BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

02-03-2010

1 SpringerLink

Google facilita l'enllaç al PDF d'SpringerLink, no a la pàgina amb la referència. Si s'intenta accedir-hi des del navegador, se'ns redirecciona directament a la pàgina correcta, però si volem obtenir el contingut amb el nostre `Browser` de Python, hi ha dues redireccions que ens acaben portant a la pàgina principal. Per exemple:

- Amb l'enllaç: `http://www.springerlink.com/index/CGBHLDY69PVGWL6D.pdf`, obtenim un `302 Found` que ens porta a `http://www.springerlink.com/link.asp?id=cgbhldy69pvgwl6d`.
- Seguint aquesta adreça, obtenim un altre `302 Found` que ens porta a `http://www.springerlink.com/default.mpx`.

Això és degut a que les cookies no estan habilitadies i no es pot emmagatzemar l'identificador de la sessió ASP d'SpringerLink.

La solució ha estat afegir un *handler* per a poder utilitzar cookies: `handlers = [PoolHTTPHandler, urllib2.HTTPCookieProcessor]`

2 Bloqueig de Google

Si s'executen moltes (en realitat, no moltes) consultes a la vegada Google passa a bloquejar el servei i tarda unes hores a desbloquejar-se.
Possibles solucions:

- Utilitzar Tor o similar per anonimitzar el tràfic. La velocitat de transferència és molt menor a l'habitual. Hi ha timeouts.
- Fer que l'usuari vagi introduint *captchas* de tant en tant. He estat treballant en un script per poder introduir a mà el captcha i desbloquejar-ho, però encara no funciona del tot.

3 Versions del mateix article

M'he trobat amb articles com ara el que porta per títol *Logical Implication and Causal Dependency* que al buscar-lo per Internet només hi ha una versió (se suposa que més nova) i que té un títol diferent. Es pot trobar a `http://www.springerlink.com/content/p0086762337rq30t/`.

Un altre exemple: l'article amb nom *Prince: An Algorith for...* es pot trobar a `http://www.springerlink.com/content/h8mja4yu8279mjq1/` amb el nom *TEX: A New Informative Generic Base of Association Rules*. Què cal fer en aquests casos? Com ho podem detectar?