Academic Data Extraction FROM Gradus

Project done by:

Jumaniyazov Suleyman - [EG6X6K]
Dijanira Muachifi - [U2FELP]
Bermejo Magarín Alberto - [H64GOG]

Project Description

This project **automates** the **collection** and **analysis** of article data from the Gradus website. It utilizes web scraping, PDF processing, and data organization to extract relevant information and consolidate it into a **structured Excel file**.

Required data

Year

Publication year of the article.

Volume

Volume number of the publication.

Number

Issue number within the volume.

File name

PDF file name of the article.

First page

Starting page number of the article.

Last page

Ending page number of the article.

Authors

Names of the article's authors.

Original title

Article's title in the original language.

Title in English

Translated title of the article.

Section code

Article's section classification (e.g., ART).

Abstract original

Abstract in the original language.

Abstract english

Abstract translated into English.

Email address

Corresponding author's email address.

MTA REPO URL

Article's link on the MTA Repository.

DOI

Digital Object Identifier for the article (from 2020).

MTMT

Link to the article in MTMT database.

Objectives

Data Collection

Scrape and download PDF files of articles available on the Gradus website

Information Extraction

Process the PDFs to extract the aforementioned data points

Result Storage

Export the data into an organized Excel file for easy access and analysis

Libraries

We have been working with Python 3.x, and we used the following libraries:

Os	For file management on the operating system.
Requests	To make HTTP requests and download files.
BeautifulSoup	For web scraping and HTML element extraction.
PyPDF2	To read and extract text from PDF files.
Pandas	For organizing and managing tabular data.
Openpyxl	To create and manipulate Excel spreadsheets.
Re	For pattern matching using regular expressions.

Code Architecture

01

Constants

Base URL of the Gradus site to access the 2020 articles.

02

process_pdf(pdf_url)

This function downloads a PDF file, processes its content, and returns the extracted text and the first and last page numbers of the article.

03

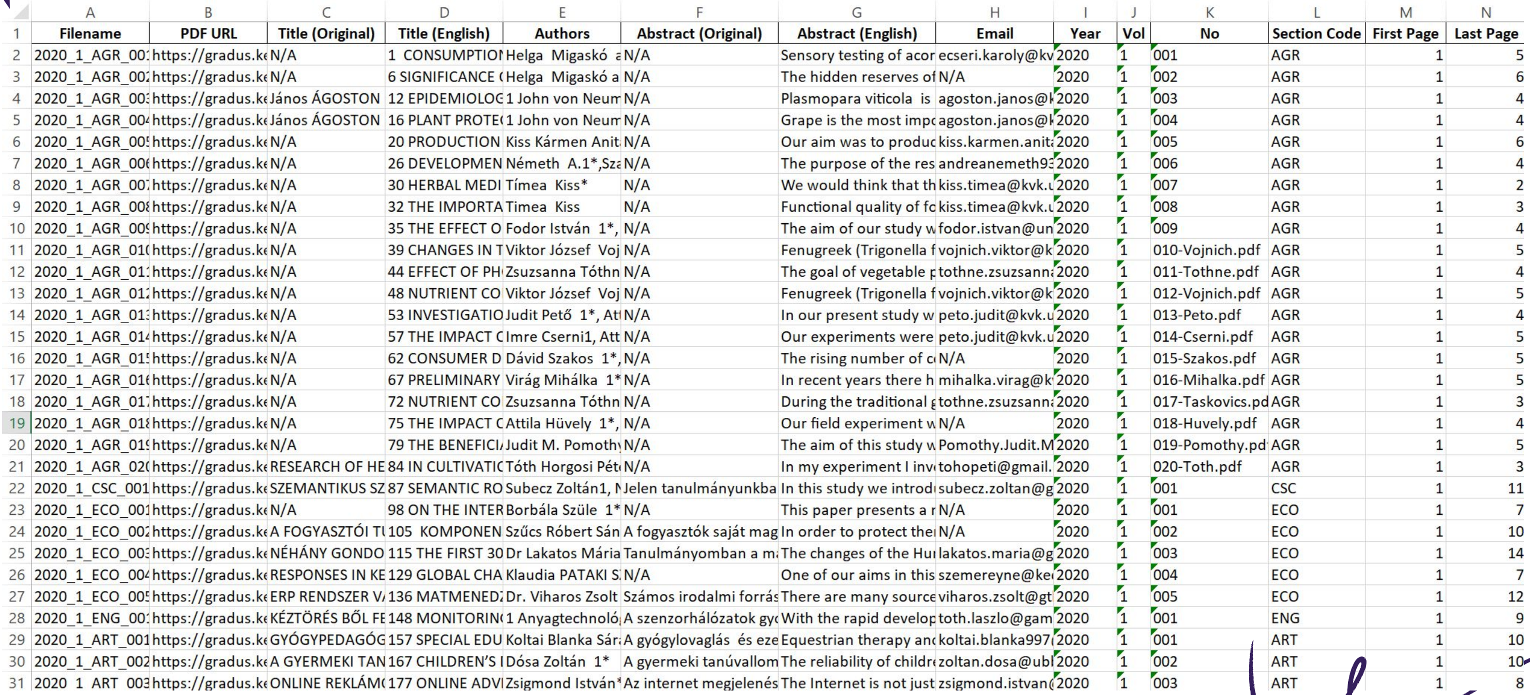
scrape_gradus()

