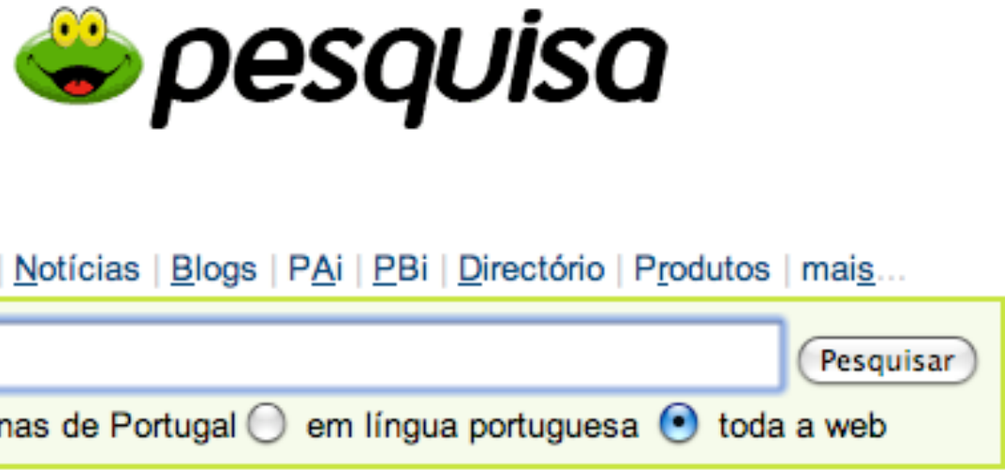
Motores de busca

João Pedro Gonçalves
jp@sapo.pt
Abril 2010

A Pesquisa do SAPO

- 50 Queries por Segundo (QPS).
- Acordo com Microsoft – Bing
- Indexador FAST com cerca de 100 milhões de páginas indexadas.



Microsoft bing™

fast

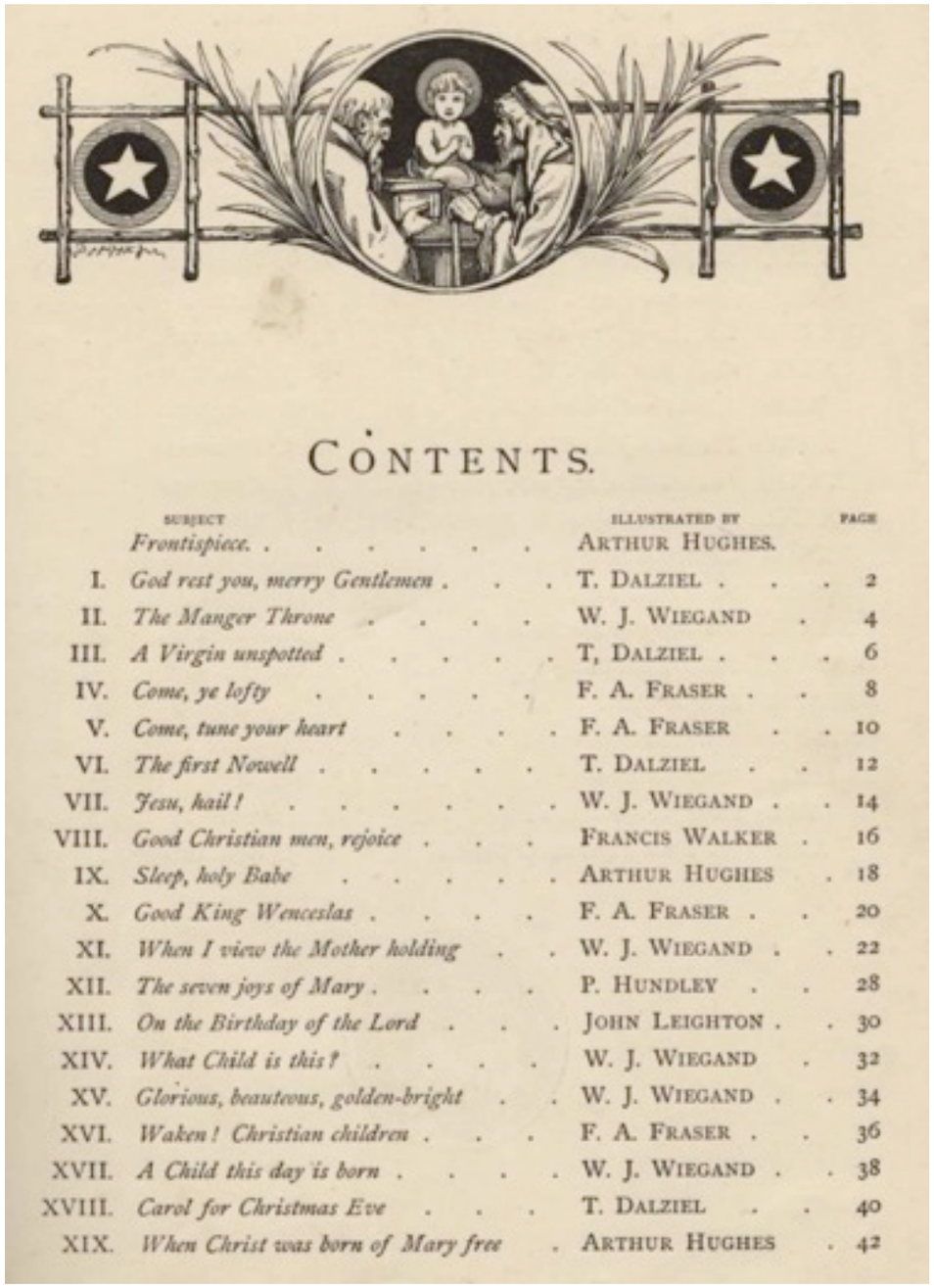
Solr

A Pesquisa do SAPO

- Frontend em mod_perl
- Perl, Python, C e C++
- Equipa de 14 programadores nas áreas de Pesquisa, Anúncios contextualizados e Directório

Directórios

por categorias

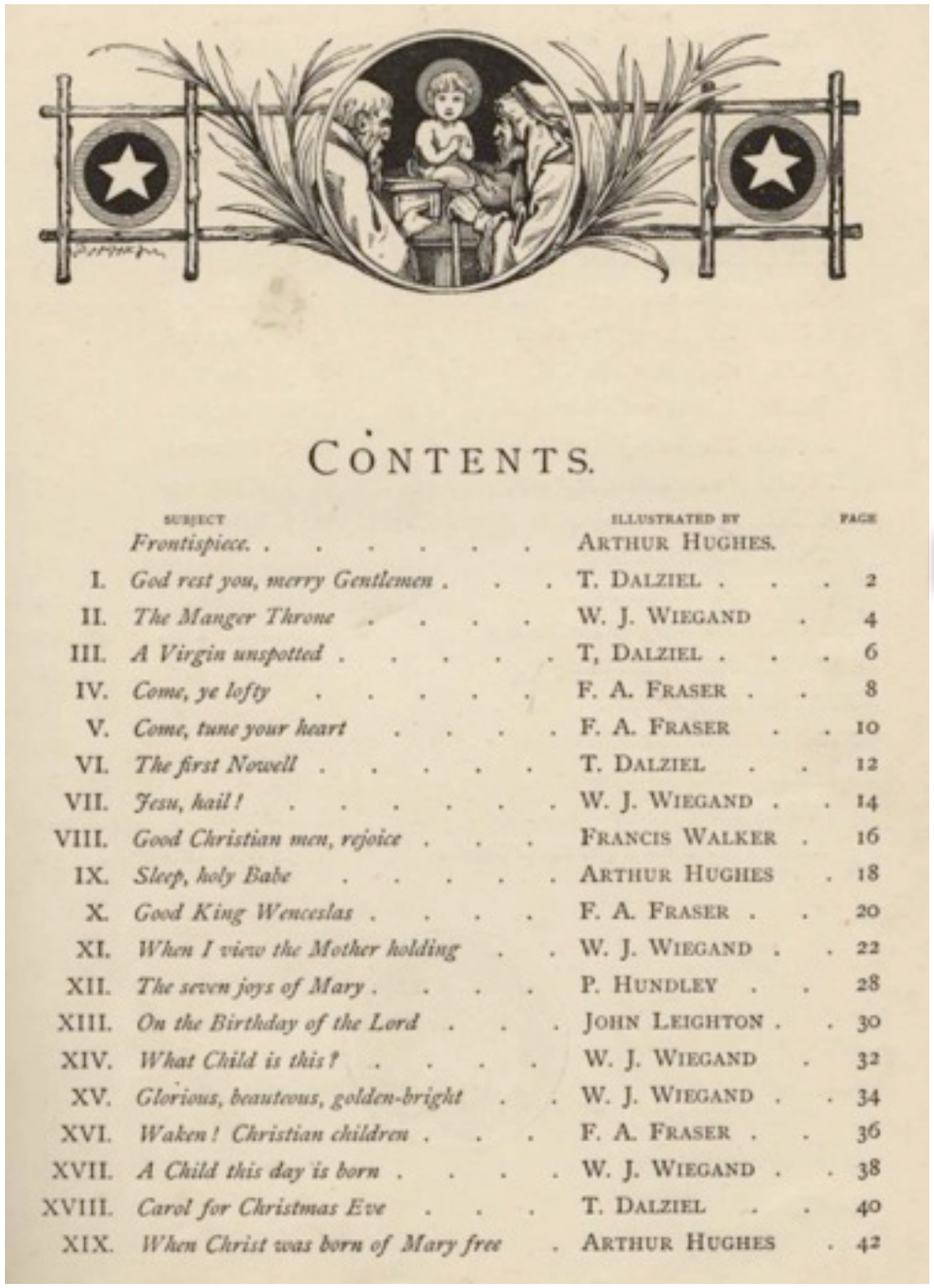


CONTENTS.

