

ARQUIVO.PT

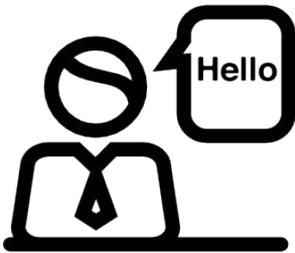
Architecture of the Portuguese Web Archive Search System

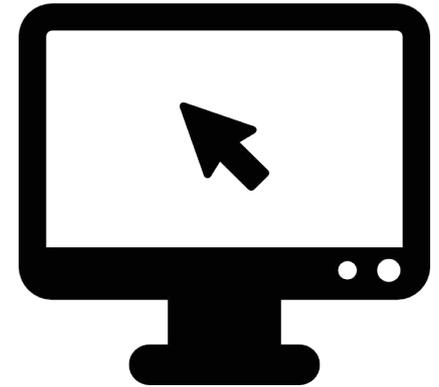
Hugo Viana

hugo.viana@fccn.pt

Presentation outline

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

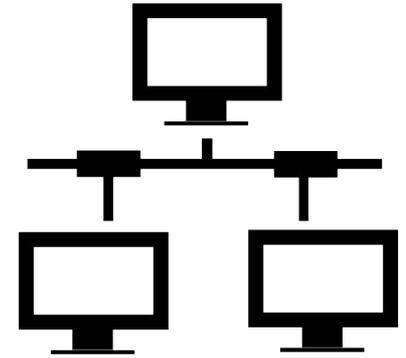




About Arquivo.pt: Searching with Arquivo.pt service

Who are we?!





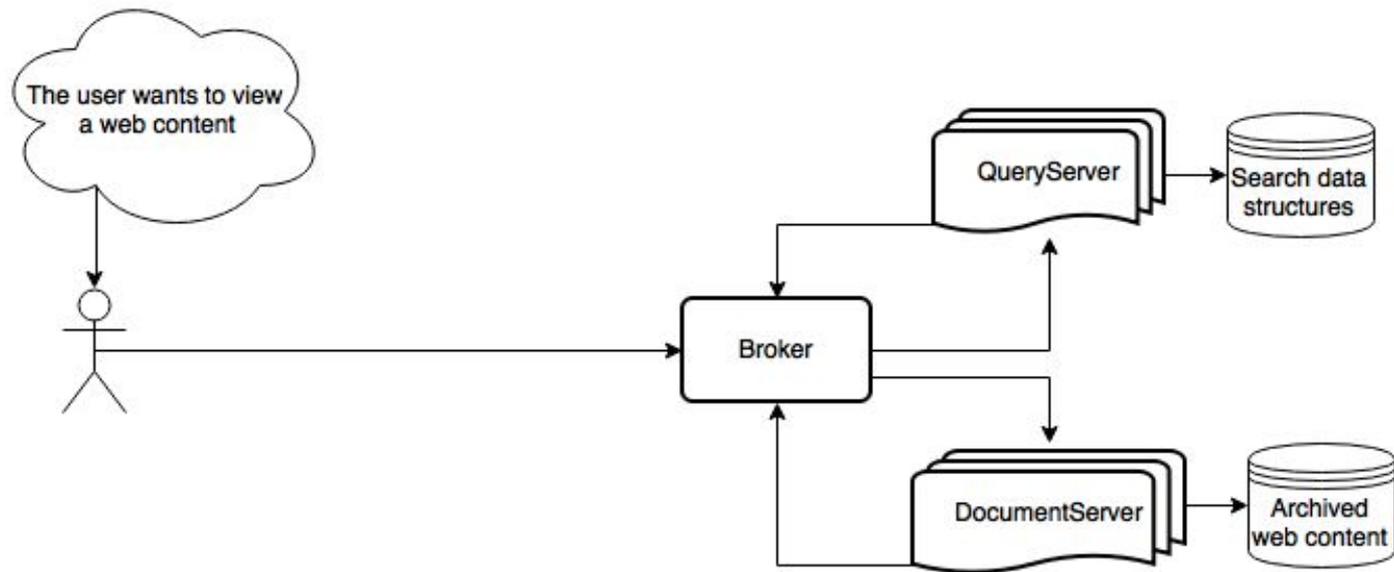
Architecture & Workflow: An overview

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 ..

