Python Data Extraction WorkPlan

What are we doing?

We want to extract data that will be used to train AI chat in the app. A student can chat with this AI and ask it for information about courses, exams, tips and tricks, contact info, university staff name, administrative tasks (e.g. renewing student ID or adding balance to Neptun) and other useful inquiries related to the University of Debrecen.

Data to extract

1. All data must be in English!
2. All data must be related to the Computer Science and Faculty of Informatics and administrative task data.

University Employee Contact Info

Description:
Includes info on professors, practical teachers, administrators, and officers. If a student faces an issue or wants to contact a Uni employee.

- Names
- Role
    - Professor
    - Assistantary
    - Administrator
    - Officer
- Department
- Email Addresses
- Phone Numbers
- Office Address
- Office Days and Hours
- Notice to visit

Example — Dr. Attila Tamás Adamkó, https://inf.unideb.hu/en/dr-attila-tamas-adamko + https://www.ik.unideb.hu/official-documents/office-hours/current-seminar-period-hu.pdf

Name	Role	Department	Email Address	Phone Number	Office Address	Office Days & Hours	Notice to visit
Dr. Adamkó Attila Tamás	Professor	University of Debrecen, Faculty of Informatics, Department of Information Technology	adamko.attila-inf.unideb.hu	+36 52 512 900 / 75226	4028 Debrecen, Kassai út 26. Faculty of Informatics building, floor 2, I226 (Lecturers' room)	Monday, 3:30 p.m; Tuesday, 3:30 p.m;	Yes; Prior application (email) is required, specifying the reason for the visit and the time required. Basically thesis consultations.

Student Administrative Tasks

Description:
Includes tasks that are related to the student's course, exam, personal items, requests and finances. Saves the hassle and stress on the student on trying to figure out how to do the administrative tasks in the deadline.

- subject registration/deregistration
- adding money to Neptun
- Submitting a request on Neptun
- etc

Tips and Tricks

Description:
Includes advice from older students on study techniques, exam strategies, and assignment help.

- Course name
- Exam strategies
- Bonus marks - Yes/No
- Recommended Books & Readings
- Project Guidelines – If the subject has a project component, details on requirements and grading.
- etc

Subject Info

Description:
General information about a subject, including requirements, availability, and assessment structure. Helps students understand what to expect from the course.

- Prerequisites – Required courses before taking this subject.
- Availability – Which semester(s) the subject is offered.
- Exam Procedure
    - Midterm – Yes/No

    - End-term – Yes/No
    - Final Exam – Yes/No
    - Retake Condition
- Exam Type
    - Project-Based
    - Written Exam (Paper-Based)
    - Online Exam
    - Oral Exam
    - Practical Exam (Lab-based, coding, etc.)
- Professor - Individual responsible for the course.
- Credits – Number of credits assigned to the subject.
- Course Type
    - Mandatory
    - Elective
    - Optional

Subject Materials

Description:
Resources and materials needed to study for the subject, including past exam questions and lecture notes.

- Past Papers – Previous exam questions and solutions.
- Course Materials – pdfs, assignments, study guides.
- Recommended Books & Readings
- Project – Yes/No
    - If Yes: Project Guidelines - details on requirements and grading.

Subject Schedule (?)

Description:
Timetable and key dates for the subject, including class times and exam schedules.

- Course Name
- Course Code - Lectures and Seminars have different course codes.
- Lecture Schedules
- Seminar Schedules
- Exam Schedules – Dates for projects, midterms, endterms, finals, and retakes.
- Lecture Instructor - Responsible for lecture and course.
- Seminar Instructor - Responsible for seminars.
- Assignment Deadlines
- Classroom/Location – Room or building where lectures and seminars take place.

Any more categories do you suggest?

Where to extract the data?

Student-based sources:

- Whatsapp groups
- Telegram groups
- Discord groups (?)

University-based sources:

- E-learning
- Neptun
- University of Debrecen Website
- Faculty of Informatics Website

Email sources (?):

….

Difficulties

- Student-based sources
    - A lot of irrelevant, outdated, or misleading information.
    - Spam and off-topic discussions.
- University-based sources
    - Access restrictions (some info might require login).
    - Official information might not be up-to-date in real-time.
- Email sources (?)

….

How much data do we need?

We'll discuss it

Tools

1. Python 3.10

Installation

Mac:

- Go to python.org
- Download the latest macOS installer
- Open the .pkg file and follow the installation steps
- Verify installation:
  python3 --version

Windows:

- Download Python Installer
  - Go to python.org
  - Download the latest Windows installer (.exe file)
- Run the Installer
  - IMPORTANT: Check "Add Python to PATH" before clicking "Install Now"
  - Follow the installation steps
- Verify Installation
  - Open Command Prompt (cmd) and run:
    python --version
  - If it doesn't work, try:
    py --version

Revision

Freshen up your python skills -

Python for Beginners - Learn Coding with Python in 1 Hour

2. Python Libraries

Still not finalized!

Documents

- PDF Documents: PyMuPDF (also known as fitz)

- Word Documents: python-docx
- PowerPoint Presentations: python-pptx

University-Based Data

1. University of Debrecen, Faculty of Informatics

Description:
These platforms often provide data through web interfaces or APIs.

- requests: To handle HTTP requests for data retrieval.
- BeautifulSoup: For parsing HTML content when APIs are not available.
- pandas: To structure and analyze the extracted data.

2. Neptun System:

Description:
Direct data extraction may be restricted due to security and privacy policies. (?)

- Manual Data Export: Use built-in export functionalities to obtain data in formats like CSV or Excel.

Student-Based Data

1. WhatsApp

Description:
WhatsApp chats can be exported as .txt files, so you need to process them line by line.

- pandas → To organize extracted messages
- re (built-in) → To parse and clean text
- json (built-in) → If dealing with structured exports

2. Telegram

Description:
Telegram allows message exports in .json format, making it easier to process.

- telethon → If using the Telegram API to fetch messages
- json (built-in) → If using Telegram's export feature

3. Discord

Description:
Discord messages can be accessed via bots (requires permissions).

- discord.py → If fetching messages from a bot
- json (built-in) → If handling exported chat logs

Installation

TODO, when finalised

Windows

Mac

Revision

3. Jupyter Notebook

Before installing Jupyter Notebook, ensure Python is installed:

```
python3 --version  # macOS
python --version   # Windows
```

Before installing Jupyter Notebook, ensure pip is installed:

```
python -m pip --version  # Windows
python3 -m pip --version  # macOS
```

If not:

pip Installation

Windows:

1. Download get-pip.py:
   curl -O <https://bootstrap.pypa.io/get-pip.py>

2. Run the installer:
   python get-pip.py

Mac:

python3 -m ensurepip --default-pip

Or, if needed:

sudo apt install python3-pip # Ubuntu/Debian
brew install python # macOS (installs pip with Python)

Installation

Mac:

1. Install Jupyter using pip (Open Terminal and run):
   pip3 install jupyter

2. Run Jupyter Notebook:
   jupyter notebook

3. (Optional) Install JupyterLab (Advanced UI):
   pip3 install jupyterlab

Windows:

1. Install Jupyter using pip (Open Command Prompt (cmd) and run):
   pip install jupyter

2. Run Jupyter Notebook:
   jupyter notebook

3. (Optional) Install JupyterLab:
   pip install jupyterlab

Anaconda (Optional for both mac and windows):

1. Download Anaconda from anaconda.com
2. Install and open Anaconda Navigator

3. Launch Jupyter Notebook from there

Revision

Freshen up on Jupyter Notebook -

[Jupyter Notebook In 10 Minutes]