	SUBJECT	ILLUSTRATED BY	PAGE
	Frontispiece.	ARTHUR HUGHES.	
I.	*God rest you, merry Gentlemen* . . .	T. DALZIEL . . .	2
II.	*The Manger Throne*	W. J. WIEGAND .	4
III.	*A Virgin unspotted*	T, DALZIEL . . .	6
IV.	*Come, ye lofty*	F. A. FRASER . .	8
V.	*Come, tune your heart*	F. A. FRASER . .	10
VI.	*The first Nowell*	T. DALZIEL . .	12
VII.	*Jesu, hail!*	W. J. WIEGAND . .	14
VIII.	*Good Christian men, rejoice* . . .	FRANCIS WALKER .	16
IX.	*Sleep, holy Babe*	ARTHUR HUGHES .	18
X.	*Good King Wenceslas*	F. A. FRASER . .	20
XI.	*When I view the Mother holding* . .	W. J. WIEGAND . .	22
XII.	*The seven joys of Mary*	P. HUNDLEY . .	28
XIII.	*On the Birthday of the Lord* . . .	JOHN LEIGHTON . .	30
XIV.	*What Child is this?*	W. J. WIEGAND .	32
XV.	*Glorious, beauteous, golden-bright* . .	W. J. WIEGAND . .	34
XVI.	*Waken! Christian children* . . .	F. A. FRASER . .	36
XVII.	*A Child this day is born*	W. J. WIEGAND . .	38
XVIII.	*Carol for Christmas Eve* . . .	T. DALZIEL . .	40
XIX.	*When Christ was born of Mary free* .	ARTHUR HUGHES .	42

Directórios

por categorias



CONTENTS.

	SUBJECT	ILLUSTRATED BY	PAGE
	Frontispiece.	ARTHUR HUGHES.	
I.	*God rest you, merry Gentlemen*	T. DALZIEL	2
II.	*The Manger Throne*	W. J. WIEGAND	4
III.	*A Virgin unspotted*	T, DALZIEL	6
IV.	*Come, ye lofty*	F. A. FRASER	8
V.	*Come, tune your heart*	F. A. FRASER	10
VI.	*The first Nowell*	T. DALZIEL	12
VII.	*Jesu, hail!*	W. J. WIEGAND	14
VIII.	*Good Christian men, rejoice*	FRANCIS WALKER	16
IX.	*Sleep, holy Babe*	ARTHUR HUGHES	18
X.	*Good King Wenceslas*	F. A. FRASER	20
XI.	*When I view the Mother holding*	W. J. WIEGAND	22
XII.	*The seven joys of Mary*	P. HUNDLEY	28
XIII.	*On the Birthday of the Lord*	JOHN LEIGHTON	30
XIV.	*What Child is this?*	W. J. WIEGAND	32
XV.	*Glorious, beauteous, golden-bright*	W. J. WIEGAND	34
XVI.	*Waken! Christian children*	F. A. FRASER	36
XVII.	*A Child this day is born*	W. J. WIEGAND	38
XVIII.	*Carol for Christmas Eve*	T. DALZIEL	40
XIX.	*When Christ was born of Mary free*	ARTHUR HUGHES	42

pesquisa

Web | Imagens | Notícias | Blogs | PAi | PBi | **Directório** | Produtos | mais...

Pesquisar

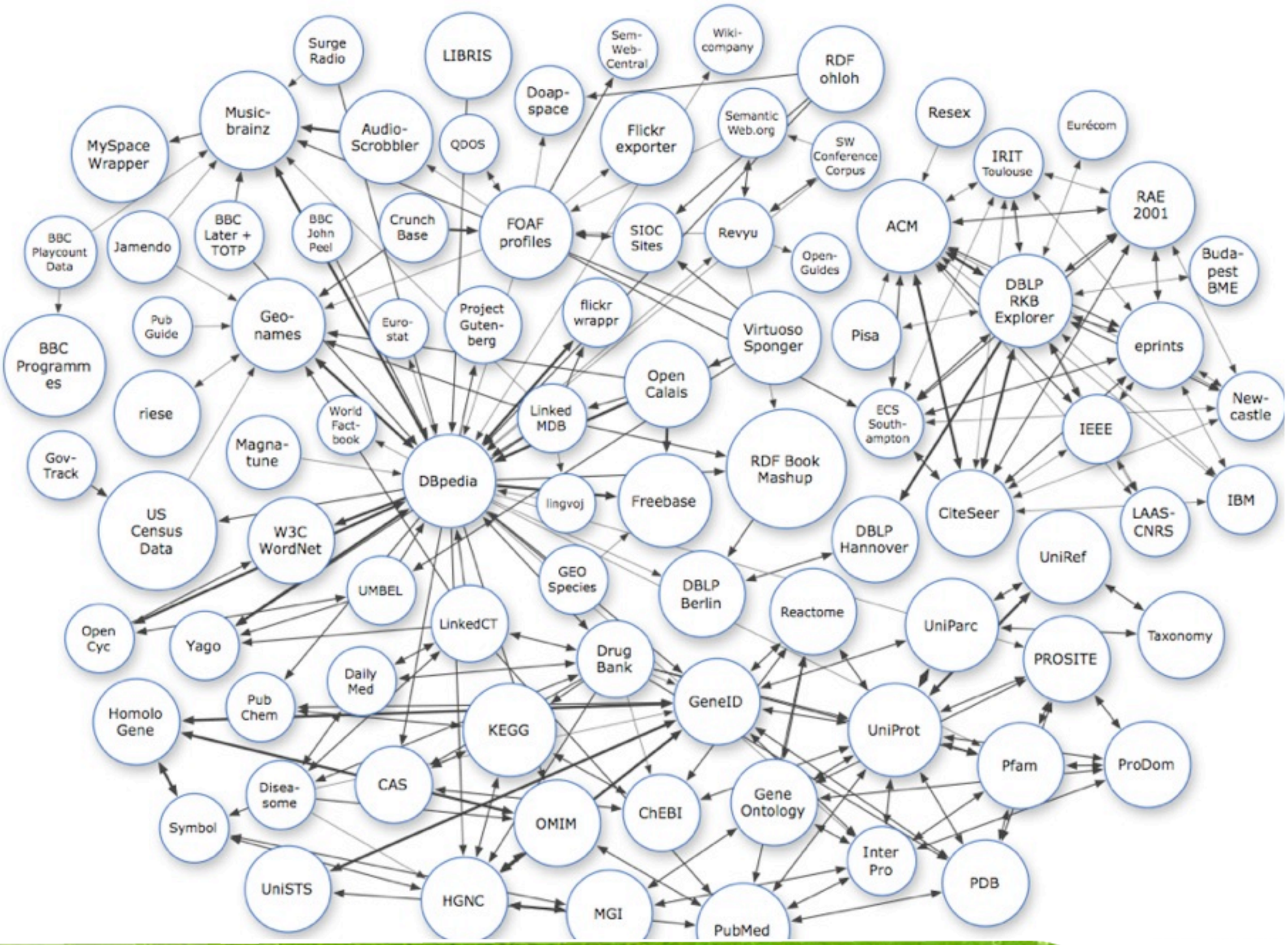
Portal SAPO | Registe o seu site | Anuncie na Pesquisa | Blog | Ajuda?

Directório

- Arte e cultura
  Cinema, Teatro, Fotografia
- Ciência
  Astronomia, Gestão, Medicina
- Comunicação Social
  Jornais, Televisões, Revistas
- Desporto
  Futebol, Motorizados, Aquáticos
- Economia e Negócios
  Imobiliário, Emprego, Comércio
- Educação
  Superior, Profissional, E-learning
- Entretenimento e Lazer
  Cinema, Desporto, Humor
- Estado e Administração
  Governo, ONGs, UE
- Nacional e Regional
  Lisboa, Porto, Faro
- Saúde
  Hospitais, Farmácias, Crianças
- Sociedade
  Política, Religião, Minorias
- Tecnologia e Internet
  Mail, Software, Biotecnologia

PT.COM © 2007

sapo.pt


Surge Radio
LIBRIS
Sem-Web-Central
Wiki-company
RDF ohloh
Music-brainz
Audio-Scrobbler
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
Resex
Eurécom
IRIT Toulouse
MySpace Wrapper
QDOS
RAE 2001
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
Buda-pest BME
BBC Programmes
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
eprints
riese
World Fact-book
Linked MDB
Open Calais
ECS South-ampton
IEEE
New-castle
Gov-Track
Magna-tune
DBpedia
RDF Book Mashup
US Census Data
W3C WordNet
lingvoj
Freebase
DBLP Hannover
CiteSeer
LAAS-CNRS
IBM
UMBEL
GEO Species
DBLP Berlin
Reactome
UniRef
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
GeneID
UniProt
Homolo Gene
KEGG
Pfam
ProDom
Disea-some
CAS
Gene Ontology
Symbol
OMIM
ChEBI
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB

WolframAlpha computational knowledge engine

portugal gdp

Input interpretation: *Mathematica form*

Portugal | GDP

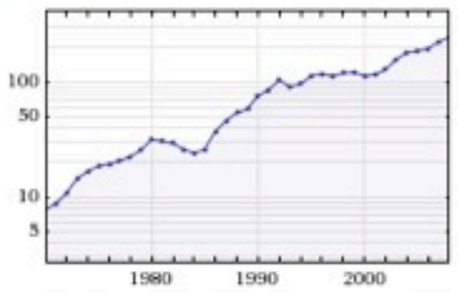
Result:

$243.5 billion per year (US dollars per year) (2008 estimate)

Local currency conversion: More

€ 178.1 billion per year (euros per year) (at current quoted rate)

GDP history: Linear scale



(from 1970 to 2008)
(in billions of US dollars per year)
(log scale)

Economic properties: More

GDP at exchange rate	**$243.5 billion per year** (2008) (world rank: 38th)
GDP at parity	**$237.3 billion per year** (2008) (world rank: 47th)
GDP in local currency	**€ 166.2 billion** (2008)
GDP per capita	**$22 805.48 per person** (2008) (world rank: 55th)
GDP real growth	**−0.04466% per year**
inflation rate	**1.914% per year**
unemployment rate	**7.6%** (2008) (world rank: 106th)

Units »

Computed by: Wolfram Mathematica Source information » Download as: PDF | Live Mathematica

WolframAlpha computational knowledge engine

portugal gdp vs italy vs greece vs spain

Input interpretation: *Mathematica form*

Portugal, Italy, Greece, Spain | GDP

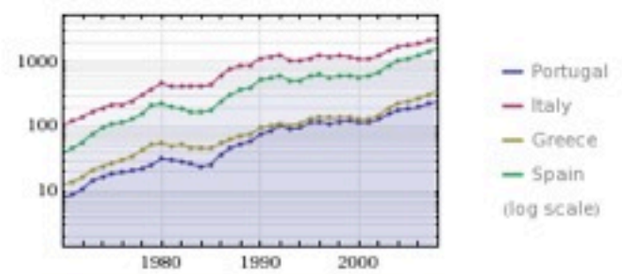
Results:

Portugal	**$243.5 billion per year** (US dollars per year) (2008 estimate)
Italy	**$2.303 trillion per year** (US dollars per year) (2008 estimate)
Greece	**$355.9 billion per year** (US dollars per year) (2008 estimate)
Spain	**$1.604 trillion per year** (US dollars per year) (2008 estimate)

Ranked values: Reverse

		visual	ratios	
1	Italy		9.458	1
2	Spain		6.588	0.6966
3	Greece		1.462	0.1545
4	Portugal		1	0.1057

GDP history: Linear scale



(from 1970 to 2008) (in billions of US dollars per year)

Computed by: Wolfram Mathematica Source information » Download as: PDF | Live Mathematica

sapo.pt

WolframAlpha computational... knowledge engine

100,000 characters in Portuguese

Input interpretation:

100 000 characters (in Portuguese)

Approximate word count:

16 260 words

(assuming 5.15 characters per word and 1 space between each word)

Printed length: More | Weight, thickness | Show assumptions

single–spaced document	33 pages	1476 lines
double–spaced document	66 pages	1476 lines

Times: More

typical typing	5 hours
typical speaking	110 minutes
silent reading	59 minutes

Data size:

100 kB (kilobytes) | 800 kb (kilobits)

(assuming 8–bit encoding)

Language translation lengths:

English	94 000 characters
French	100 000 characters
German	100 000 characters
Spanish	100 000 characters
Italian	110 000 characters
Mandarin Chinese	50 000 characters

(estimates based on typical translations)

Computed by: Wolfram Mathematica Source information » Download as: PDF | Live Mathematica

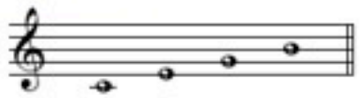
WolframAlpha computational... knowledge engine

C major 7th chord

Input interpretation:

C major seventh chord

Music notation: Play notes

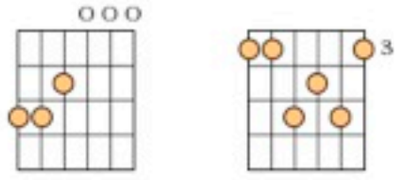


Note names:

C | E | G | B

Keyboard display: Play chord

Guitar chord voicings: More



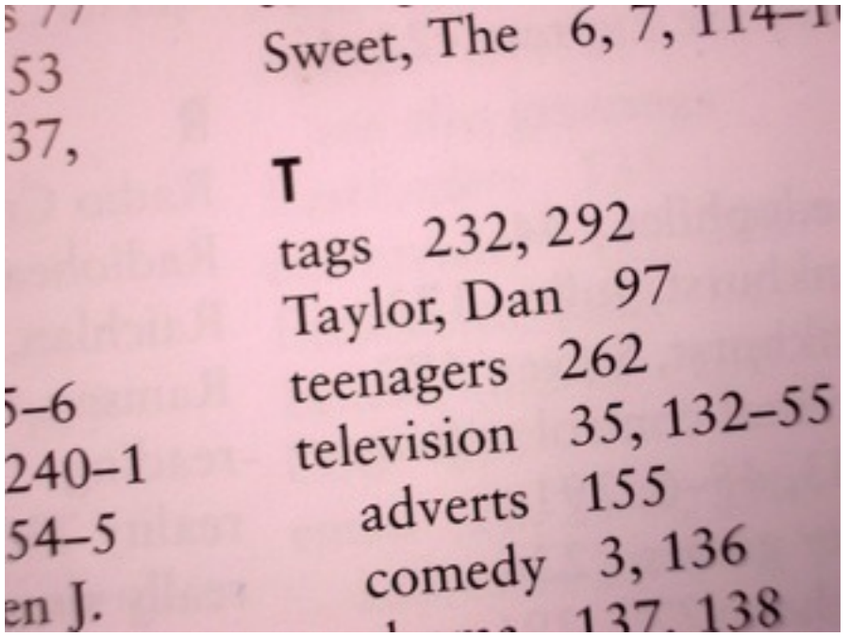
Computed by: Wolfram Mathematica Download as: PDF | Live Mathematica

sapo.pt

Wednesday, June 2, 2010

Pesquisar

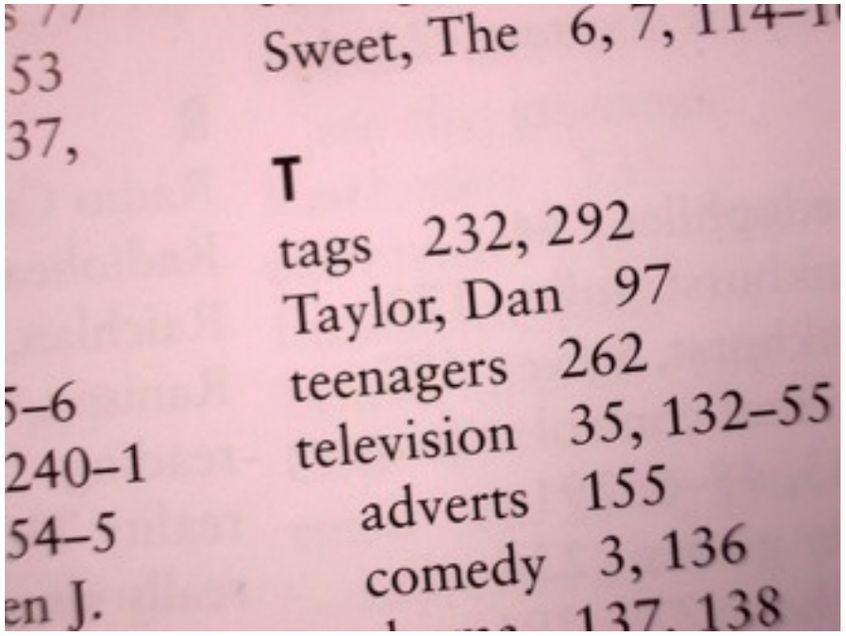
por índice invertido



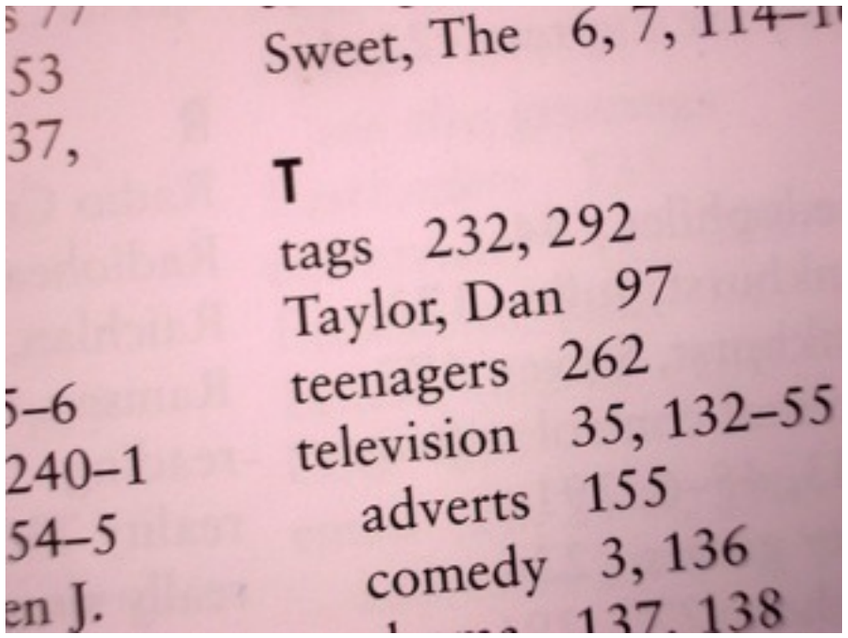
sapo.pt

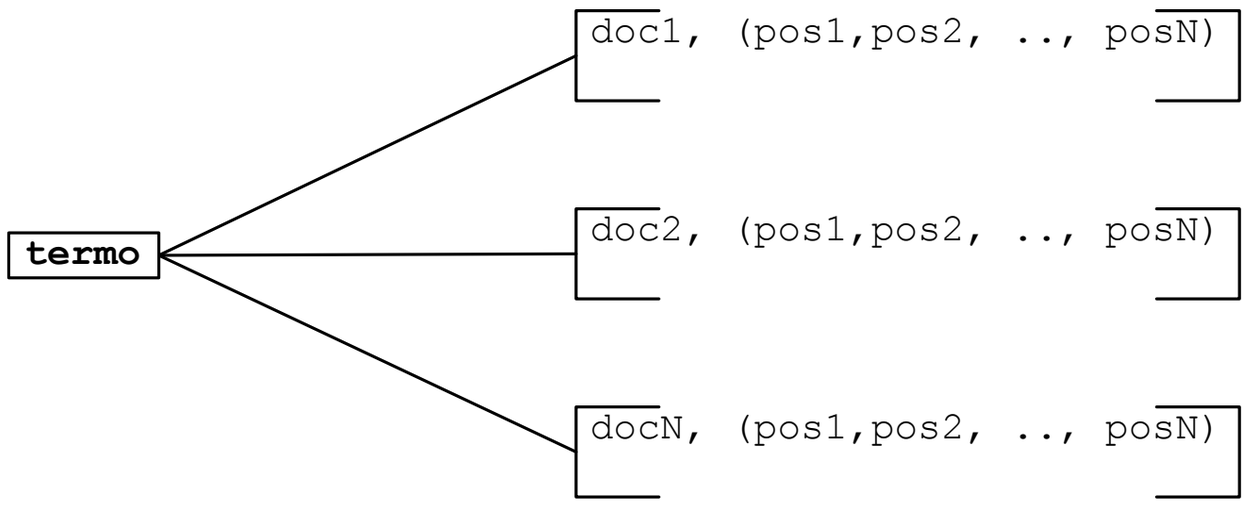
Pesquisar

por índice invertido



sapo.pt





Encontrar

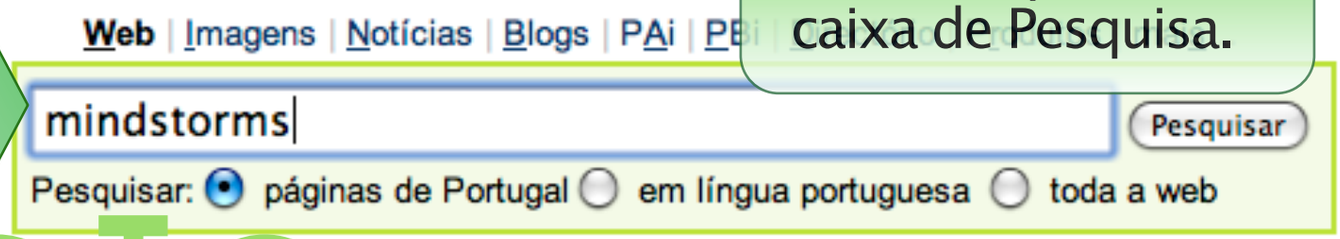
ciclo Search/Find/Obtain

Encontrar

ciclo Search/Find/Obtain

pesquisa

Acesso rápido à caixa de Pesquisa.