Broker: Home page

[Português](#) [English](#) [Help](#)

 [Advanced search](#)

Search pages from the past

Millions of contents archived since 1996

[Meet the service](#)



Broker: Full-text search results



fccn

between:

01/01/1996



and:

31/12/2015



Results 1 to 10 from 1,446

<http://www.fccn.pt/>

6 November, 2013 - other dates

<http://www.fccn.pt/>

[FCCN - Fundação para a Computação Científica Nacional](#)

25 September, 2009 - other dates

FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos FCCN ... da Ciência e Tecnologia da CPLP visitam FCCN Os ministros da Ciência e Tecnologia, da Comunidade dos Países de Língua Portuguesa realizaram uma visita à FCCN, na sequência da Cimeira de Lisboa ...

<http://www.fccn.pt/>

[301 Moved Permanently](#)

9 June, 2010 - other dates

301 Moved Permanently Moved Permanently The document has moved here ...

<http://exameinformatica.aeiou.pt/fccn-muda-regras-mas-nao-liberaliza-registo-...>

[speedmeter - Speedmeter](#)

21 January, 2011 - other dates

O Speedmeter é uma funcionalidade desenvolvida internamente pela FCCN que permite efectuar medições ... dos resultados apresentados, pelo que, em caso algum, a FCCN será responsabilizada pelo conteúdo dos ...

<http://speedmeter.fccn.pt/v1> . Condições de utilização Informação técnica IPv6 Contactos © 2009 ...

<http://speedmeter.fccn.pt/>

Broker: Url search results



fccn.pt

between:

01/01/1996



and:

31/12/2015



Did you want to see webpages with the text: <http://fccn.pt?>

Versions of the archived web pages

We archived 451 versions of the Web page <http://fccn.pt> from 1 January, 1996 and 12 February, 2011

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

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/wayback/20141126195806/http://www.fccn.pt/pt/

Arquivo da Web Portuguesa - ligações exteriores, formulários e caixas de pesquisa poderão não funcionar corretamente. URL: http://www.fccn.pt/pt/ Data: 19:58:06 26 Novembro, 2014 [lescondet]

FCCN Outros sites da FCCN Utilizador não autenticado login Ainda não tenho conta. Preciso de ajuda. Google Pesquisa Personalizada PT EN

Está a usar IPv4. O seu endereço é: 193.136.192.166

Imprimir Enviar Partilhar Rss

FCCN FCT

Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

A FCCN Rede Académica Serviços Suporte a Utilizadores Eventos Imprensa

Centro Avançado de Ciber Defesa lança Portal para a comunidade ACDC

O ACDC (Advanced Cyber Security Center) criou um portal online desenhado para ajudar a comunidade na partilha de informação e soluções para acelerar processos de deteção de botnets e para agilizar a implementação de abordagens inovadoras, para prevenção e proteção de dispositivos móveis e fixos.
[Leia a notícia completa.](#)

09 10:00 - 16:30
Dez [Open Day STV](#)

04 Nov 2014
[Concurso para novo slogan b-on](#)

15 Out 2014
[Semana Internacional do Acesso Aberto 2014](#)

13 Out 2014
[DANTE e TERENA unem-se para criar a GÉANT Association](#)

1 2 3

b-on
Biblioteca on-line
[Entrar na b-on](#)
biblioteca do conhecimento online

Speedmeter
Teste a velocidade da sua rede
[Testar ligação](#)

Zappiens
Publique o seu vídeo
[Entrar no Zappiens](#)

Arquivo da Web Portuguesa
Pesquise no Passado
[Entrar no arquivo da web](#)

Investigadores e Professores do Ensino Superior

Estudantes do Ensino Superior
[Serviços e recursos para estudantes](#)

Público em Geral
[Serviços e recursos para o público em geral](#)

How all this flows?

