BibTeX Bibliography Index Maker: Notes

Ramon Xuriguera

1 BibTeX

Aspectes del format BibTeX a tenir en compte:

- Com podem distingir entre diferents tipus d'entrada (article, book, inproceedings, etc.) a partir del fitxer?
- Format dels noms. Un nom consisteix de diferents parts: First, von, Last, Jr. El token *von* o *de la* cal posar-los en minúscules. Per tal que el nom es reconegui, cal que tingui el format: von Last, Jr, First. D'aquesta manera, si hi ha més d'un cognom no passa res.
- Caràcters Unicode entre claus per poder ser utilitzats correctament amb l'estil *alpha*. Per exemple: `Jos{\'{e}}`
- Per prevenir que BibTeX canviï un text a minúscules, cal posar el text entre claus.
- Si hi ha massa autors, truncar la llista amb *et al.*
- Utilitzar abreviatures de tres lletres per als mesos
- Utilitzar el camp `key` per a organitzacions amb un nom llarg, de manera que s'utilitzin les inicials de l'organització al fer una cita.

2 Extracció del contingut dels fitxers PDF

Aspectes a considerar:

- Caràcters Unicode
- Glyphs com ara *fi* corresponen a més d'una lletra
- Les llibreries extreuen el text per línies
- Podem obtenir les següents dades del fitxer sense haver-ne d'extreure el text: número de pàgines, títol, autor, assumpte i paraules clau. Si aquests camps no s'han omplert al generar el fitxer, estaran en blanc.

2.1 Software

- xPDF Proporciona eines executables des de la línia de comandes per extreure el text. Només converteix a text pla, però no separa els diferents paràgrafs en línies diferents. Permet obtenir el resultat en diferents codificacions de caràcters, però no conserva els text correcte. És a dir, continua tenint problemes per extreure accents, etc. Amb l'eina `pdftohtml` podem obtenir informació d'algunes de les paraules en negreta, però separa cada línia amb una etiqueta br.

- PDFBox Llibreria escrita en Java i publicada sota la llicència *Apache License v2.0*. Actualment es troba a la incubadora d'Apache. Permet obtenir el text i les metadades d'un fitxer. Separa els resultats per línia segons es troben en el document, no per paràgrafs.

2.2 Procediment

3 Articles Db

Google Scholar/Google Search API
Enllaç directe a la cita en format BibTeX. Actualment no proporcionen cap API per obtenir-ho. Caldria fer-ho amb `wget`.

DBLP
DBLP proporciona una API molt simple per cercar autors autors i obtenir les seves publicacions, però no al revés. Al cercar per autor, es mostren les claus de cada publicació, però no el títol.

Portal ACM

CiteSeerX

Arxiv.org
`http://arxiv.org/help/api/index`

4 JabRef

Java, llicència: LGPL.
És possible crear plug-ins amb *Java Plug-in Framework*. L'última versió d'aquest framework és de fa més de dos anys, actualment es troba en estat *frozen* i és pre-OSGi. En el fòrum i el tracker de *JabRef* no hi ha cap missatge en el que es discuteixin canvis sobre aquest tema.
Extension-points permesos:

- **ImportFormat** Add importers to JabRef accessible from the 'Import into ... database'.
- EntryFetcher Add access to databases like Citeseer or Medline to the Web Search menu.
- ExportFormatTemplate Add a template based export like the ones accessible using the Manage Custom Exports.
- ExportFormat Add an export filter to JabRef's export dialog, that is more complicated than the simple template based one.
- ExportFormatProvider A more powerful way to add export formats to JabRef.
- LayoutFormatter Add formatters that can be used in the layout based exporters.
- **SidePanePlugin** Add a side pane component that can do any kinds of operations. The panel is accessed from a Plugins menu in JabRef's main window.

5 Llenguatge

La idea seria desenvolupar una aplicació de línia de comandes amb *Jython*, Python sobre la JVM i utilitzar

A Exemples de text extret

A continuació es mostren algunes capçaleres d'articles i el resultat obtingut a l'extreure'n el text:

- Text 1 PDF:

Lorentz symmetry violation, dark matter and dark energy

Luis Gonzalez-Mestres[a]

[a]LAPP, Université de Savoie, CNRS/IN2P3, B.P. 110, 74941 Annecy-le-Vieux Cedex, France

Taking into account the experimental results of the HiRes and AUGER collaborations, the present status of bounds on Lorentz symmetry violation (LSV) patterns is discussed. Although significant constraints will emerge, a wide range of models and values of parameters will still be left open. Cosmological implications of allowed LSV patterns are discussed focusing on the origin of our Universe, the cosmological constant, dark matter and dark energy. Superbradyons (superluminal preons) may be the actual constituents of vacuum and of standard particles, and form equally a cosmological sea leading to new forms of dark matter and dark energy.

1. Patterns of Lorentz symmetry violation particle. For $p \gg mc$, one has:

Text pla:

```
arXiv:0912.0725v1 [astro-ph.CO] 3 Dec 2009

Lorentz symmetry violation, dark matter and dark energy
Luis Gonzalez-Mestresa
a
LAPP, Universit de Savoie, CNRS/IN2P3, B.P. 110, 74941 Annecy-le-Vieux Cedex, Fr

Taking into account the experimental results of the HiRes and AUGER collaboratio

1. Patterns of Lorentz symmetry violation A formulation of Planck-scale Lorentz
```

HTML:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"><HTML>
<HEAD>
<TITLE></TITLE>
</HEAD>
<BODY>
<A name=1></a>Lorentz symmetry violation, dark matter and dark energy<br>
Luis Gonzalez-Mestresa<br>
aLAPP, Universite de Savoie, CNRS/IN2P3, B.P. 110, 74941 Annecy-le-Vieux Cedex, F
Taking into account the experimental results of the HiRes and AUGER collaboration
bounds on Lorentz symmetry violation (LSV) patterns is discussed. Although signif
1. Patterns of Lorentz symmetry violation<br>
```

- Text 2