Tempo de carregamento da páginas de destino está fora de controle do motor de busca.

Resultado pretendido rápido de encontrar na página.

sapo.pt

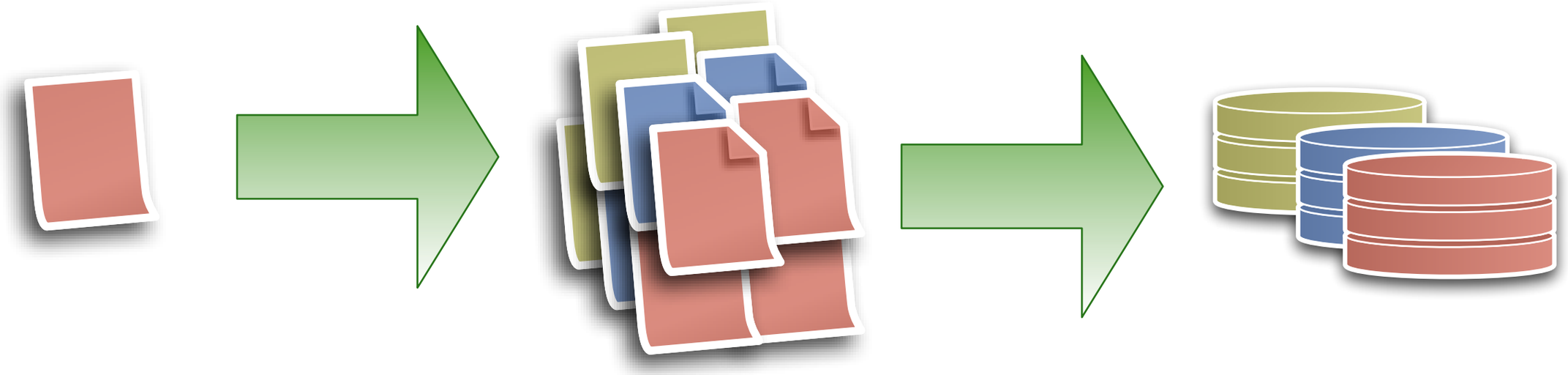
Volume da Informação - Desafios

Volume da Informação – Desafios

~15 KB
por documento

100
milhões de páginas

1,4 Tb de Indice de
Pesquisa

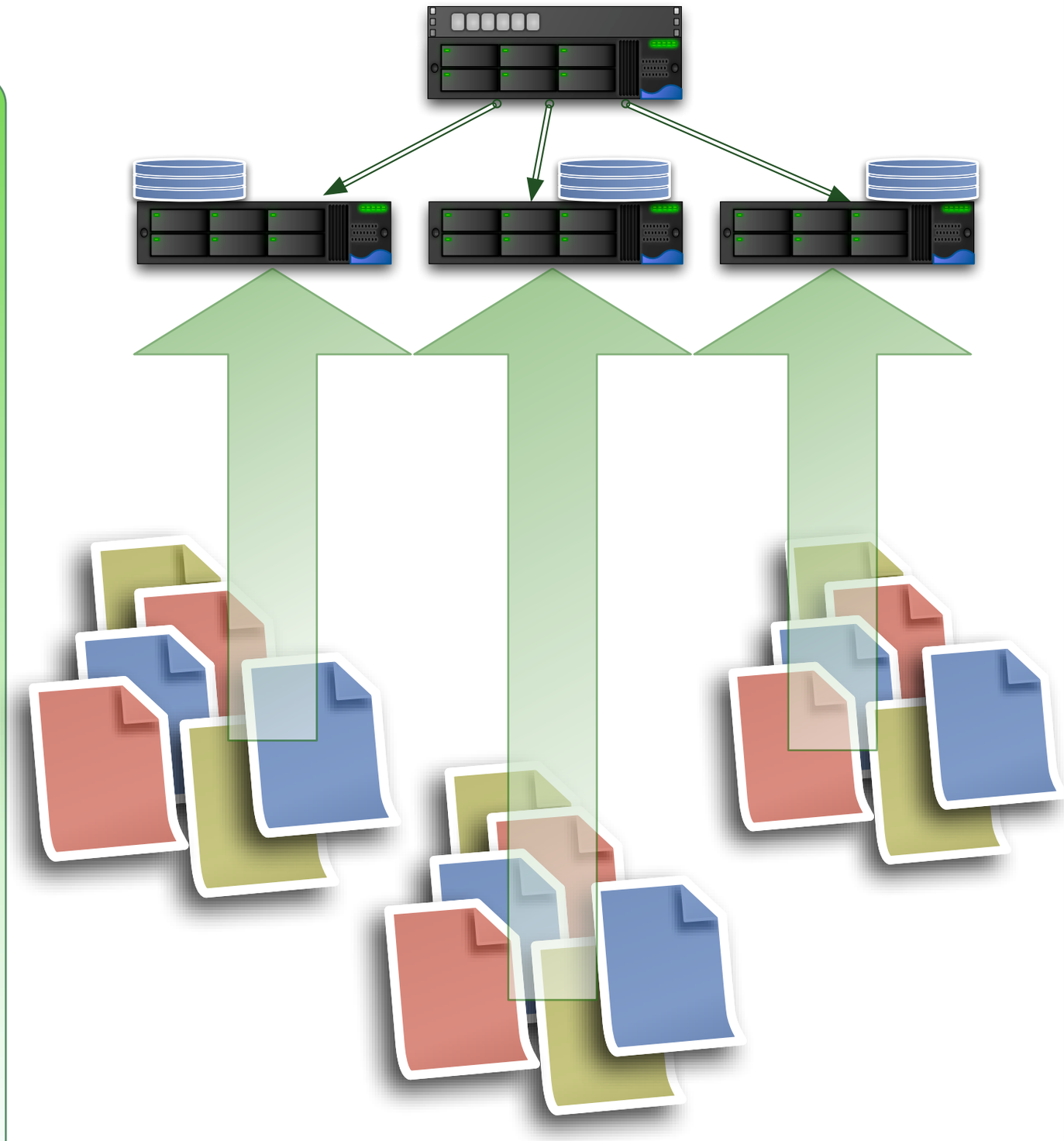


No caso Português, garantir 50 Pesquisas por segundo, cada uma em menos de 500ms por Pesquisa, com um Índice de 1,4 Terabyte.

Crawlers

Crawlers

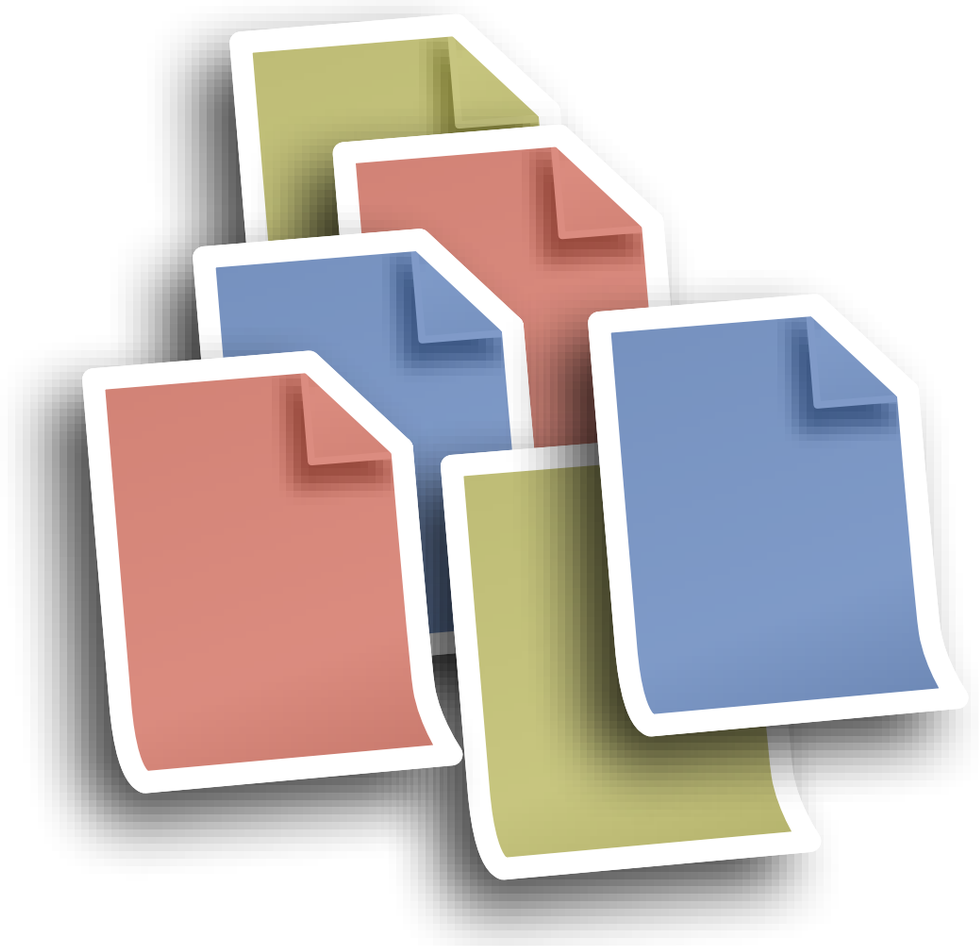
- Percorrem as páginas que irão ser indexadas.
- Limite de pedidos por site.
- Vários sites em simultâneo, optimizado para a largura de banda disponível.
- Crawling distribuído.
- Master distribui tarefas, cada nó é responsável pelo seu armazenamento.