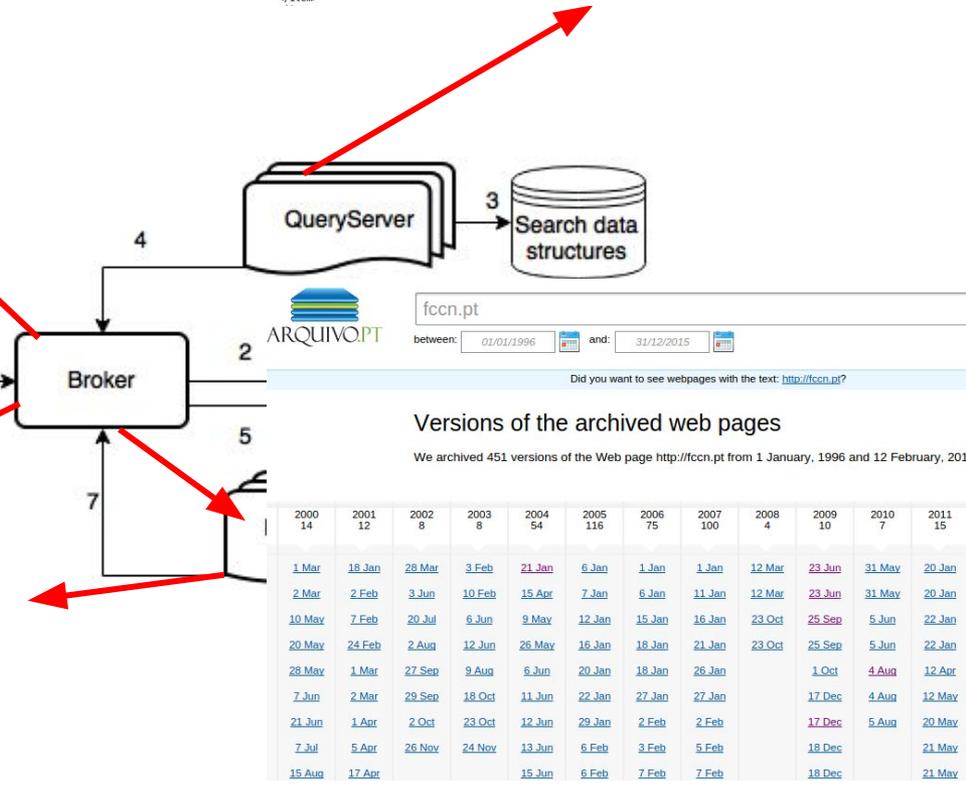
```

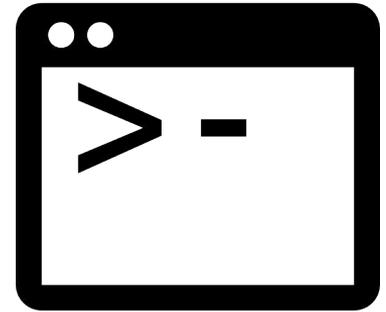
<description>PMA search results for query: fccn</description>
<link>http://arquivo.pt/</link>
<opensearch:totalResults>1446620</opensearch:totalResults>
<opensearch:startIndex>0</opensearch:startIndex>
<opensearch:itemsPerPage>10</opensearch:itemsPerPage>
<opensearch:Query role="request" searchTerms="fccn" 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:arccoffset>86899984</pwa:arccoffset>
<pwa:digest>4118c98f00b204e9809998c78427e</pwa:digest>
<pwa:timestamp>2013110622044000</pwa:timestamp>
<pwa:contentLength>417</pwa:contentLength>
<pwa:primaryType>text</pwa:primaryType>
<pwa:subType>html</pwa:subType>
</item>
<item>
<title>FCCN - Fundação para a Computação Científica Nacional</title>
<source url="http://www.fccn.pt/">Original URL of FC CN - Fundação para a Computação Científica Nacional</source>
<link>http://arquivo.pt/wayback/20090926073120/http://www.fccn.pt/</link>
<pwa:id>3335</pwa:id>
<pwa:index>22</pwa:index>
<pwa:arcname>IAH-20090925190110-05968-awp01.fccn.pt/<pwa:arcname>
<pwa:arccoffset>54764399</pwa:arccoffset>
<pwa:digest>3e53f640520963ca72d8e2cade29b9dd</pwa:digest>
<pwa:timestamp>20090925191844000</pwa:timestamp>
<pwa:contentLength>2523</pwa:contentLength>
<pwa:primaryType>text</pwa:primaryType>
<pwa:subType>html</pwa:subType>
</item>

