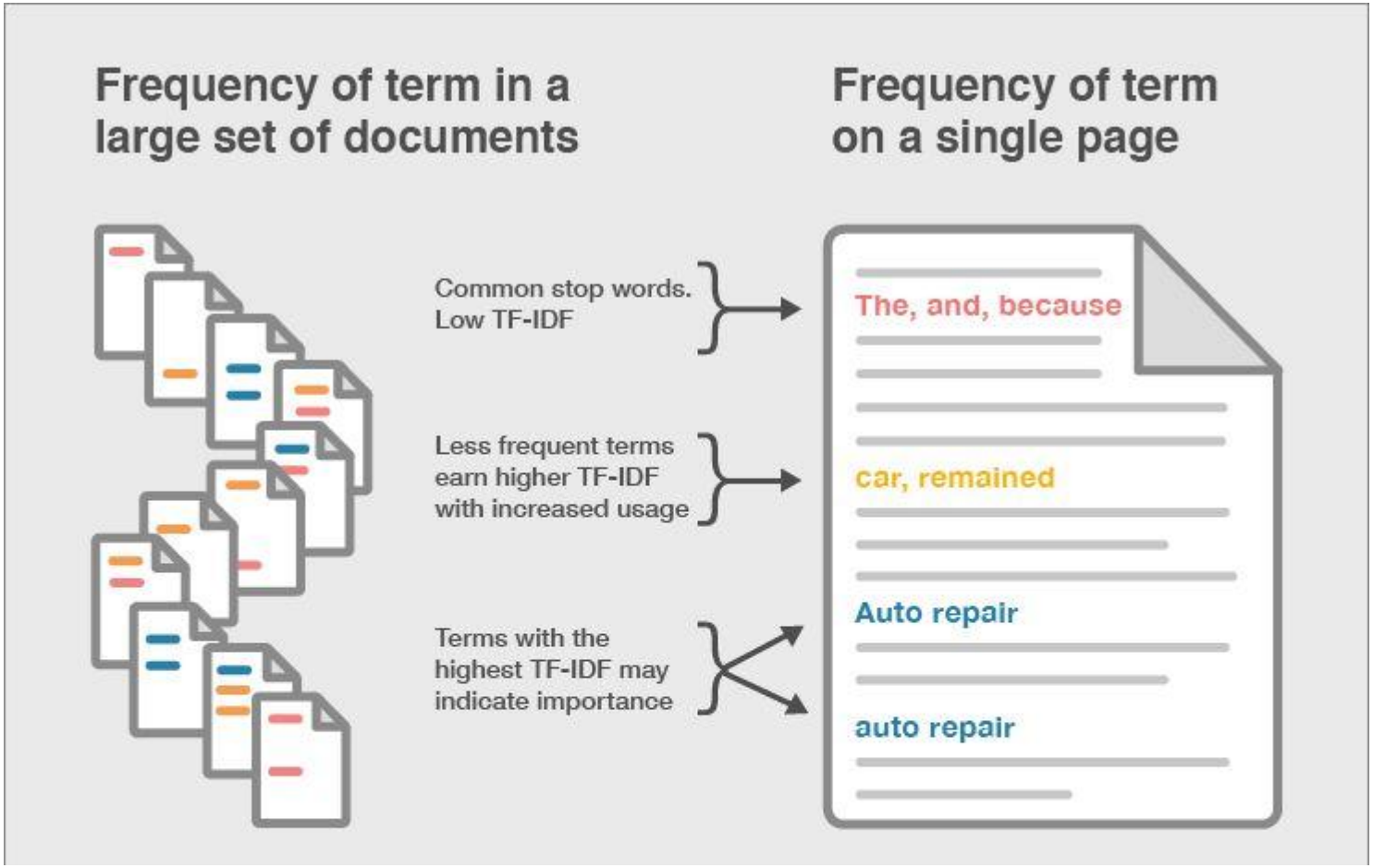
- **tf (term frequency in document):**

$$tf_t = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

- **idf (inverse document frequency):**

$$idf_t = log \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

Nutch - An overview



MinSpanCovOrd (Minimum Span Coverage Ordered)

Example of a document:

Search for "President of the United States of America"

The President of the United States of America is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The President is considered one of the world's most powerful people, leading the world's only contemporary superpower. The role includes being the commander-in-chief of the world's most expensive military with the largest nuclear arsenal and leading the largest economy by real and nominal GDP. The office of the president holds significant hard and soft power both in **of the United States** and abroad.

