Praktični zadatak – Ekstrakcija podataka iz PDF dokumenata

Kolegij: Uvod u teorijsko računarstvo
Nositelj predmeta: Prof. dr. sc. Sanda Martinčić-Ipšić
Asistent: Andrija Poleksić, mag. inf.

Autor: Ivan Pribanić
Godina: 2026

---

1. Opis zadatka

Cilj zadatka je automatski izlučiti strukturirane podatke iz nepristupačnih PDF dokumenata znanstvenih radova iz domene klimatskih promjena te ih spremiti u više čitljivih formata.

PDF dokumenti nisu prilagođeni automatskoj obradi jer ne sadrže jasnu semantičku strukturu (naslovi, autori, sažetak, sadržaj). Zbog toga je potrebno koristiti heuristike i alate za ekstrakciju teksta kako bi se dobili korisni podaci.

---

2. Izvor podataka

Korišten je skup PDF dokumenata znanstvenih radova iz časopisa iz područja klimatskih promjena i atmosferske znanosti.

Izvor: - PDF datoteke preuzete sa sustava Merlin (Nature Climate and Atmospheric Science) - Datoteke su pohranjene lokalno u direktoriju: /pdfs - Svaka PDF datoteka predstavlja jedan znanstveni rad.

---

3. Korišteni alati i biblioteke

Projekt je implementiran u Python Jupyter Notebook okruženju.

Korištene biblioteke:

PyMuPDF (fitz)

- Ekstrakcija teksta iz PDF dokumenata
- Brzo čitanje pojedinih stranica bez učitavanja cijelog dokumenta

Pandas

- Obrada tabličnih podataka
- Spremanje podataka u CSV i Parquet formate

- Transformacija strukture podataka

JSON

- Spremanje rezultata u strukturirani JSON format

Pathlib

- Upravljanje datotekama i direktorijima

Regular Expressions (re)

- Čišćenje teksta
- Prepoznavanje godina
- Heuristička ekstrakcija autora i naslova

---

4. Primjena velikih jezičnih modela (LLM)

Veliki jezični model (ChatGPT) korišten je za:

- Dizajn heuristika za ekstrakciju autora i naslova
- Optimizaciju obrade PDF tekstualnog sadržaja
- Pomoć pri rješavanju problema s encoding greškama
- Popravljanje heuristike prepoznavanja autora
- Oblikovanje i formatiranje dokumentacije
- Analizu strukture zadatka i zahtjeva

---

5. Struktura spremljenih podataka (Shema)

Implementirana je proširena shema podataka:

Polje	Tip	Opis
file_name	STRING	Naziv PDF datoteke
title	STRING	Naslov znanstvenog rada
authors_raw	STRING	Autori u tekstualnom obliku (razdvojeni sa ;)
authors	LIST[STRING]	Lista autora (samo u JSON export)
year	STRING	Godina objave rada
abstract	STRING	Sažetak rada (ako je pronađen)
content	STRING	Dio glavnog sadržaja rada
num_pages	INTEGER	Broj stranica PDF dokumenta
has_tables	BOOLEAN	Detekcija pojave tablica u tekstu
extraction_issues	STRING	Bilješke o problemima ekstrakcije

---

6. Format izlaznih podataka

Podaci su spremljeni u četiri različita formata:

CSV

- Pogodan za tabličnu analizu
- Autori spremljeni kao jedan string(razdvojena sa ';')
- Datoteka: `results.csv`

JSON

- Strukturni format pogodan za daljnju automatsku tekstualnu obradu
- Autori spremljeni kao lista stringova
- Datoteka: `results.json`

Parquet

- Kolumno orijentirani format sa dobrim performansama
- Optimiziran za brzu obradu velikih skupova podataka
- Datoteka: `results.parquet`

XML

- Hijerarhijski tekstualni format pogodan za razmjenu podataka između različitih sustava i aplikacija.
- Posebna obrada je primijenjena kako bi se uklonili nevažeći kontrolni znakovi koji mogu uzrokovati probleme pri parsiranju XML datoteke.
- Datoteka: `results.xml`

---

7. Način ekstrakcije podataka

Naslov rada

- Ekstrakcija s prve stranice PDF-a
- Ignoriraju se meta oznake (ARTICLE, OPEN, DOI)
- Naslov se sastavlja iz prvih nekoliko linija iznad autora

Autori

Autori se izvlače pomoću heuristika:

- Traže se linije odmah ispod naslova
- Filtriraju se footnote oznake, brojevi i simboli
- Imena se razdvajaju pomoću `;`, `,` i `and`
- Duplicirani autori se uklanjaju

Rezultat: - `authors_raw` → tekstualni zapis - `authors` → lista (koristi se samo za JSON)

Godina objave

Godina se pokušava pronaći pomoću:

- Uzorka `(YYYY)`
- Linija s oznakama `Published`, `Received`, `Accepted`
- Copyright oznaka

Ako godina nije pronađena, polje ostaje prazno.

Sažetak (Abstract)

Sažetak se izlučuje:

- Iz dijela teksta prije pojave riječi `INTRODUCTION`
- Čiste se meta linije (npj, Published, ARTICLE)
- Ograničava se maksimalna duljina

Glavni sadržaj

Zbog performansi:

- Ne učitava se cijeli PDF
- Čita se ograničen broj početnih stranica
- Tekst se normalizira i čisti

---

8. Bilješke o ekstrakciji podataka

Tijekom obrade uočeni su sljedeći problemi:

1. Nepotpuni autori

- Kod nekih radova nisu svi autori ispravno izvučeni
- Razlog: složeni layout, stupci i footnote oznake

Rješenje: - Korišten heuristički pristup - Prednost dana točnosti imena iznad potpunosti

---

2. Gubitak dijela teksta

- Matematičke jednadžbe i specijalni simboli često nisu pravilno izvučeni
- Neki PDF-ovi sadrže slike umjesto pravog teksta (skeniranje)

---

3. Ligature i encoding problemi

Primjeri: - effects -> effects - first -> first

Rješenje: - Normalizacija ligatura prije obrade podataka

---

4. Različita struktura PDF dokumenata

- Nisu svi radovi jednako formatirani
- Neki nemaju jasno označen abstract
- Neki sadrže zaglavlja umetnuta u tijelo teksta

---

9. Zaključak

Projekt uspješno demonstrira:

- Automatsku ekstrakciju podataka iz PDF dokumenata
- Pretvaranje nestrukturiranog teksta u strukturirane zapise
- Spremanje podataka u više standardnih formata

Najveći izazovi ostaju:

- Normaliziranje i obrada autora
- Kompleksni layout PDF dokumenata
- Matematičke jednadžbe
- Varijacije u strukturi radova