```



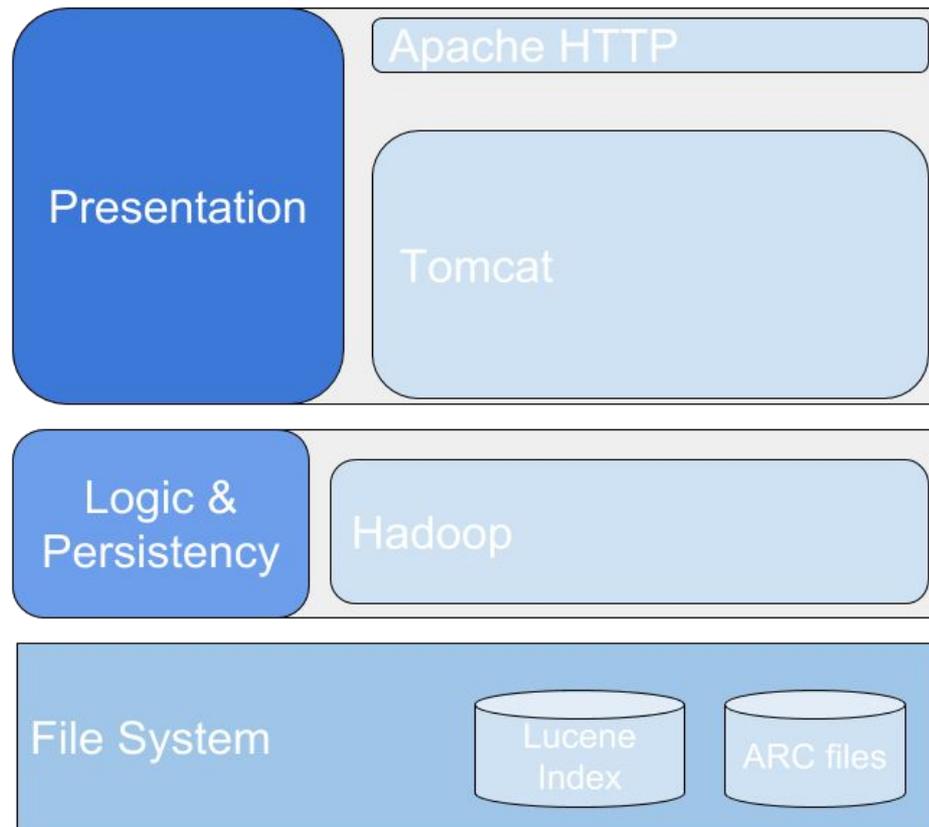
The user wants to view a web content

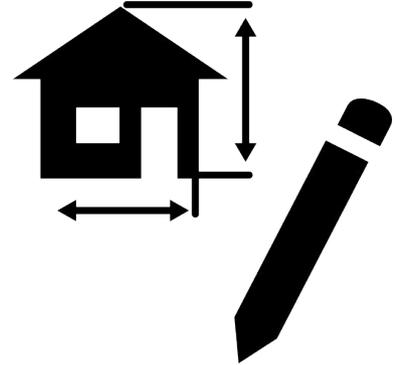




Workflows: Software components

Software components for the search system: workflow



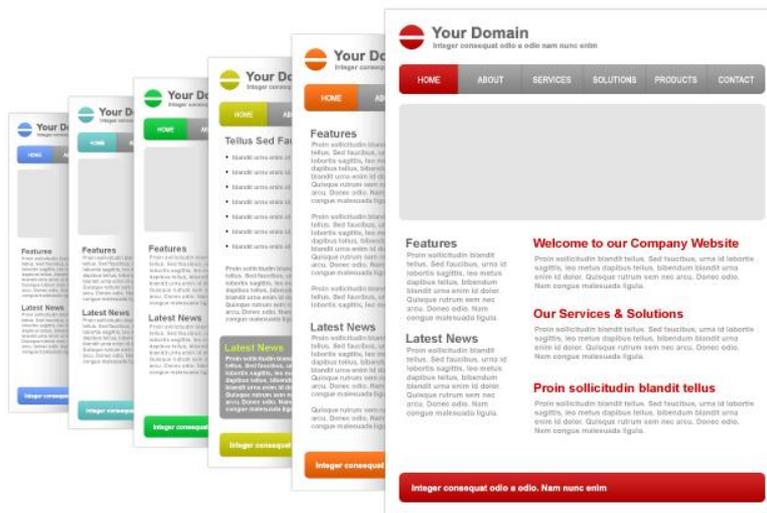


Operations Procedures

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

Invert index: What is this?

INDEX

- ABC, 164, 321*n*
academic journals, 262, 280–82
Adobe eBook Reader, 148–53
advertising, 36, 45–46, 127, 145–46, 167–68, 321*n*
Africa, medications for HIV patients in, 257–61
Agee, Michael, 223–24, 225
agricultural patents, 313*n*
Aibo robotic dog, 153–55, 156, 157, 160
AIDS medications, 257–60
air traffic, land ownership vs., 1–3
Akerlof, George, 232
Alben, Alex, 100–104, 105, 198–99, 295, 317*n*
alcohol prohibition, 200
Alice's Adventures in Wonderland (Carroll), 152–53
Anello, Douglas, 60
animated cartoons, 21–24
antiretroviral drugs, 257–61
Apple Corporation, 203, 264, 302
architecture, constraint effected through, 122, 123, 124, 318*n*
archive.org, 112
see also Internet Archive
archives, digital, 108–15, 173, 222, 226–27
Aristotle, 150
Armstrong, Edwin Howard, 3–6, 184, 196
Arrow, Kenneth, 232
art, underground, 186
artists:
 publicity rights on images of, 317*n*
 recording industry payments to, 52, 58–59, 74, 195, 196–97, 199, 301, 329*n*–30*n*