Document Pipeline

Document Pipeline

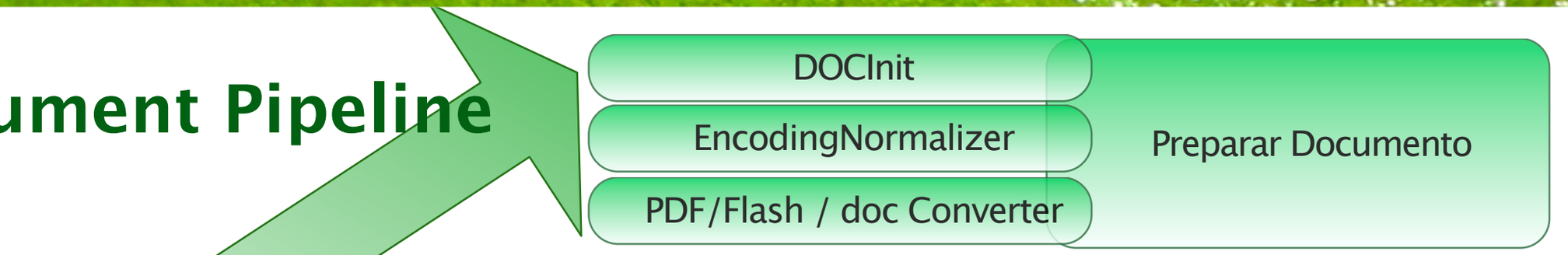
- Conteúdo dos crawlers é enviado para a Document Pipeline.
- Envio em batch dos documentos.
- Cada url é um objecto Document.
- No início do processo, apenas sabemos a url e o conteúdo do documento não processado.



