

```

!pip install -q -U bitsandbytes accelerate transformers pymupdf

import os
import fitz
import re
import json
import pandas as pd
from datetime import datetime
from google.colab import drive
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer,
BitsAndBytesConfig

drive.mount('/content/drive')

FOLDER_PATH = '/content/drive/MyDrive/modul_praktikum_sains_data'

model_id = "Qwen/Qwen2.5-1.5B-Instruct"

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.float16,
)

print(f"Sedang memuat AI {model_id}...")
try:
    tokenizer = AutoTokenizer.from_pretrained(model_id,
trust_remote_code=True)
    model = AutoModelForCausalLM.from_pretrained(
        model_id,
        quantization_config=bnb_config,
        device_map="auto",
        trust_remote_code=True
    )
except Exception as e:
    print("\n Error memuat model. Pastikan Anda menggunakan Runtime T4 GPU!")
    raise e

ABBREVIATION_MAP = {
    "Data Mining": "DM",
    "Pemodelan Stokastik": "PS",
    "Pergudangan Data": "PD",
    "Komputasi Paralel": "KP",
    "Analisis Data Statistik": "ADS",
    "Teknologi Basis Data": "TBD",
    "Basis Data": "BD",
    "Algoritma Pemrograman": "AP",
    "Deep Learning": "DL",

```

```

    "Machine Learning": "ML"
}

DAFTAR_MATKUL_AI = list(ABBREVIATION_MAP.keys())

def get_pdf_content(file_path):
    """Membaca teks halaman awal PDF untuk konteks AI"""
    try:
        doc = fitz.open(file_path)
        text = ""
        for i in range(min(2, len(doc))):
            text += doc[i].get_text()
        doc.close()
        return re.sub(r'\s+', ' ', text).strip()[:2500]
    except:
        return ""

def ask_ai(text_content):
    """Mengirim Prompt ke AI untuk ekstraksi JSON"""

    system_prompt = "Kamu adalah asisten akademik ahli Sains Data.  
Tugasmu mengekstrak metadata modul praktikum."

    user_prompt = f"""
    Analisis teks modul praktikum berikut:
    ---
    {text_content}
    ---

    Ekstrak informasi berikut dalam format JSON Valid:
    1. "mata_kuliah_full": Pilih SATU yang paling relevan dari daftar
    ini: {DAFTAR_MATKUL_AI}.
    2. "judul_topik": Judul spesifik modul ini. (HAPUS kata 'Modul',
    'Praktikum', 'Bab', 'Percobaan', atau Angka urutan di awal. Ambil inti
    judul teknisnya saja).
    3. "deskripsi": Buat ringkasan DETIL (3-4 kalimat). Jelaskan
    TUJUAN praktikum, METODE/ALGORITMA yang digunakan, dan TOOLS/LIBRARY
    yang dipakai (jika ada). Gunakan Bahasa Indonesia formal.

    Jawab HANYA JSON.
    """

    messages = [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": user_prompt}
    ]

    text = tokenizer.apply_chat_template(messages, tokenize=False,
    add_generation_prompt=True)
    model_inputs = tokenizer([text], return_tensors="pt").to("cuda")

```

```

generated_ids = model.generate(
    model_inputs.input_ids,
    max_new_tokens=450,
    temperature=0.1,
    do_sample=True
)

response = tokenizer.batch_decode(generated_ids,
skip_special_tokens=True)[0]

return response.split(user_prompt)[-1] if user_prompt in response
else response

def parse_json_response(response):
    """Membersihkan output AI menjadi Dictionary Python"""
    try:
        match = re.search(r'\{.*\}', response, re.DOTALL)
        if match: return json.loads(match.group(0))
    except: pass
    return None

if not os.path.exists(FOLDER_PATH):
    print(f"❌ Path Salah: {FOLDER_PATH}")
else:
    files = [os.path.join(FOLDER_PATH, f) for f in
os.listdir(FOLDER_PATH) if f.lower().endswith(".pdf")]
    results = []

    print(f"\n🔄 Memproses {len(files)} file PDF dengan AI...\n")

    for f_path in files:
        filename = os.path.basename(f_path)

        raw_text = get_pdf_content(f_path)
        if not raw_text: continue

        ai_raw = ask_ai(raw_text)
        data = parse_json_response(ai_raw)

        if data:
            full_mk = data.get("mata_kuliah_full", "Lainnya")

            l1_abbr = ABBREVIATION_MAP.get(full_mk, "OTH")

            l2_topic = data.get("judul_topik", "Topik Umum")
            desc = data.get("deskripsi", "Tidak ada deskripsi.")
        else:
            print(f"⚠️ Gagal parsing JSON: {filename}")
            l1_abbr, l2_topic, desc = "ERR", "Manual Check", "Error AI