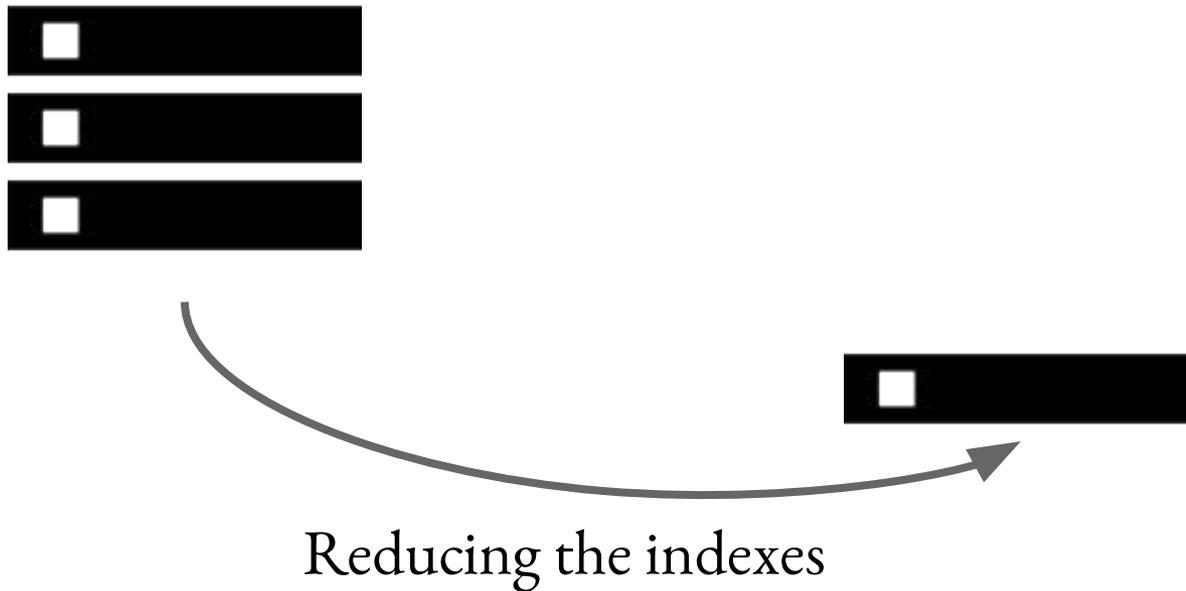
Invert index: An example

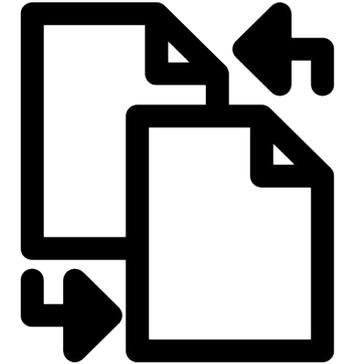
ID	Text
1	Baseball is played during summer months.
2	Summer is the time for picnics here.
3	Months later we found out why.
4	Why is summer so hot here
↑	Sample document data

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.





Ranking features used in the 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