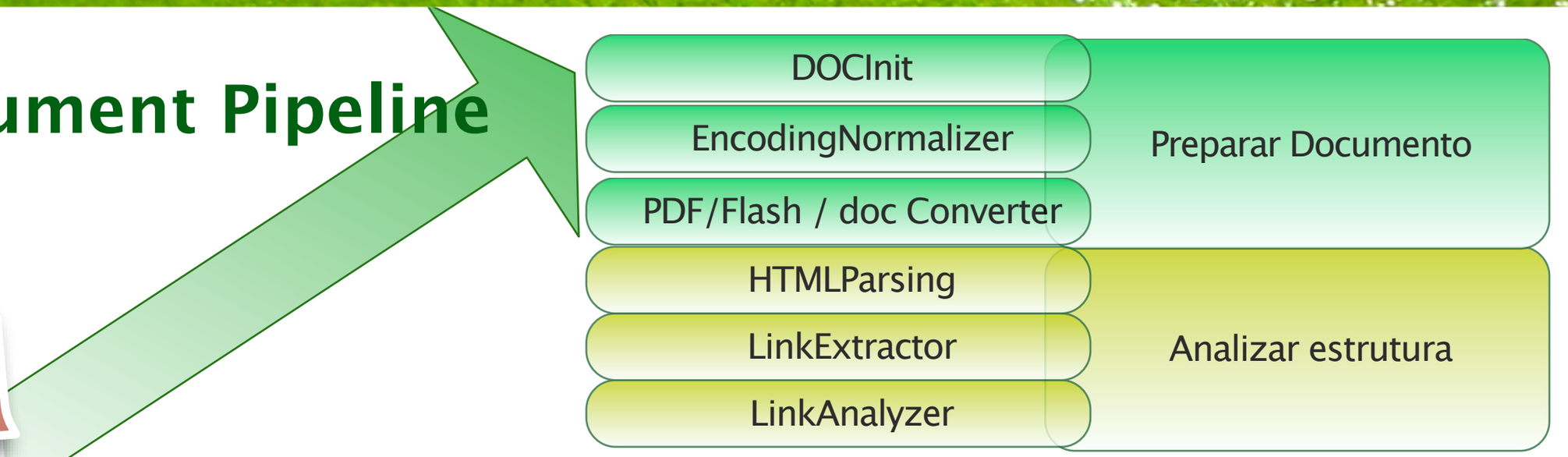
sapo.pt

Document Pipeline

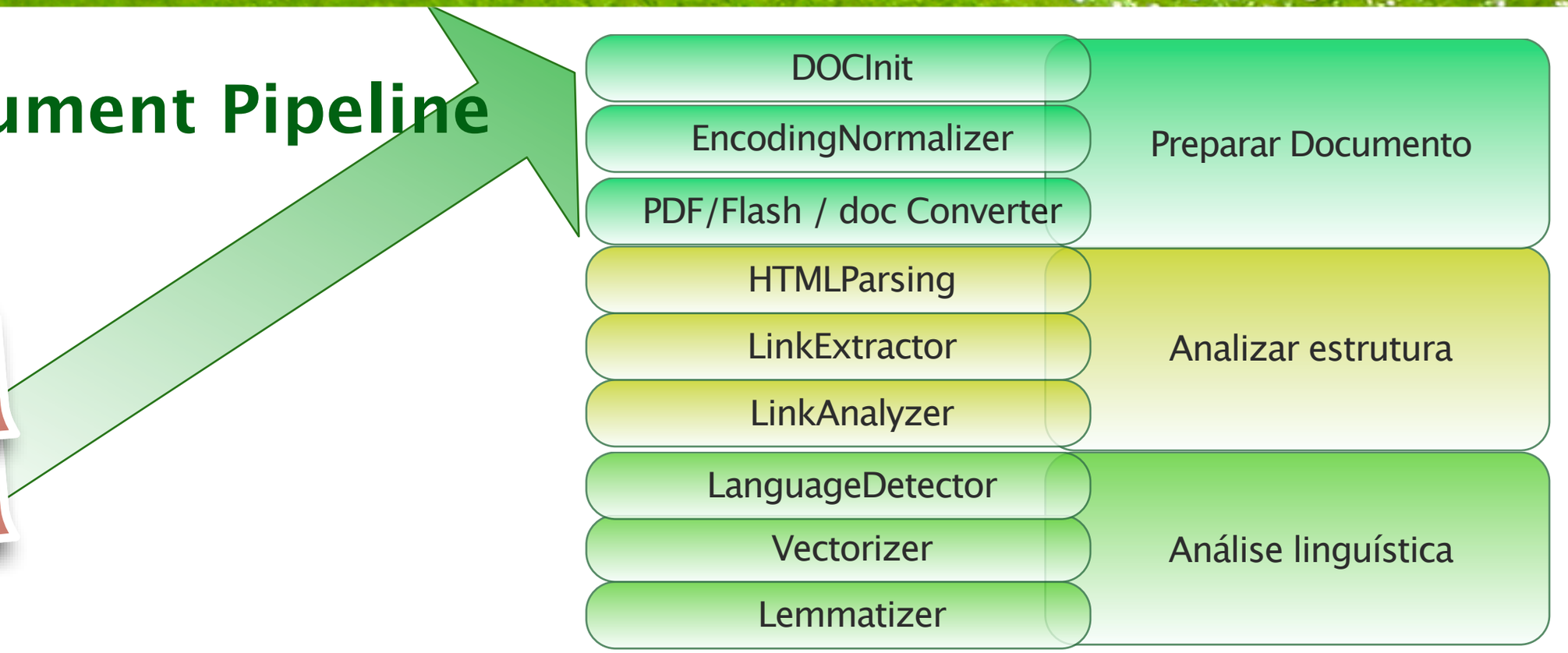
Document Pipeline



Document Pipeline



Document Pipeline



Document Pipeline

- DOCInit
- EncodingNormalizer
- PDF/Flash / doc Converter

Preparar Documento

- HTMLParsing
- LinkExtractor
- LinkAnalyzer

Analizar estrutura

- LanguageDetector
- Vectorizer
- Lemmatizer

Análise linguística

- GetLocations
- GetPersonNames
- GetDateAndTime

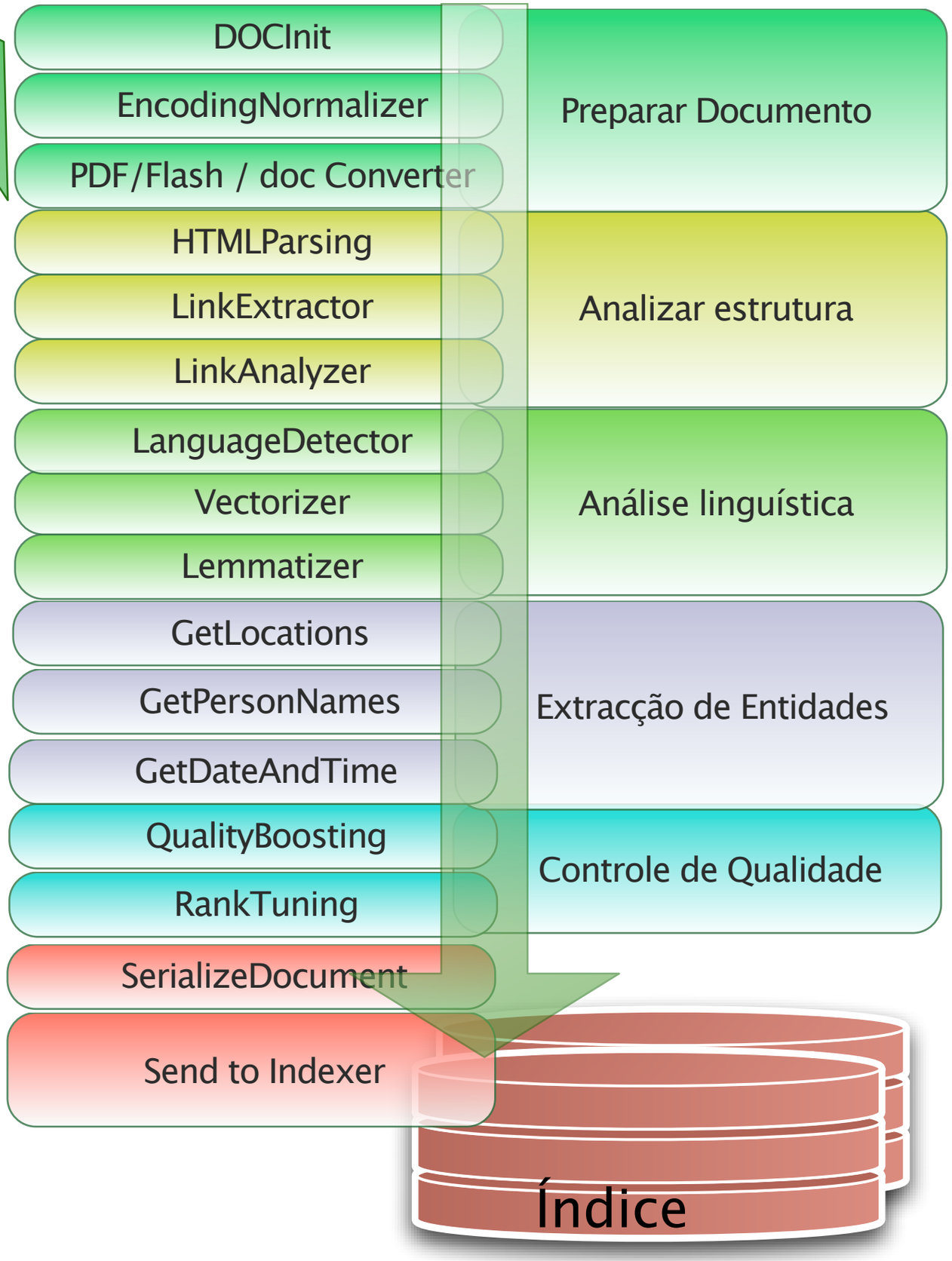
Extracção de Entidades

Document Pipeline

- **Preparar Documento**
  - DOCInit
  - EncodingNormalizer
  - PDF/Flash / doc Converter
- **Analizar estrutura**
  - HTMLParsing
  - LinkExtractor
  - LinkAnalyzer
- **Análise linguística**
  - LanguageDetector
  - Vectorizer
  - Lemmatizer
- **Extracção de Entidades**
  - GetLocations
  - GetPersonNames
  - GetDateAndTime
- **Controle de Qualidade**
  - QualityBoosting
  - RankTuning

Document Pipeline

- Cada componente na cadeia é um módulo isolado que altera o objecto Document.
- Módulos desenvolvidos em Python.



sapo.pt

Nokia, TMN e Sapo continuam a liderar preferências

Publicado por Casa dos Bits há 20 horas e 27 minutos | 0 comentários

Nokia, TMN, Sapo, HP e Canon lideram, cada uma na sua categoria, as preferências dos portugueses, segundo o estudo "Marcas de Confiança" de 2010, das Selecções do Reader's Digest, relativamente à área das tecnologias.

Na décima edição da análise, a Nokia mantém-se em primeiro (68%), em clara vantagem face às marcas que ocupam a segunda e terceira posições, a Samsung (13%) e a Sony Ericsson (9%). Relativamente a 2009, as marcas de confiança são as mesmas, mas as percentagens eram outras - 73, 8,5 e seis por cento, respectivamente.

A TMN continua a ser o operador de rede móvel mais popular, com um primeiro lugar reforçado, já que 49 por cento dos inquiridos votaram na marca, face aos 45 por cento que o tinham feito em 2009. Segue-se a Vodafone com 36 por cento e a Optimus, com 13 por cento.

Na categoria de Empresas de Serviço de Internet, o Sapo conseguiu superar os resultados do ano anterior (38%), somando 45 por cento dos votos. A Zon Multimédia perdeu três por cento relativamente a 2009, apresentando agora 14 por cento, enquanto o MEO ganhou o terceiro lugar, com oito por cento, destronando a Cabovisão.

Na área dos computadores pessoais, a Hewlett-Packard viu a confiança na marca reforçada, consolidando o primeiro lugar com 35 por cento, mais cinco por cento do que em 2009. A Toshiba, com 17 por cento, e a Asus, com 12 por cento, mantêm, respectivamente, a segunda e a terceira posições nesta categoria.

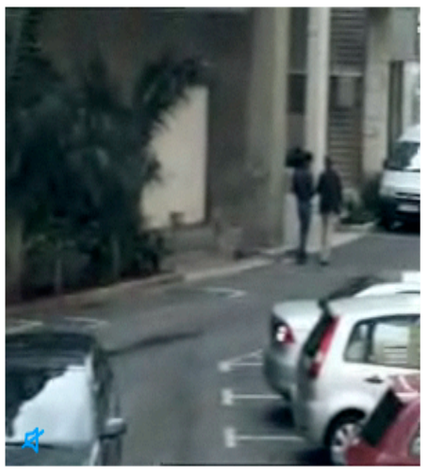
A Canon (34%), a Sony (24%) e a Olimpus (12%) são, por esta ordem, as três marcas de confiança dos portugueses na categoria Máquinas Fotográficas.

A análise "Marcas de Confiança" das Selecções do Reader's Digest foi promovida em 16 países, entre os quais Alemanha, Áustria, Bélgica, Espanha, Finlândia, França, Holanda, Hungria e Polónia. Em Portugal, este estudo foi realizado através de questionário postal endereçado a 12.200 assinantes da revista Selecções do Reader's Digest.

Comentar este artigo »

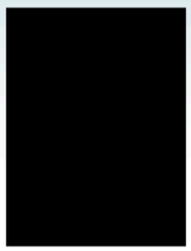
Últimas Em destaque Mais comentados

- Vídeos do MyWay na Reuters
- Portugal em 6º em vírus nos PCs
- Android chega às 50 mil aplicações
- Anti-vírus comparados
- 5 perguntas a... Paulo Costa
- TomTom anuncia novidades



Sugestões Montra

Sugestão TeK: Vai uma boleia para chegar ao destino?



Poupar dinheiro, ter companhia e uma atitude ambientalmente mais responsável são as consequências mais directas de partilhar o carro com terceiros. Mas a alternativa também pode ser uma saída em dias de greve.

Ler artigo

Outras sugestões »

Title

Author, Date

Keywords

Body

Nokia, TMN e Sapo continuam a liderar preferências

Publicado por Casa dos Bits há 20 horas e 27 minutos | 0 comentários

Nokia, TMN, Sapo, HP e Canon lideram, cada uma na sua categoria, as preferências dos portugueses, segundo o estudo "Marcas de Confiança" de 2010, das Selecções do Reader's Digest, relativamente à área das tecnologias.

Na décima edição da análise, a Nokia mantém-se em primeiro (68%), em clara vantagem face às marcas que ocupam a segunda e terceira posições, a Samsung (13%) e a Sony Ericsson (9%). Relativamente a 2009, as marcas de confiança são as mesmas, mas as percentagens eram outras - 73, 8,5 e seis por cento, respectivamente.

A TMN continua a ser o operador de rede móvel mais popular, com um primeiro lugar reforçado, já que 49 por cento dos inquiridos votaram na marca, face aos 45 por cento que o tinham feito em 2009. Segue-se a Vodafone com 36 por cento e a Optimus, com 13 por cento.

Na categoria de Empresas de Serviço de Internet, o Sapo conseguiu superar os resultados do ano anterior (38%), somando 45 por cento dos votos. A Zon Multimédia perdeu três por cento relativamente a 2009, apresentando agora 14 por cento, enquanto o MEO ganhou o terceiro lugar, com oito por cento, destronando a Cabovisão.

Na área dos computadores pessoais, a Hewlett-Packard viu a confiança na marca reforçada, consolidando o primeiro lugar com 35 por cento, mais cinco por cento do que em 2009. A Toshiba, com 17 por cento, e a Asus, com 12 por cento, mantêm, respectivamente, a segunda e a terceira posições nesta categoria.

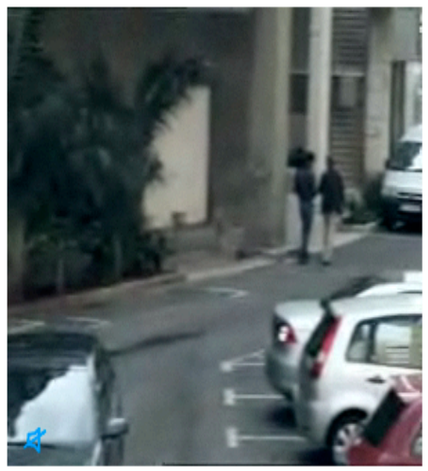
A Canon (34%), a Sony (24%) e a Olimpus (12%) são, por esta ordem, as três marcas de confiança dos portugueses na categoria Máquinas Fotográficas.

A análise "Marcas de Confiança" das Selecções do Reader's Digest foi promovida em 16 países, entre os quais Alemanha, Áustria, Bélgica, Espanha, Finlândia, França, Holanda, Hungria e Polónia. Em Portugal, este estudo foi realizado através de questionário postal endereçado a 12.200 assinantes da revista Selecções do Reader's Digest.

Comentar este artigo »

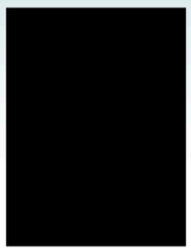
Últimas Em destaque Mais comentados

- Vídeos do MyWay na Reuters
- Portugal em 6º em vírus nos PCs
- Android chega às 50 mil aplicações
- Anti-vírus comparados
- 5 perguntas a... Paulo Costa
- TomTom anuncia novidades



Sugestões Montra

Sugestão TeK: Vai uma boleia para chegar ao destino?



Poupar dinheiro, ter companhia e uma atitude ambientalmente mais responsável são as consequências mais directas de partilhar o carro com terceiros. Mas a alternativa também pode ser uma saída em dias de greve.

Ler artigo

Outras sugestões »

Indexação

```
<field name="title" fullsort="yes" tokenize="auto" lemmatize="yes" >
  <vectorize default="10:0" />
</field>
<field name="body" tokenize="auto" max-result-size="1024"
       fallback-ref="teaser" result="dynamic" index="no" lemmatize="yes">
  <vectorize default="5:5" alternative="{ja,ko,zh,szh,tzh}:5:0" />
</field>
<field name="teaser" index="no" />
<field name="description" element-name="meta_description" result="no" tokenize="auto" />
<field name="anchortext" result="no" tokenize="auto" />
<field name="keywords" element-name="meta_keywords_off" result="no" tokenize="auto" />
```

- Cada índice é gerado localmente.
- Perfil do Índice permite definir quais os campos mais relevantes e qual o peso.