ARQUIVO.PT

MinSpanCovOrd (Minimum Span Coverage Unordered)

Example of a document:

Search for "President of the United States of America"

United States of America The President of the is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces.

The President is considered one of the world's most powerful people, leading the world's only contemporary superpower. The role includes being the commander-in-chief of the world's most expensive military with the largest nuclear arsenal and leading the largest economy by real and nominal GDP. The office of the president holds significant hard and soft power both in **of the United States** and abroad.

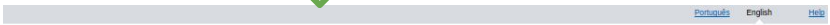
What is used for indexing ?

- ➔ Cluster with 17 slaves with and installation of Apache Hadoop-0.14.5;
- ➔ CentOS release 6.6 (Final)
- ➔ Red Hat Enterprise Linux Server release 5.5 (Tikanga) -Master

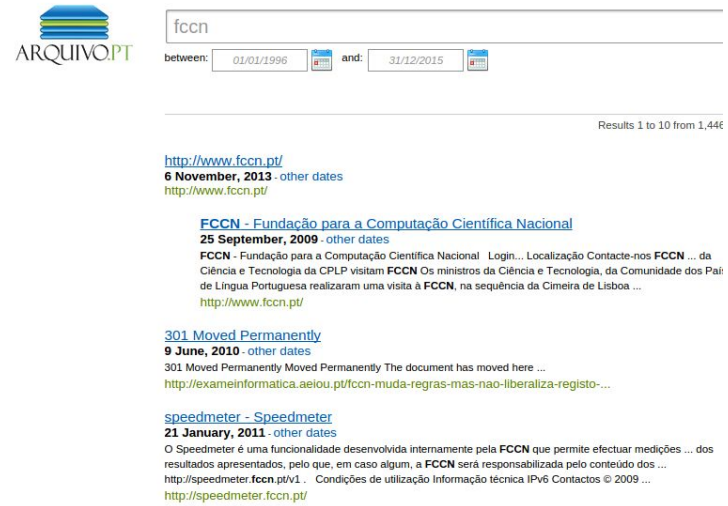
Wrapping up..

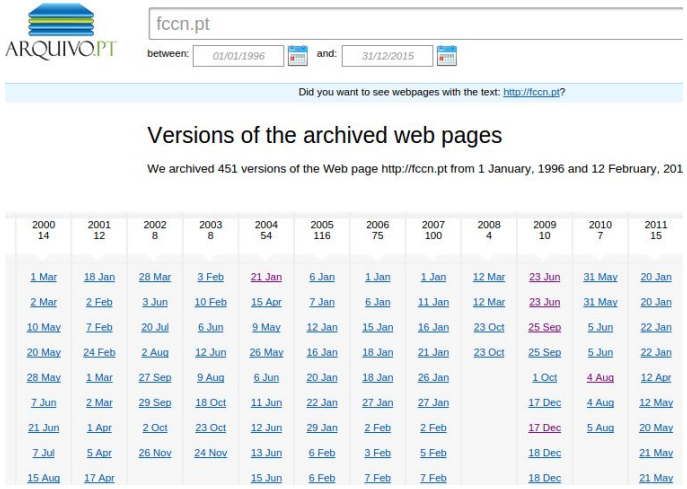
1º

2º

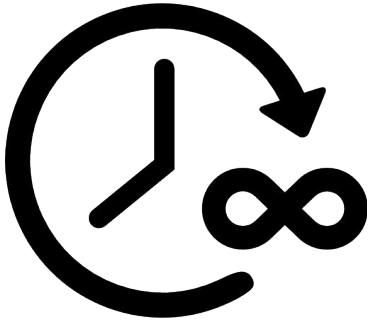


3º





ARQUIVO.PT



Future work

Future work

- Report about the feasibility of migrating Arquivo.pt search system to SOLR;v1
- Report about the feasibility of migrating Arquivo.pt search system to SOLR;v2

Want to know more?

- https://github.com/arquivo/pwa-technologies/wiki
- https://github.com/arquivo/pwa-technologies/blob/master/Report.pdf
- http://sobre.arquivo.pt/about-the-archive/publications-1/publications?set_language=en

ARQUIVO.PT

ARQUIVO.PT

Thank you!

hugo.viana@fccn.pt