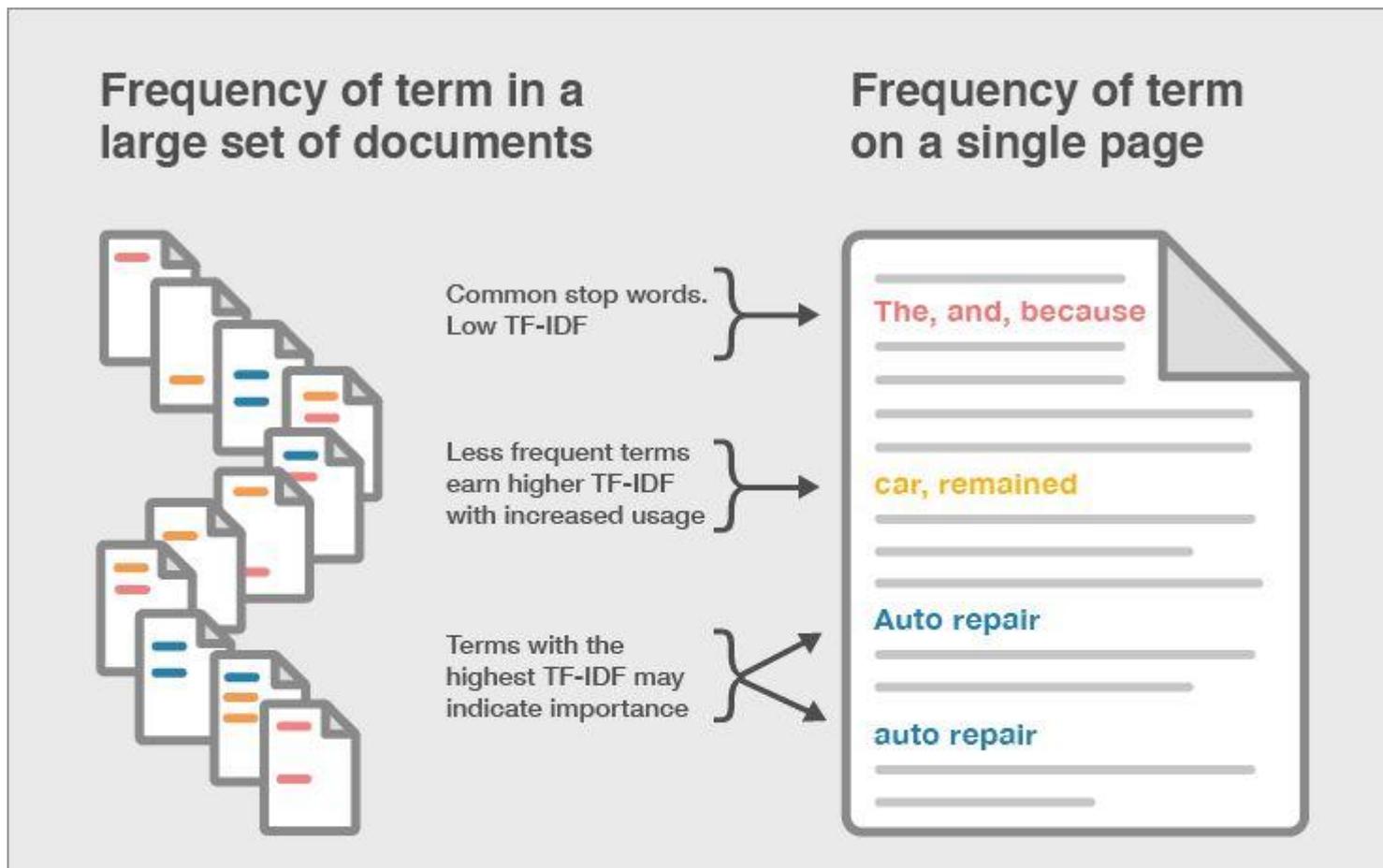
- **tf (term frequency in document):**

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

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

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

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.

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..

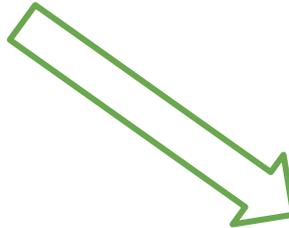
1º



2º



3º



fcfn

between: and:

Results 1 to 10 from 1,446

<http://www.fcfn.pt/>
6 November, 2013 - other dates
<http://www.fcfn.pt/>

FCCN - Fundação para a Computação Científica Nacional
25 September, 2009 - other dates

FCCN - Fundação para a Computação Científica Nacional. Login... Localização Contacte-nos FCCN ... da Ciência e Tecnologia da CPLP visitam FCCN Os ministros da Ciência e Tecnologia, da Comunidade dos Países de Língua Portuguesa realizaram uma visita à FCCN, na sequência da Cimeira de Lisboa ...
<http://www.fcfn.pt/>

301 Moved Permanently
9 June, 2010 - other dates

301 Moved Permanently Moved Permanently The document has moved here ...