Indexação

```
<context weight="50">
  <field-weight field-ref="body" value="5" />
  <field-weight field-ref="description" value="30" />
  <field-weight field-ref="urlkeywords" value="40" />
  <field-weight field-ref="keywords" value="50" />
  <field-weight field-ref="title" value="60" />
</context>
```

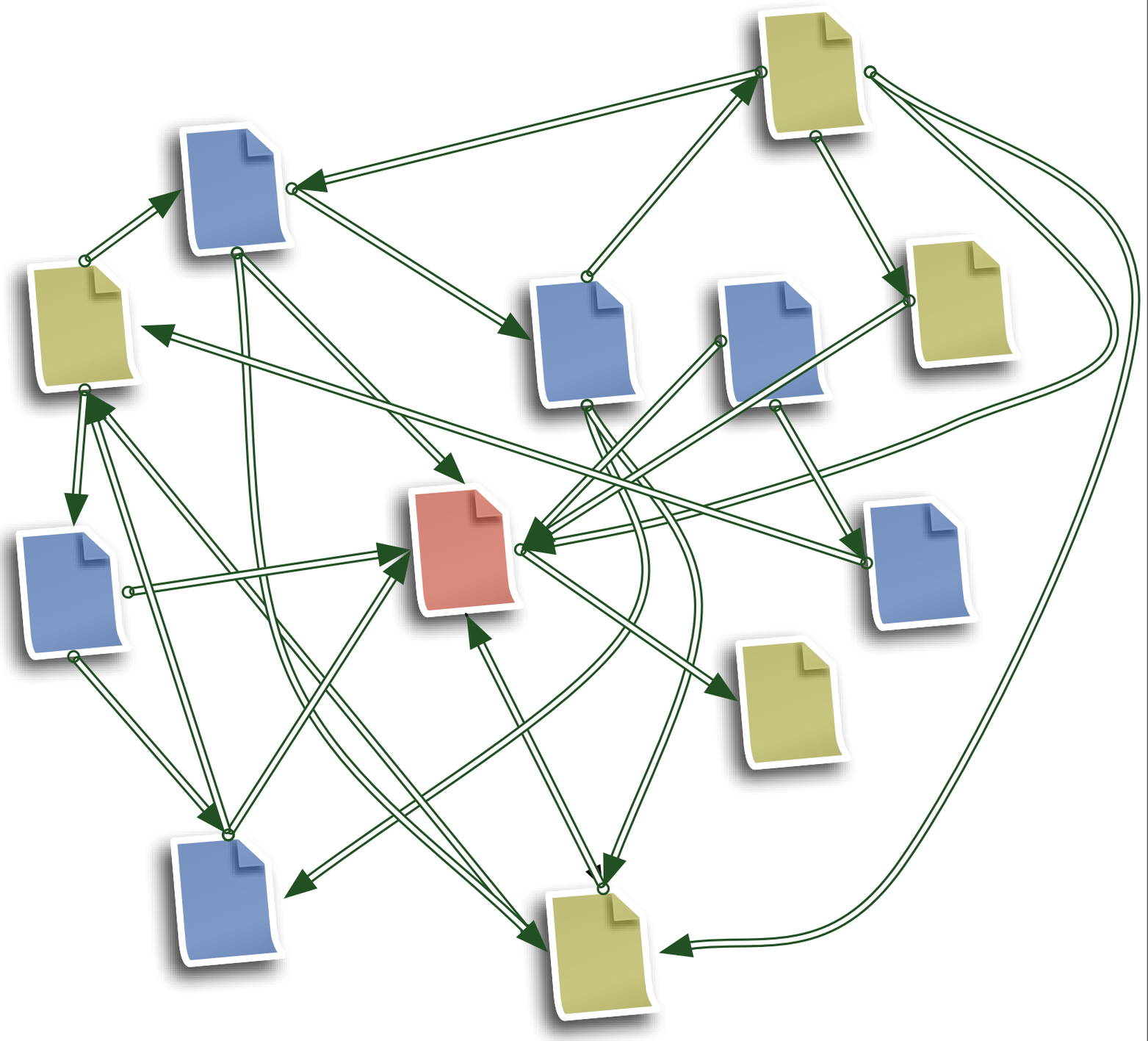
- Cada índice é gerado localmente.
- Perfil do Índice permite definir quais os campos mais relevantes e qual o peso.

Confiar

PageRank

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1$$

- Análise estrutural da Web.
- Autoridade nem sempre qualidade
- Comportamento dos utilizadores ajuda no ranking.
- Problemas com Spam.

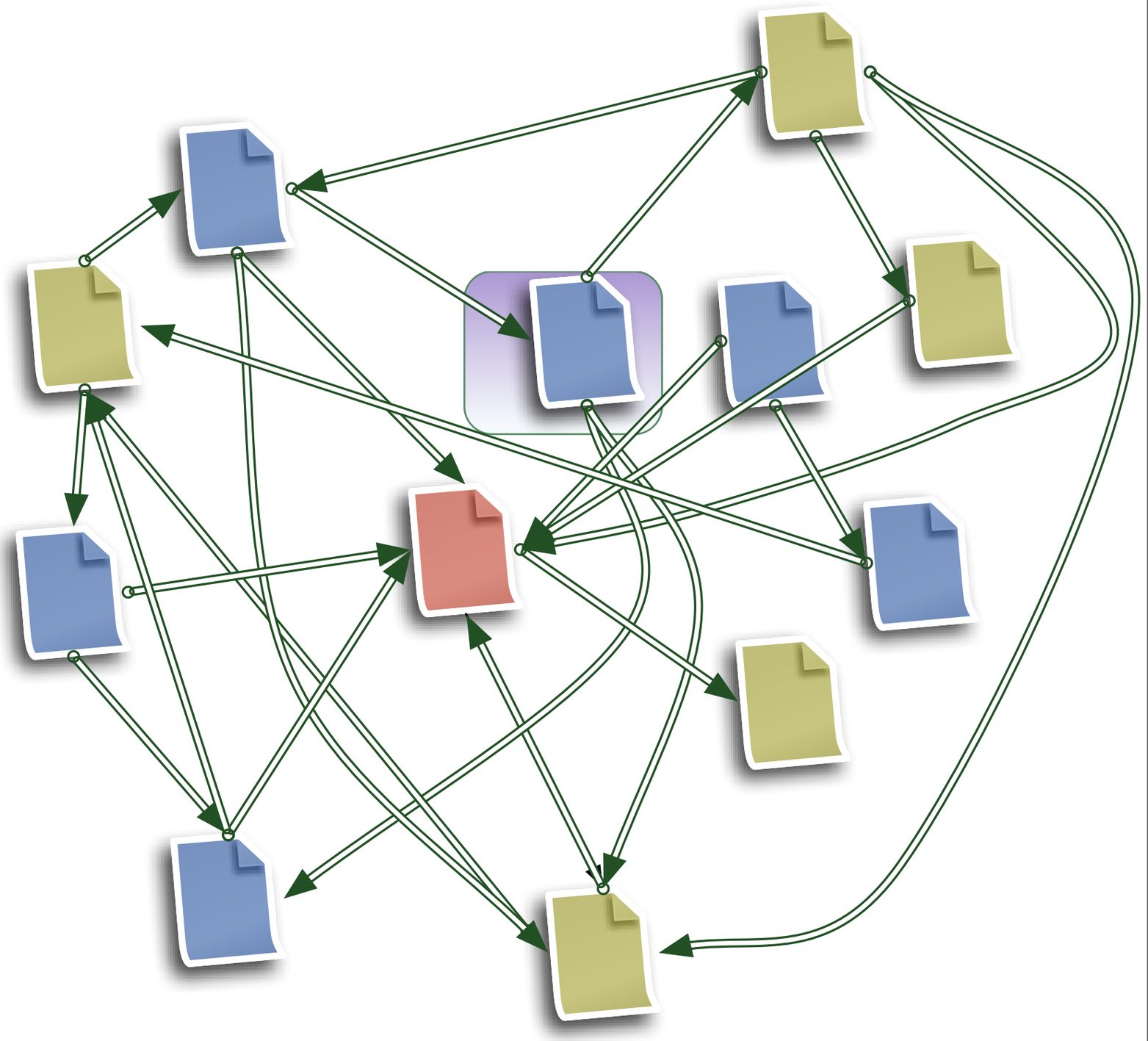


Confiar

PageRank

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1$$

- Análise estrutural da Web.
- Autoridade nem sempre qualidade
- Comportamento dos utilizadores ajuda no ranking.
- Problemas com Spam.

