Université de Montréal

Processus de recherche de documents sur le web

Guy Lapalme
IFT3225

Exemple tirée de la section 12.1.2 de *Web Application Architecture*, L. Shklar & R. Rosen, Wiley, p 349-357

Programme python illustrant ces calculs

rali

Documents

Doc 1	Search technologies have been around for over forty years. Over this time, their user base expanded first from scientists and technologists to information professionals, and finally from information professionals to pretty much everyone.
Doc 2	Math and physics students are familiar with the challenge of finding the unambiguous "right answer". The same is not true for information retrieval. Finding the "right document" may be as much art a science.
Doc 3	Many serial killers do not suffer from psychosis and appear to be quite normal. Search for such serial killers can take years, even with the latest police technologies, and the results are often shocking..
StopWords	have been for over this their from and for the with same is not quite true may be as do and such can even often to are a of of
Requête	**the promise of search technologies**

Documents

Doc 1 — Search technologies ~~have been~~ around ~~for over~~ forty years. ~~Over~~ this time, ~~their~~ user base expanded first ~~from~~ scientists ~~and~~ technologists ~~to~~ information professionals, ~~and~~ finally from information professionals ~~to~~ pretty much everyone.

Doc 1 - sw — search technologies around forty years time user base expanded first scientists technologists information professionals finally information professionals pretty much everyone

Doc 1 -sw+lem — search **technology** around forty year time user base expand first **science technology** information **professional final** information **professional** pretty much everyone

Requête — ~~the~~ promise ~~of~~ search technologies

Requête -sw+lem — promise search **technology**

TERM	D1	D2	D3	Q
everyone	1	-	-	-
art	-	1	-	-
right	-	2	-	-
result	-	-	1	-
shock	-	-	1	-
year	1	-	1	-
serial	-	-	2	-
suffer	-	-	1	-
technology	2	-	1	1
find	-	2	-	-
unambiguous	-	1	-	-
information	2	1	-	-
police	-	-	1	-
appear	-	-	1	-
forty	1	-	-	-
much	1	1	-	-
take	-	-	1	-
pretty	1	-	-	-
answer	-	1	-	-
document	-	1	-	-
final	1	-	-	-
math	-	1	-	-
around	1	-	-	-
normal	-	-	1	-
familiar	-	1	-	-
killer	-	-	2	-
base	1	-	-	-
user	1	-	-	-
student	-	1	-	-
expand	1	-	-	-
search	1	-	1	1
science	1	1	-	-
challenge	-	1	-	-
many	-	-	1	-
psychosis	-	-	1	-
time	1	-	-	-
professional	2	-	-	-
latest	-	-	1	-
physics	-	1	-	-
retrieval	-	1	-	-
first	1	-	-	-

Vecteurs des termes

différences avec
requête
D1=1 [2]
D2=2 [3]
D3=0 [1]

différences avec
requête pondérée (requête *1000)
D1=1997 [1]
D2=2000 [3]
D3=1998 [2]

rali

TERM	D1	D2	D3	Q	
everyone	1	-	-	-	
art	-	1	-	-	
right	-	2	-	-	
result	-	-	1	-	
shock	-	-	1	-	
year	1	-	1	-	
serial	-	-	2	-	
suffer	-	-	1	-	
technology	2	-	1	1	
find	-	2	-	-	
unambiguous	-	1	-	-	
information	2	1	-	-	
police	-	-	1	-	
appear	-	-	1	-	
forty	1	-	-	-	
much	1	1	-	-	
take	-	-	1	-	
pretty	1	-	-	-	
answer	-	1	-	-	
document	-	1	-	-	
final	1	-	-	-	
math	-	1	-	-	
around	1	-	-	-	
normal	-	-	1	-	
familiar	-	1	-	-	
killer	-	-	2	-	
base	1	-	-	-	
user	1	-	-	-	
student	-	1	-	-	
expand	1	-	-	-	
search	1	-	1	1	
science	1	1	-	-	
challenge	-	1	-	-	
many	-	-	1	-	
psychosis	-	-	1	-	
time	1	-	-	-	
professional	2	-	-	-	
latest	-	-	1	-	
physics	-	1	-	-	
retrieval	-	1	-	-	
first	1	-	-	-	
Synt0	1	1	-	1	search technology / information retrieval
Synt1	-	-	1	-	police technology
Synt2	2	-	-	-	information professional