```

## Response"

```
print(f"[ OK] {filename}")
print(f"    L1 (Code) : {l1_abbr}")
print(f"    L2 (Topik): {l2_topic}")
print("-" * 40)

results.append({
    'file_name': filename,
    'layer_1_code': l1_abbr,
    'layer_2_topic': l2_topic,
    'description': desc,
    'file_path': f_path,
    'created_at': datetime.now().strftime('%Y-%m-%d %H:%M:
%S'),
    'updated_at': datetime.now().strftime('%Y-%m-%d %H:%M:%S')
})

df = pd.DataFrame(results)
output_csv = '/content/drive/MyDrive/indexing_final_db.csv'
df.to_csv(output_csv, index=False)
print(f"\n Selesai! Data tersimpan di: {output_csv}")
```

```
44.0/44.0 kB 1.3 MB/s eta
0:00:00
59.4/59.4 MB 11.7 MB/s eta
0:00:00
12.0/12.0 MB 39.3 MB/s eta
0:00:00
24.1/24.1 MB 38.0 MB/s eta
0:00:00
emuat AI Qwen/Qwen2.5-1.5B-Instruct...
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/
_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your
settings tab (https://huggingface.co/settings/tokens), set it as
secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to
access public models or datasets.
warnings.warn(
```

```
{"model_id": "24a73a23b168404eb9f8716cc2ca1f36", "version_major": 2, "vers
ion_minor": 0}
```

```
{"model_id": "3a4a5359d9c44ad2bbca420d80807e76", "version_major": 2, "vers
ion_minor": 0}
```

```
{"model_id":"eae10248923b4cbd85e4b17b67271177","version_major":2,"version_minor":0}
```

```
{"model_id":"e1f2de9f23ab430392a1b07ad60f8a28","version_major":2,"version_minor":0}
```

```
{"model_id":"1b5803245e6a48b8a77e098de937ced7","version_major":2,"version_minor":0}
```

```
{"model_id":"c9b2da0628ed494094d0a5636dd3bf70","version_major":2,"version_minor":0}
```

```
{"model_id":"04af4027844c4ff3880e10d84cc62ae3","version_major":2,"version_minor":0}
```

□ Memulai indexing dari:  
/content/drive/MyDrive/modul\_praktikum\_sains\_data

□ Folder: Pergudangan Data | □ File: MODUL PRAKTIKUM 2 Pergudangan Data.pdf

The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your input's `attention\_mask` to obtain reliable results.

□ Topik: Desain Konseptual Pergudangan Data

-----  
□ Folder: Pergudangan Data | □ File: MODUL PRAKTIKUM 3 Pergudangan Data.pdf

□ Topik: Desain Logikal Gudang Data

-----  
□ Folder: Pergudangan Data | □ File: 1. MODUL PRATIKUM 4 - PHYSICAL DESIGN.pdf

□ Topik: Design Fisik Pergudangan Data

-----  
□ Folder: Pergudangan Data | □ File: dok-misi-dw.pdf

□ Topik: Misi 2: Desain Fisikal dan Development

-----  
□ Folder: Pemodelan Stokastik | □ File: modul 2-Pemodelan-Stokastik.pdf

□ Topik: Rantai Markov

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 1-Pemodelan-Stokastik.pdf

□ Topik: Proses Stokastik

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 4 - Proses Poisson.pdf

□ Topik: Proses Poisson

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 7 Pemodelan

Stokastik.pdf

□ Topik: Rantai Markov Waktu Kontinu

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 8 Praktikum Pemodelan Stokastik.pdf

□ Topik: Teori Antrean (M/M/1 dan M/M/c)

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 5 Pemodelan Stokastik.pdf

□ Topik: Distribusi Eksponensial dan Proses Poisson

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 6 Pemodelan Stokastik.pdf

□ Topik: Generalizations of the Poisson Process

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 3 Pemodelan Stokastik.pdf

□ Topik: Distribusi Eksponensial

-----  
□ Folder: Pemodelan Stokastik | □ File: Modul 9 Praktikum Pemodelan Stokastik.pdf

□ Topik: Limiting Probabilities and Time-Reversibility

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 5-Jackknife.pdf

□ Topik: Resampling Methods: Jackknife

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 1 Algoritma Iteratif .pdf

□ Topik: Algoritma Iteratif

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 4 Pengembangan Pemrograman Fungsi R.pdf

□ Topik: Pengembangan Pemrograman Fungsi R

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 3Pembangkitan Bilangan Acak.pdf