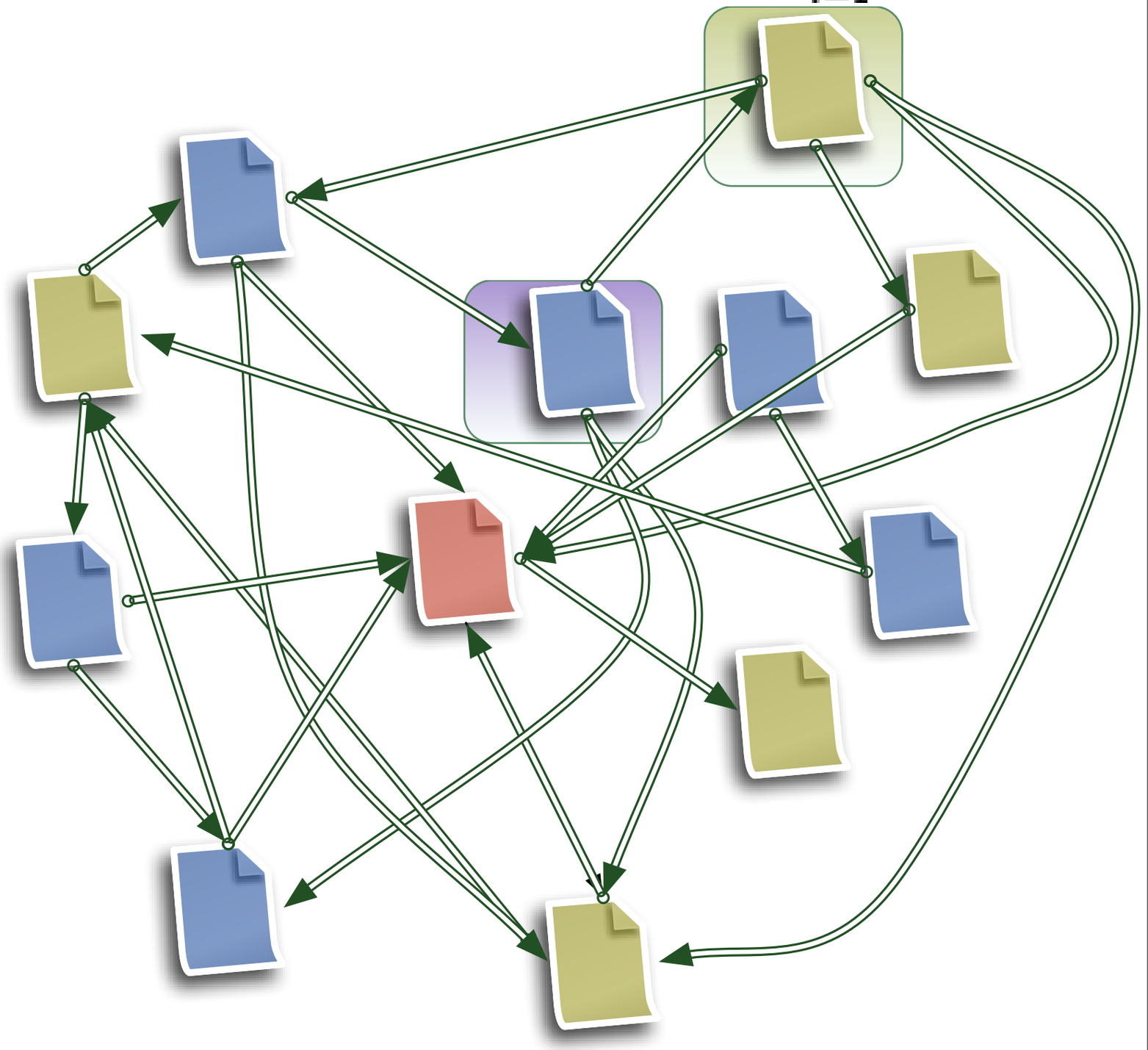


Confiar

PageRank

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1$$

- Análise estrutural da Web.
- Autoridade nem sempre qualidade
- Comportamento dos utilizadores ajuda no ranking.
- Problemas com Spam.

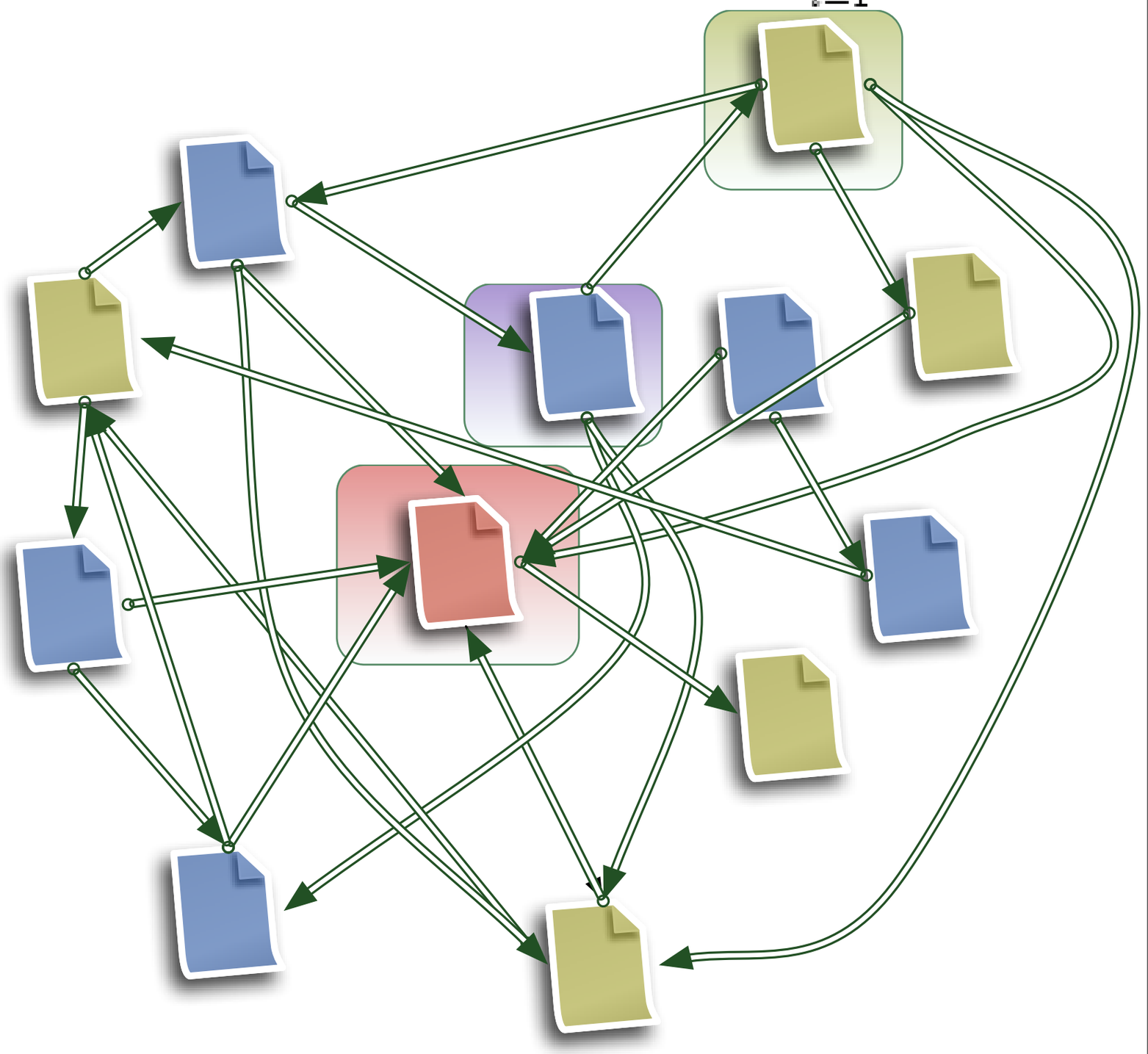


Confiar

PageRank

$$\sum_{i=1}^{N} \ell(p_i, p_j) = 1$$

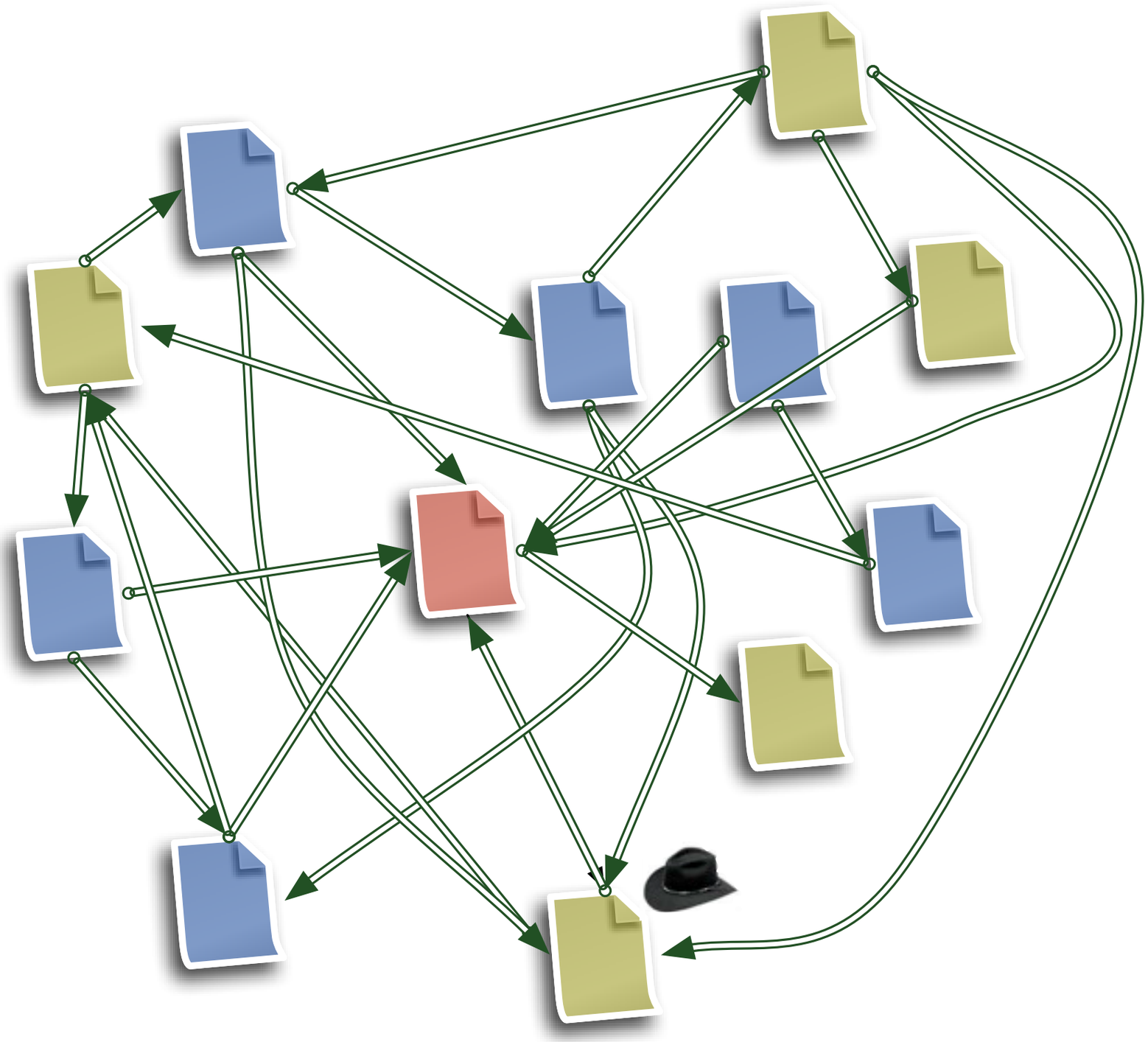
- Análise estrutural da Web.
- Autoridade nem sempre qualidade
- Comportamento dos utilizadores ajuda no ranking.
- Problemas com Spam.



Confiar

SEO

- Como manipular a confiança a favor do meu site?
- White hat
  - Trazer resultados mais relevantes, seguindo as regras.
- Black hat
  - Trazer mais tráfego, com recursos a meios menos legítimos.



sapo.pt

Obrigado!

João Pedro Gonçalves
jp@sapo.pt
Abril 2010