UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007-2013

Platformă de e-learning și curriculă e-content pentru învățământul superior tehnic

AEACD

21. Analiza de Ierarhie

Elemente principale

- In momentul in care vorbim despre ierarhia unui document ne referim la ierarhia paginiilor si nu la ierarhia continutului paginiilor, desi acest al doilea tip de ierarhie ce a fost discutat in cursurile anterioare ajuta la obtinerea analizei dorite in acest capitol
- Astfel:
  - Coperta (pot fi continute mai multe pagini)
  - Cuprins
  - Parte introductiva
  - Continut

Elemente principale

- Pentru detectia tuturor acestor elemente este necesara analiza continutului paginiilor individuale, dupa cum s-a mentionat
- Cele mai importante surse de informatie sunt date de:
  - Numarul paginii;
  - Tipul textului continut si cantitatea (titluri, paragrafe de text, imagini)

Coperta

- De cele mai multe ori copertiile pot fi gasite deoarece contin putin text, majoritatea fiind titluri, cu font mult mai mare decat restul paginiilor
- De asemenea daca limbajul in care este redactat documentul este cunoscut se pot cauta cuvinte cheie precum: “Editura”, “Vol.” etc

Cuprinsul

- Cuprinsul poate fi detectat foarte usor deoarece majoritatea liniilor de text, daca au fost detectate corect vor contine atat litere cat si cifre, daca OCRul a functionat relativ corect
- De asemenea cuvinte cheie pot fi cautate pentru a detecta daca este sau nu curpinsul documentului

Analiza de continut

- Continutul documentului poate contine deseori o parte introductiva, mai ales in cazul cartiilor
- Folosirea unui alt tip de numerotare al paginiilor (de exemplu caractere romane in loc de caractere arabe) pot conduce la detectia acestei parti
- De asemenea cautarea in titlu al diferitelor cuvinte cheie sau in anumite cazuri pagina va contine in partea superioara un antet, deseori delimitat printr-un separator ce va contine numele capitolului

Analiza de continut

- Un alt factor important in analiza de continut o reprezinta paginiile de legatura
- Exista anumite pagini ce contin fie doar numele capitolelor sau numele capitolelor plus alte elemente
- Aceste nume se gasesc usor deoarece sunt detectate ca titluri si contin de cele mai multe ori cuvinte cheie precum “Capitol” in cazul in care se cunoaste limba si detectia OCR a fost facuta corect

Utilizarea tiparelor de documente

- Odata descoperita structura unui document, aceasta poate fi de cele mai multe ori folosita ca tipar pentru alte documente de acelasi tip ce au caracteristici comune, de exemplu editura
- Astfel, se poate castiga foarte mult timp si de asemenea se poate imbunatati calitatea anumitor documente cunoscand caracteristicile lui pe baza documentelor tipar