□ Topik: Pembangkitan Bilangan Acak

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 2.pdf

□ Topik: Data Wrangling

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 6 .pdf

□ Topik: Metode Numerik Integrasi

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 7 - Numerical Integration 2.pdf

□ Topik: Metode Numerik Integral: Simpson 3/8, Boole, dan Romberg

-----  
□ Folder: Komputasi Statistik | □ File: Modul Praktikum 8 - Simulasi

Monte Carlo.pdf

△ Gagal parsing JSON AI

□ Folder: Komputasi Statistik | □ File: Modul Praktikum 9 - Bootstrap.pdf

□ Topik: Resampling Methods: Bootstrap

□ Folder: Komputasi Paralel | □ File: 2. Modul Praktikum Komputasi Paralel.docx.pdf

□ Topik: Konsep Dasar Komputasi Paralel

□ Folder: Komputasi Paralel | □ File: 3. Modul Praktikum Komputasi Paralel.docx.pdf

□ Topik: Model Komputasi Paralel

□ Folder: Komputasi Paralel | □ File: 4. Modul Praktikum Komputasi Paralel1.docx.pdf

□ Topik: Algoritma Paralel

□ Folder: Komputasi Paralel | □ File: 5. Modul Praktikum Komputasi Paralel11.pdf

□ Topik: Komunikasi Sistem Paralel

□ Folder: Komputasi Paralel | □ File: 6. Modul Praktikum Komputasi Paralel111.pdf

□ Topik: Pemrograman Paralel Berbasis Shared Memory dengan Python (multiprocessing)

□ Folder: Komputasi Paralel | □ File: 9. Modul Praktikum Komputasi Paralel11111 (1).pdf

□ Topik: Komputasi Paralel dalam Machine Learning

□ Folder: Komputasi Paralel | □ File: 7. Modul 7.pdf

□ Topik: Message Passing Interface (MPI)

□ Folder: Komputasi Paralel | □ File: 8. Modul 8.pdf

□ Topik: Evaluasi Performa Paralel

□ Folder: Deep Learning | □ File: Modul 1 Prak Deep Learning .pdf

□ Topik: Feedforward Neural Network

□ Folder: Deep Learning | □ File: Modul 5 Prak Deep Learning .pdf

□ Topik: Convolutional Neural Networks (CNN)

□ Folder: Deep Learning | □ File: Modul 6 Prak Deep Learning .pdf

□ Topik: Modern CNN

□ Folder: Deep Learning | □ File: Modul 7 Prak Deep Learning .pdf

□ Topik: Long Short-Term Memory (LSTM) dan Gated Recurrent Unit

(GRU)

- □ Folder: Deep Learning | □ File: Modul 8 Prak Deep Learning .pdf  
□ Topik: Pengantar Attention Mechanism  
-----
- Folder: Deep Learning | □ File: Modul 2 Prak Deep Learning .pdf  
□ Topik: Backpropagation Neural Network  
-----
- Folder: Deep Learning | □ File: Modul 4 Prak Deep Learning .pdf  
□ Topik: Regularisasi pada Deep Learning  
-----
- Folder: Deep Learning | □ File: Modul 3 Prak Deep Learning .pdf  
□ Topik: Hyperparameter & Optimizer  
-----
- Folder: Data Mining | □ File: Modul 1 Praktikum Data Mining.pdf  
□ Topik: Data Understanding and Data Preparation  
-----
- Folder: Data Mining | □ File: Modul 2 Praktikum Data Mining.pdf  
□ Topik: Praktikum Data Mining - Data Cleaning & Integration  
-----
- Folder: Data Mining | □ File: Modul 3 Praktikum Data Mining.pdf  
□ Topik: Data Reduction  
-----
- Folder: Data Mining | □ File: Modul 4 Praktikum Data Mining.pdf  
□ Topik: Naive Bayes for Classification  
-----
- Folder: Data Mining | □ File: Modul 5 Praktikum Data Mining.pdf  
□ Topik: Decision Tree for Classification  
-----
- Folder: Data Mining | □ File: Modul 6 Praktikum Data Mining.pdf  
□ Topik: Unsupervised Learning: DBSCAN  
-----
- Folder: Data Mining | □ File: Modul 7 Praktikum Data Mining.pdf  
□ Topik: Association Rule Mining  
-----
- Folder: Data Mining | □ File: Modul 8 Praktikum Data Mining.pdf  
□ Topik: Deteksi Komunitas dengan NetworkX dan Algoritma Girvan-Newman  
-----
- Folder: Data Mining | □ File: Modul 9 Praktikum Data Mining.pdf  
□ Topik: Clustering & Outlier Detection  
-----
- Folder: Analisis Data Statistik | □ File: Modul-Praktikum-2.pdf  
□ Topik: Praktikum 2 Importing data and Visualizing Data  
-----
- Folder: Analisis Data Statistik | □ File: Modul-Praktikum-3.pdf  
□ Topik: Visualisasi Data Kategorik dan Transformasi Variabel Kategorikal  
-----



