BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

09-02-2010

1 Lligadures

En alguns casos, l'eina `pdftotext` té problemes per reconéixer alguns caràcters com ara les lligadures: *fi*, *ffi*, etc. Pel que he vist, aquests caràcters es substitueixen per dues cometes ". Solució: podem fer la cerca a l'Scholar amb una expressió regular del tipus: `r = re.search('([\w]+ ){6,8}',str)`. On \w equival a `[a-zA-Z0-9_]`

2 Com trobar l'*abstract*?

Si el paràgraf comença amb les paraules *Abstract*, *Summary*, *In this paper*, etc. no hi ha cap problema. Però si no ho fa, és més complicat.

Podem dividir el text en paràgrafs, però reconéixer quin és el primer caràcter de text és més complicat. Per exemple, donada la primera pàgina d'un article, obtenim els quatre paràgrafs següents:

- Neurocomputing 35 2000 3}26
- Class separability estimation and incremental learning using boundary methods Jose-Luis Sancho *, William E. Pierson , Batu Ulug , H AnmH bal R. Figueiras-Vidal , Stanley C. Ahalt ATSC-DI, Escuela Politecnica Superior. Universidad Carlos III Leganes-Madrid, Spain & & Department of Electrical Engineering, The Ohio State University Columbus, OH 43210, USA Received 7 January 1999; revised 5 April 1999; accepted 10 April 2000
- Abstract In this paper we discuss the use of boundary methods (BMs) for distribution analysis. We view these methods as tools which can be used to [...] estimation; Gradient algorithms; Sample selection strategies
- * Corresponding author. Tel.: #34 [...] 2 9 3 - 9

Provar utilitzant una expressió regular, si no s'obtenen bons resultats, fer PoS tagging de tots els paràgrafs del text i mirar la relacio $\frac{noms}{verbs}$ per saber quin és el primer paràgraf de text.

3 Extracció d'informació

Si podem extreure la referència bibtex sencera: Wrappers especials.
Sinó: Wrappers per camp per tal de reaprofitar al màxim.

Si coneixem la pàgina, utilitzem els wrappers corresponents.
Sinó, els provem tots per mirar d'obtenir tants camps com puguem.