Scrapes Gradus for PDFs, extracts article details (titles, abstracts, authors, emails), and organizes them into dictionaries.

04

main()

Coordinates the program, collects data with scrape_gradus(), converts it to a DataFrame, and saves it as gradus_articles.xlsx.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Filename	PDF URL	Title (Original)	Title (English)	Authors	Abstract (Original)	Abstract (English)	Email	Year	Vol	No	Section Code	First Page	Last Page
2	2020_1_AGR_00	https://gradus.ke	N/A	1 CONSUMPTIO	Helga Migaskó	N/A	Sensory testing of acor	ecseri.karoly@kv	2020	1	001	AGR	1	5
3	2020_1_AGR_00	https://gradus.ke	N/A	6 SIGNIFICANCE (	Helga Migaskó a	N/A	The hidden reserves of	N/A	2020	1	002	AGR	1	6
4	2020_1_AGR_00	https://gradus.ke	János ÁGOSTON	12 EPIDEMIOLOG	1 John von Neum	N/A	Plasmopara viticola is	agoston.janos@k	2020	1	003	AGR	1	4
5	2020_1_AGR_00	https://gradus.ke	János ÁGOSTON	16 PLANT PROTE	1 John von Neum	N/A	Grape is the most impo	agoston.janos@k	2020	1	004	AGR	1	4
6	2020_1_AGR_00	https://gradus.ke	N/A	20 PRODUCTION	Kiss Kármen Anit	N/A	Our aim was to produc	kiss.karmen.anit	2020	1	005	AGR	1	6
7	2020_1_AGR_00	https://gradus.ke	N/A	26 DEVELOPMEN	Németh A.1*,Sza	N/A	The purpose of the res	andreanemeth9	2020	1	006	AGR	1	4
8	2020_1_AGR_00	https://gradus.ke	N/A	30 HERBAL MEDI	Tímea Kiss*	N/A	We would think that th	kiss.timea@kvk.u	2020	1	007	AGR	1	2
9	2020_1_AGR_00	https://gradus.ke	N/A	32 THE IMPORTA	Timea Kiss	N/A	Functional quality of fo	kiss.timea@kvk.u	2020	1	008	AGR	1	3
10	2020_1_AGR_00	https://gradus.ke	N/A	35 THE EFFECT O	Fodor István 1*,	N/A	The aim of our study w	fodor.istvan@un	2020	1	009	AGR	1	4
11	2020_1_AGR_01	https://gradus.ke	N/A	39 CHANGES IN T	Viktor József Voj	N/A	Fenugreek (Trigonella f	vojnich.viktor@k	2020	1	010-Vojnich.pdf	AGR	1	5
12	2020_1_AGR_01	https://gradus.ke	N/A	44 EFFECT OF PH	Zsuzsanna Tóthn	N/A	The goal of vegetable p	tothne.zsuzsann	2020	1	011-Tothne.pdf	AGR	1	4
13	2020_1_AGR_01	https://gradus.ke	N/A	48 NUTRIENT CO	Viktor József Voj	N/A	Fenugreek (Trigonella f	vojnich.viktor@k	2020	1	012-Vojnich.pdf	AGR	1	5
14	2020_1_AGR_01	https://gradus.ke	N/A	53 INVESTIGATIO	Judit Pető 1*, At	N/A	In our present study w	peto.judit@kvk.u	2020	1	013-Peto.pdf	AGR	1	4
15	2020_1_AGR_01	https://gradus.ke	N/A	57 THE IMPACT C	Imre Cserni1, Att	N/A	Our experiments were	peto.judit@kvk.u	2020	1	014-Cserni.pdf	AGR	1	5
16	2020_1_AGR_01	https://gradus.ke	N/A	62 CONSUMER D	Dávid Szakos 1*,	N/A	The rising number of c	N/A	2020	1	015-Szakos.pdf	AGR	1	5
17	2020_1_AGR_01	https://gradus.ke	N/A	67 PRELIMINARY	Virág Mihálka 1*	N/A	In recent years there h	mihalka.virag@k	2020	1	016-Mihalka.pdf	AGR	1	5
18	2020_1_AGR_01	https://gradus.ke	N/A	72 NUTRIENT CO	Zsuzsanna Tóthn	N/A	During the traditional g	tothne.zsuzsann	2020	1	017-Taskovics.pd	AGR	1	3
19	2020_1_AGR_01	https://gradus.ke	N/A	75 THE IMPACT C	Attila Hüvely 1*,	N/A	Our field experiment w	N/A	2020	1	018-Huvely.pdf	AGR	1	4