<http://exameinformatica.aeiou.pt/fccn-muda-regras-mas-nao-liberaliza-registo-...>

speedmeter - Speedmeter
21 January, 2011 - other dates

O Speedmeter é uma funcionalidade desenvolvida internamente pela FCCN que permite efectuar medições ... dos resultados apresentados, pelo que, em caso algum, a FCCN será responsabilizada pelo conteúdo dos ...
<http://speedmeter.fcfn.pt/v1> Condições de utilização Informação técnica IPv6 Contactos © 2009 ...
<http://speedmeter.fcfn.pt/>



Type word or URL Search

Search pages from the past

Millions of contents archived since 1996
[Meet the service](#)



fcfn.pt

between: and:

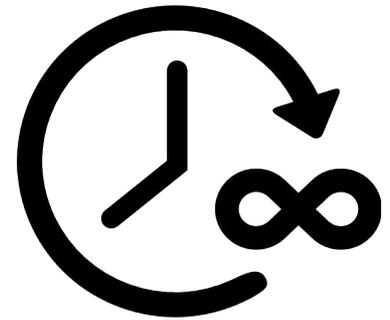
Did you want to see webpages with the text: <http://fcfn.pt/>?

Versions of the archived web pages

We archived 451 versions of the Web page <http://fcfn.pt> from 1 January, 1996 and 12 February, 2011

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





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

Thank you!

hugo.viana@fccn.pt