Assignment – NLP Data Curation

Course: Natural Language Processing
Deadline: 20th September
Submission: A compressed folder containing extracted JSON data + retrieved PDFs (if any) + short documentation.

Objective

The purpose of this assignment is to practice data extraction and curation from multiple real-world legal sources of Pakistan. Students are divided into 5 groups based on roll numbers, but this is not a group task. Each student must individually extract and compile the complete data assigned to their group.
You are allowed to use automated tools (e.g., Python libraries such as BeautifulSoup, Selenium, PyPDF2, pdfplumber, requests, Tesseract OCR).
Evaluation Criteria:

- Data quality
- Completeness
- JSON structure adherence

Bonus Marks: If you extract and include additional useful fields or metadata that are not present in the sample JSON files.

Submission Guidelines

1. Each student must submit their **own Code File, JSON file, PDF folder and Readme.txt in Zip format** named according to their roll number and assigned task.
2. Store **downloaded PDFs** (if any) in a separate folder named PDFs_<RollNo>.
3. Submit a **README.txt / README.md** with:
    - Tools used
    - Steps followed
    - Issues faced

Evaluation Criteria

- **40% Data Completeness** – Did you extract all required records?
- **30% JSON Formatting & Accuracy** – Does your output follow the given sample format?
- **20% File Organisation** – Proper folder structure, naming conventions, and documentation.
- **10% Innovation** – Bonus marks for extracting **extra-rich data** not mentioned in the sample files.

Final Note: Start early, as some sources involve **large datasets and scanned PDFs**. Submissions received after September 20th will not be accepted.

Roll No. G1

Source: Pakistan Code: https://pakistancode.gov.pk/english/LGu0xVD.php

- Extract **all laws and codes** available on the website.
- Follow the structure in **PakistanCode.json** (provided as an example).
- Store each law with **title, section, sub-sections, and full text**.
- Save the final file as: PakistanCode_<RollNo>.json.

Roll No. G2

Source: Islamabad High Court

Click on *Advanced Search* and select the initial date from which you want to extract the data up to the present. You will then get a list of cases. Store them in the provided format.

- Focus: **Argument mining** from **live court proceedings** (text-based & video-based).
- Extract: Case number, case title, parties involved, advocates, arguments (text), and, if possible, metadata from videos.
- Follow the structure in **Islamabad High Court.Json**
- Save the final file as: ISBHighCourt_<RollNo>.json.

  Additional marks for saving the data from this link: https://mis.ihc.gov.pk/frmCausLst from the initial date to the present.

Roll No. G3

Sources: Sindh High Court Case Search

Go to the **Sindh High Court Case Search Portal**.

On the main page, you will see the **major courts** (e.g., Karachi, Sukkur, Hyderabad, Larkana, etc.).

- For each **major court**, create a **separate JSON file** (e.g., SindhCourt_Karachi.json, SindhCourt_Sukkur.json, etc.).

After selecting a major court, you will be directed to another page where you must choose the court option from the dropdown menu.

- All cases from these dropdown courts must be collected and put in the same JSON file as the major court.
- Example: SindhCourt_Karachi.json contains all of the cases from Karachi's subcourts.

Each case must include:

- **Summary**
- **Details**
- **Tagline** (use "NA" if not available)

Make sure you capture fields like:

- Case No
- Case Year
- Case Category
- Bench
- Court
- Status
- Disposal Date
- Hearing Dates, etc.

Repeat the process for every **major court** on the main page until all JSON files are created.

Roll No. G4

Source: Supreme Court of Pakistan

- Useful Links: **Case Details**, **Judgments**
- For the Case Judgement, you must select the Year, then search and collect the data from 1980 to 2025.

  Each record contains:

  - **Case details** (subject, case no, title, judge, dates).
  - **Files** renamed into a structured folder (`judgments/judgment_caseNo.pdf`).
  - **File size** preserved.
  - **Optional tagline** included if available.

- For the Case Information link, you must provide the data for at least two fields, select Year and Registry, with all possible combinations for the collection of data

Folder convention for students

- `memopdfs/` → all petition/appeal memos, renamed as `memo_<CaseNo>.pdf`
- `judgementpdfs/` → all judgments/orders, renamed as `judgement_<CaseNo>.pdf`

- Extract metadata for each case:
  - Case Title, Case No, Case Subject, Author Judge, Status, Dates, Citations, Download links.
- Total records available: **~3326 judgments**.
- Follow the structure in **SupremeCourt_CaseInfo.Json** & **SupremeCourt_Judgments.Json**
- Save the final file as: SupremeCourt_<RollNo>.json.

Roll No. G5

Source: Federal Shariat Court – Laws

- Collect:
    - **9 Rules/Codes** (8–100 page PDFs)
    - **4 Acts/Ordinances** (4–10 page PDFs)
    - **3 Orders/Instructions** (~200 page PDFs)
- Follow the structure in **ShariahLaw.json**.
- Save final file as: ShariahLaw_<RollNo>.json. (**Remaining**)

Source: Reported Judgements

Source: Leading Judgements

- Extract metadata for **all reported judgements (12,530 cases)** and **20 leading judgements**.
- Fields: Case No, Case Title, Subject, Author Judge, Judgment Date, Citations, Download link.
- Follow the structure in **ShariahCourtJudgements.json**.
- Save final file as: ShariahJudgements_<RollNo>.json.