□ Folder: Analisis Data Statistik | □ File: Praktikum Analisis Regresi ADS.pdf  
□ Topik: Praktikum Analisis Regresi Linier Sederhana  
-----

□ Folder: Analisis Data Statistik | □ File: Praktikum-ANOVA.pdf  
□ Topik: Praktikum Analisis Data Statistik - Analisis Ragam (ANOVA)  
-----

□ Folder: Analisis Data Statistik | □ File: Modul-Praktikum-5.pdf  
□ Topik: Selang Kepercayaan  
-----

□ Folder: Analisis Data Statistik | □ File: Modul-Praktikum-6.pdf  
□ Topik: Uji Hipotesis  
-----

□ Folder: Algoritma Pemrograman | □ File: MODUL 1 variabel dan tipe data.pdf  
□ Topik: Variabel dan Tipe Data  
-----

□ Folder: Algoritma Pemrograman | □ File: 1. Modul 2- Value- assign-ekspresi- IO.pdf  
□ Topik: Input/Output dan Expressions  
-----

□ Folder: Algoritma Pemrograman | □ File: Modul 3 Praktikum Algoritma Pemrograman.pdf  
□ Topik: Pengkondisian dengan IF dalam Pemrograman Python  
-----

□ Folder: Algoritma Pemrograman | □ File: Modul Praktikum Python File-1-.pdf  
□ Topik: Menggunakan Python dengan File  
-----

□ Folder: Algoritma Pemrograman | □ File: Modul-Praktikum-6-Rekursif\_.pdf  
□ Topik: Fungsi Rekursif  
-----

□ Folder: Algoritma Pemrograman | □ File: Modul 4 Praktikum Algoritma Pemrograman.pdf  
□ Topik: Perulangan (Iterasi/Looping) dalam Bahasa Pemrograman Python 3  
-----

□ Folder: Algoritma Pemrograman | □ File: Modul 8 Praktikum Algoritma Pemrograman.pdf  
□ Topik: Pengenalan Algoritma Pemrograman  
-----

□ Folder: Analisis Multivariat | □ File: Modul\_Tutorial\_Analisis\_Korelasi\_Kanonik.pdf  
□ Topik: Analisis Korelasi Kanonik  
-----

□ Folder: Analisis Multivariat | □ File: Modul\_Tutorial\_Anmul\_Analisis\_Korespondensi.pdf  
□ Topik: Analisis Multivariat

-----  
□ Folder: Analisis Multivariat | □ File: Modul Analisis Biplot.pdf  
□ Topik: Biplot

-----  
□ Folder: Analisis Multivariat | □ File: Modul Tutorial Analisis  
Diskriminan.pdf  
□ Topik: Analisis Diskriminan

-----  
□ Folder: Analisis Multivariat | □ File: MODUL 5 ANALISIS  
MULTIVARIAT.pdf  
□ Topik: Analisis Profil

-----  
□ Folder: Analisis Multivariat | □ File: MODUL 4 ANALISIS  
MULTIVARIAT.pdf  
□ Topik: Analisis Multivariat - MANOVA

-----  
□ Folder: Analisis Multivariat | □ File: MODUL 1 ANALISIS  
MULTIVARIAT.pdf  
□ Topik: Analisis Multivariat Sebaran Normal Multivariat

-----  
□ Folder: Analisis Multivariat | □ File: MODUL 2 ANALISIS  
MULTIVARIAT.pdf  
□ Topik: Analisis Multivariat

-----  
□ Folder: Analisis Multivariat | □ File: MODUL 3 ANALISIS  
MULTIVARIAT.pdf  
□ Topik: MANOVA Satu Arah

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 2 Praktikum - Tuning  
Skema (1) (1).pdf  
□ Topik: Schema Tuning

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 3 Praktikum - Index  
Tunning (2).pdf  
□ Topik: Naive Bayes for Classification

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 5 Praktikum -  
Concurrency Control (KONTROL KONKURENSI) (1).pdf  
□ Topik: Concurrency Control

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 8 Praktikum - Graph  
Database (1) (1).pdf  
□ Topik: Graph Database

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 9 Praktikum - Document  
Database v2 (2) (1).pdf  
□ Topik: Document Database

