IA 2017-2018, B5

Participanți:

Bivol Daniel

Ghergu Andreea

Lungeanu Ionut

Matrana Karina

Palihovici Andrei

Prodan Marius

Tanasa Petru

Trifan Tamana

Turcu Nicusor

Ulinici Cristina

Vasilache Diana

Tutore inteligent

Grupa 1, submodul 1 – Pregătirea textului

Descriere:

Scopul acestui submodul este prelucrarea unui text extras din fișiere de diferite tipuri (pdf, docx, etc). Prelucrarea constă în parsarea textului, extragerea cuvintelor și curățarea de imagini și orice alte elemente irelevante contextului, cum ar fi numerotarea paginilor, headers, footers sau alte elemente de acest tip. Informațiile vor fi trimise către submodulul 2, printr-un fișier stabilit în prealabil, care va conține cuvintele extrase, semnele de punctuație corespunzătoare, precum și alte informații necesare submodulului 2.

# Extensie:

Presupunem că avem o carte scanată (pot fi și o serie de imagini) în format pdf. Deci textul este efectiv încadrat într-o imagine, nu poate fi selectat sau prelucrat direct de către alte programe ca entitate text. În acest caz aplicația noastră permite extragerea textului din imagine printr-un parser OCR (optical character recognition).

Scenarii de utilizare:

# (Scenarii care pot apărea la parsarea textului)

* Cuvintele separate prin cratimă care apar la sfârșit și început de linie sunt selectate fără cratimă.
* Imaginile sunt ignorate. Descrierea unei imagini trebuie prelucrată și ea (?)
* Diagramele, schemele și tabelele sunt trasnformate în format text(cât de mult posibil) pentru a putea extrage conținut relevant (?)
* Textul care apare sub forma unei enumerări este încadrat în propoziția anterioară, grupurile de cuvinte de la fiecare enumerare fiind separate printr un semn de punctuație ( , sau ; )
* Titlurile și denumirile de sub-capitole sunt transmise drept propoziții separate.
* Numerotarea paginilor este ignorată.
* Footer-ele și header-ele sunt ignorate.
* Watermark –urile sunt ignorate.
* Notele la subsol sunt parsate și încadrate în propoziția potrivită contextual.
* Cuprinsul și referințele sunt ignorate.