20	2020_1_AGR_01	https://gradus.ke	N/A	79 THE BENEFICI	Judit M. Pomoth	N/A	The aim of this study w	Pomothy.Judit.M	2020	1	019-Pomothy.pd	AGR	1	5
21	2020_1_AGR_02	https://gradus.ke	RESEARCH OF HE	84 IN CULTIVATI	Tóth Horgosi Pét	N/A	In my experiment I inv	tohopeti@gmail.	2020	1	020-Toth.pdf	AGR	1	3
22	2020_1_CSC_001	https://gradus.ke	SZEMANTIKUS SZ	87 SEMANTIC RO	Subecz Zoltán1, N	Jelen tanulmányunkba	In this study we introd	subecz.zoltan@g	2020	1	001	CSC	1	11
23	2020_1_ECO_001	https://gradus.ke	N/A	98 ON THE INTER	Borbála Szüle 1*	N/A	This paper presents a r	N/A	2020	1	001	ECO	1	7
24	2020_1_ECO_002	https://gradus.ke	A FOGYASZTÓI T	105 KOMPONEN	Szűcs Róbert Sán	A fogyasztók saját mag	In order to protect the	N/A	2020	1	002	ECO	1	10
25	2020_1_ECO_00	https://gradus.ke	NÉHÁNY GONDO	115 THE FIRST 30	Dr Lakatos Mária	Tanulmányomban a m	The changes of the Hu	lakatos.maria@g	2020	1	003	ECO	1	14
26	2020_1_ECO_00	https://gradus.ke	RESPONSES IN KE	129 GLOBAL CHA	Klaudia PATAKI S	N/A	One of our aims in this	szemereyne@ke	2020	1	004	ECO	1	7
27	2020_1_ECO_00	https://gradus.ke	ERP RENDSZER V	136 MATMENED	Dr. Viharos Zsolt	Számos irodalmi forrás	There are many source	viharos.zsolt@gt	2020	1	005	ECO	1	12
28	2020_1_ENG_00	https://gradus.ke	KÉZTÖRÉS BŐL FE	148 MONITORIN	1 Anyagtechnoló	A szenzorhálózatok gy	With the rapid develop	toth.laszlo@gam	2020	1	001	ENG	1	9
29	2020_1_ART_001	https://gradus.ke	GYÓGYPEDAGÓG	157 SPECIAL EDU	Koltai Blanka Sár	A gyógylovaglás és eze	Equestrian therapy an	koltai.blanka997	2020	1	001	ART	1	10
30	2020_1_ART_002	https://gradus.ke	A GYERMEKI TAN	167 CHILDREN'S	Dósa Zoltán 1*	A gyermeki tanúvallom	The reliability of childr	zoltan.dosa@ub	2020	1	002	ART	1	10
31	2020_1_ART_003	https://gradus.ke	ONLINE REKLÁM	177 ONLINE ADV	Zsigmond István*	Az internet megjelenés	The Internet is not just	zsigmond.istvan	2020	1	003	ART	1	8

CONCLUSION

This project demonstrates the power of automation in handling and processing large amounts of academic data from online sources. By leveraging Python libraries like BeautifulSoup and PyPDF2, we were able to efficiently extract, process, and organize article metadata and content. The final product, an Excel file, provides a structured and user-friendly format for analyzing the gathered information. This project serves as a foundation for further improvements and expansions, showcasing the potential for future applications in academic research and data management.

Contributions

Member	Task	Observations
Jumaniyazov Suleyman	Implementation of the process_pdf function	Focused on extracting simple metadata from PDFs
Muachifi Dijanira	Development of the scrape_gradus function	Web scraping and HTML data collection
Bermejo Magarín Alberto	Creation of Excel export and testing	Final data organization

Thanks for your attention!

Do you have any questions?