Vecteurs avec syntagmes

requête initiale
D1=1 [1]
D2=2 [3]
D3=1 [1]

rali

TERM	D1	D2	D3	Q	
everyone	1	-	-	-	
art	-	1	-	-	
right	-	2	-	-	
result	-	-	1	-	
shock	-	-	1	-	
year	1	-	1	-	
serial	-	-	2	-	
suffer	-	-	1	-	
technology	2	-	1	1	
find	-	2	-	-	
unambiguous	-	1	-	-	
information	2	1	-	-	
police	-	-	1	-	
appear	-	-	1	-	
forty	1	-	-	-	
much	1	1	-	-	
take	-	-	1	-	
pretty	1	-	-	-	
answer	-	1	-	-	
document	-	1	-	-	
final	1	-	-	-	
math	-	1	-	-	
around	1	-	-	-	
normal	-	-	1	-	
familiar	-	1	-	-	
killer	-	-	2	-	
base	1	-	-	-	
user	1	-	-	-	
student	-	1	-	-	
expand	1	-	-	-	
search	1	-	1	1	
science	1	1	-	-	
challenge	-	1	-	-	
many	-	-	1	-	
psychosis	-	-	1	-	
time	1	-	-	-	
professional	2	-	-	-	
latest	-	-	1	-	
physics	-	1	-	-	
retrieval	-	1	-	-	
first	1	-	-	-	
Synt0	10	10	-	1	search technology / information retrieval
Synt1	-	-	10	-	police technology
Synt2	10	-	-	-	information professional

Vecteurs avec syntagmes

requête initiale
D1=1 [1]
D2=2 [3]
D3=1 [1]

requête pondérée (requête *1000)
D1=2987 [1] (syntagme *10)
D2=2990 [2]
D3=2998 [3]

rali

Université de Montréal

SEO
Search Engine Optimisation

Tenter d'améliorer le référencement de son site

rali

Trucs *légaux*
White hat SEO

- mettre du contenu intéressant pour que les gens lui fassent des liens
- simplification de la structure des URL
- site maps
- enlever les références circulaires
- éviter les multiples URL qui référencent la même ressource

rali

Trucs *moins légaux*
«Black hat SEO»

- insérer des mots-clés invisibles
- mettre des faux mots dans le meta-tag
- créer des pages qui ne sont visibles que des *web crawlers*
- créer des *link farms*

rali

Trucs *moins légaux*
«Black hat SEO»

- insérer des mots-clés invisibles
- mettre des faux mots dans le meta-tag
- créer des pages qui ne sont visibles que des *web crawlers*
- créer des *link farms*

Guerre entre les moteurs de recherche
et les gens qui veulent s'afficher

rali

Orientation des crawlers dans son site

- robot.txt : permissions pour des parties du site

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tmp/
Disallow: /~joe/
```

- sitemap
  - propriétés d'une partie du site
  - fréquences des changements
  - dernières modifications

rali

Exemple de sitemap

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
   <url>
      <loc>http://www.example.com/catalog?item=12&amp;desc=vacation_hawaii</loc>
      <changefreq>weekly</changefreq>
   </url>
   <url>
      <loc>http://www.example.com/catalog?item=73&amp;desc=vacation_new_zealand</loc>
      <lastmod>2004-12-23</lastmod>
      <changefreq>weekly</changefreq>
   </url>
   <url>
      <loc>http://www.example.com/catalog?item=74&amp;desc=vacation_newfoundland</loc>
      <lastmod>2004-12-23T18:00:15+00:00</lastmod>
      <priority>0.3</priority>
   </url>
   <url>
      <loc>http://www.example.com/catalog?item=83&amp;desc=vacation_usa</loc>
      <lastmod>2004-11-23</lastmod>
   </url>
</urlset>
```

rali