Sr.	RollNo	Name		RollNo	Name	
1	21I-0447	Usman Afzal		21I-0394	Fahad Jameel	
2	21I-0453	Awais Naeem		21I-0424	Ali Farooq	
3	21I-0457	Abdul Moiz		21I-0433	Nida Azam	
4	21I-0489	Ali Fayyaz		21I-0741	Laraib Noor	
5	21I-0499	Muhammad Ejaz		21I-0746	Abdullah Ashfaque	
6	21I-0520	Munib Akhtar		21I-0773	Abbas Ali	G1
7	21I-0547	Haider Rizvi		21I-0792	Jahan Zaib	
8	21I-0570	Taha Ahmed		21I-0810	Ajwaad Asghar	
9	21I-0720	Saim Mubarak		21I-1384	Ali Musharaf	
10	21I-0847	Babar Shaheen		21I-2520	Anas Imran	
11	21I-0848	Muhammad Ahmad		21I-2594	Shahzor Jamali	
12	21I-2532	Qazi Mohibunnabi		21I-2964	Shayan Salam	
13	21I-2987	Ataa Ur Rasool		21I-2975	Muhammad Huzaifa	
14	22I-0776	Wajahat Ullah		22I-0748	Zirwah Khalil	
15	22I-0785	Abdul Rehman		22I-0755	Arrij Fawwad	
16	22I-0788	Bilal Javed		22I-0805	Shaheer Zaman	
17	22I-0808	Alishba Naveed		22I-0811	Bilal Naveed	G2
18	22I-0809	Fiza Wajid		22I-0813	Hassan Imran	
19	22I-0825	Ahmed Ali		22I-0816	Amin Adnan	
20	22I-0835	Saad Mursaleen		22I-0870	Muhammad Taha	
21	22I-0894	Ali Salman		22I-0871	Nabeed Haider	
22	22I-0899	Ayesha Ejaz		22I-0874	Haseeb Sultan	
23	22I-0907	Anas Rashid		22I-0918	Hassaan Afzal	
24	22I-0919	Hasan Ali		22I-0928	Ibrahim Azhar	
25	22I-0933	Saif Ur Rehman		22I-0935	Qusai	
26	22I-0946	Zohaib Khan		22I-0961	Hamza Naveed	
27	22I-0948	Muhammad Rafay		22I-1017	Muneeb Ur Rehman	
28	22I-0959	Hafsa Imtiaz		22I-1047	Hamid Ali	G3
29	22I-0962	Sohaib Sattar		22I-1048	Shahmeer Ahmed	
30	22I-0965	Rehan Tariq		22I-1057	Azeem Ashfaq	
31	22I-1000	Adan Malik		22I-1098	Hamna Arshad	
32	22I-1007	Umer Farooq		22I-1125	Muaz Ahmed	
33	22I-1022	Muhammad Rayyan		22I-1129	Rimsha Azam	
34	22I-1040	Muhammad Khan		22I-1133	Umama Bajwa	
35	22I-1041	Abeer Jawad		22I-1145	Raima Tariq	
36	22I-1053	Huzaifa Nasir		22I-1155	Ubaida Tariq	
37	22I-1067	Ahmed Javed		22I-1165	Hamza Riaz	
38	22I-1130	Hamza Munim		22I-1171	Sheharyar Ahmed	
39	22I-1134	Ahmad Aqeel		22I-1177	Usman Haroon	G4
40	22I-1156	Arham Khalid		22I-1214	Haziq Naeem	
41	22I-1166	Athaar Fatima		22I-1227	Asjad Ullah	
42	22I-1179	Aamna Saeed		22I-1239	Tauha Imran	
43	22I-1190	Saad Nasim		22I-1280	Sarim Rasheed	
44	22I-1199	Ali Aamir		22I-1288	Ahmad	
45	22I-1244	Areeba Riaz		22I-1296	Eraj Zaman	
46	22I-1247	Zayyam Hassan		22I-1297	Bilal Tariq	
47	22I-1281	Aisha Siddiqa		22I-1305	Muhammad Danish	
48	22I-1308	Ambreen Arshad		22I-1326	Abdul Momin	
49	22I-1312	Saif Ullah		22I-1332	Najamuddin Hassan	
50	22I-1355	Waseemullah Zahid		22I-1333	Abdul Hadi	G5
51	22I-2330	Hamza Ijaz		22I-1338	Rabab Fatima	
52	22I-2427	Arshman Khawar		22I-1354	Ameer Hamza	
53	22I-8223	Sameed Ahmed		22I-2123	Muhammad Mujtaba	
54	22K-5082	Ayan Shahid		22I-2354	Ifrah Naseem	
56	22I-8222	Ammar Hussain		22I-8220	Shahzaib Rauf	