-----  
□ Folder: Teknologi Basis Data | □ File: Modul 4 Praktikum.pdf

Folder: Teknologi Basis Data | File: Modul 6 Praktikum - Spatial Temporal Database (1).pdf

📁 Folder: Teknologi Basis Data | 📄 File: Modul 7 Praktikum.pdf

Folder: Teknologi Basis Data | File: Modul 1 Praktikum - Persiapan Tunning.pdf

□ SUKSES! 78 dokumen terindeks.

```
import pandas as pd
```

```
{
  "summary": {
    "name": "df",
    "rows": 78,
    "fields": [
      {
        "column": "layer_1_code",
        "properties": {
          "dtype": "category",
          "num_unique_values": 7,
          "samples": [
            "PD",
            "PS",
            "AMV"
          ],
          "semantic_type": "",
          "description": ""
        },
        "column": "subject_name",
        "properties": {
          "dtype": "category",
          "num_unique_values": 10,
          "samples": [
            "Analisis Multivariat",
            "Pemodelan Stokastik",
            "Data Mining"
          ],
          "semantic_type": "",
          "description": ""
        },
        "column": "layer_2_topic",
        "properties": {
          "dtype": "string",
          "num_unique_values": 76,
          "samples": [
            "Rantai Markov",
            "Backpropagation Neural Network",
            "Generalizations of the Poisson Process"
          ],
          "semantic_type": "",
          "description": ""
        },
        "column": "description",
        "properties": {
          "dtype": "string",
          "num_unique_values": 78,
          "samples": [
            "Modul ini membahas konsep dasar dari LSTM dan GRU, serta perbedaan mereka. LSTM adalah jenis RNN yang dirancang untuk mengatasi masalah vanishing gradient, sedangkan GRU adalah alternatif yang lebih sederhana. Mereka menggunakan struktur khusus seperti unit memori dan gerbang-gerbang untuk menyimpan dan mengambil informasi dari data berurutan. Metode ini sangat berguna dalam analisis dan prediksi berbasis data.",
            "Modul ini membahas tahap desain konseptual dalam pembangunan gudang data. Dalam hal ini, desain konseptual dirujuk pada pemahaman"
          ]
        }
      ]
    }
  }
}
```

kebutuhan bisnis, identifikasi proses bisnis, dan perancangan model data konseptual berupa pemetaan antara fakta dan dimensi. Mahasiswa diharapkan dapat memahami konsep dasar desain konseptual, mengidentifikasi kebutuhan data dari proses bisnis, dan membuat model konseptual dengan pemetaan fakta dan dimensi.

"Dalam modul praktikum ini, Anda belajar tentang pengantar pada mekanisme Attention. Dalam hal ini, Anda akan mempelajari definisi formal, komponen utama dari mekanisme Attention, serta proses normalisasi untuk mengatur bobot perhatian. Selain itu, Anda juga akan mengetahui bagaimana menggunakan mekanisme Attention dalam algoritma seperti Nadaraya-Watson Kernel Regression dan Scaled Dot-Product Attention. Perlu diingat bahwa bahasan ini melibatkan pemahaman umum tentang mekanisme Attention dan konsep-konsep dasarnya."

```
{
  "semantic_type": "",
  "description": "",
  "column": "file_name",
  "properties": {
    "dtype": "string",
    "num_unique_values": 78,
    "samples": [
      "Modul 7 Prak Deep Learning .pdf",
      "MODUL PRAKTIKUM 2 Pergudangan Data.pdf",
      "Modul 8 Prak Deep Learning .pdf"
    ],
    "semantic_type": "",
    "description": "",
    "column": "file_path",
    "properties": {
      "dtype": "string",
      "num_unique_values": 78,
      "samples": [
        "/content/drive/MyDrive/modul_praktikum_sains_data/Deep Learning /Modul 7 Prak Deep Learning .pdf",
        "/content/drive/MyDrive/modul_praktikum_sains_data/Pergudangan Data/MODUL PRAKTIKUM 2 Pergudangan Data.pdf",
        "/content/drive/MyDrive/modul_praktikum_sains_data/Deep Learning /Modul 8 Prak Deep Learning .pdf"
      ],
      "semantic_type": "",
      "description": "",
      "column": "created_at",
      "properties": {
        "dtype": "object",
        "num_unique_values": 78,
        "samples": [
          "2025-11-29 13:42:59",
          "2025-11-29 13:36:26",
          "2025-11-29 13:43:10"
        ],
        "semantic_type": "",
        "description": ""
      ]
    ]
  },
  "type": "dataframe",
  "variable_name": "df"
}
```