# Proyek UTS Praktikum Penelusuran Informasi
# UTS Project: Multi-Dataset Information Retrieval System (CLI-Based)

### Deskripsi Singkat

Pada proyek UTS ini, setiap tim akan membangun sistem Information Retrieval (IR) berbasis Command-Line Interface (CLI) yang mampu melakukan proses pencarian dan ranking dokumen dari berbagai sumber teks nyata.

Terdapat 5 dataset utama yang wajib digunakan:

1. `etd-usk` (tesis/disertasi Universitas Syiah Kuala)
2. `etd-ugm` (tesis/disertasi Universitas Gadjah Mada)
3. `kompas` (berita harian nasional)
4. `tempo` (majalah berita dan opini)
5. `mojok` (artikel populer dan satir)

Setiap tim bertanggung jawab membangun sistem yang dapat melakukan:

- Preprocessing dan tokenisasi teks.
- Representasi dokumen (BoW).
- Pembentukan index dokumen (Whoosh).
- Pencarian berbasis query pengguna.
- Ranking hasil berdasarkan `cosine similarity`.

### Tujuan

1. Memahami pipeline lengkap sistem penelusuran informasi dari preprocessing hingga ranking.
2. Mengintegrasikan konsep *Vector Space Model* ke dalam sistem nyata.
3. Melatih kemampuan kolaboratif dalam pengembangan sistem IR.
4. Meningkatkan ketepatan pencarian dengan pendekatan representasi teks yang efisien.

### Aturan Tim

- Proyek dikerjakan secara **kelompok berisi 3 orang**.
- Anggota dapat memilih tim sendiri.
- Tiap tim wajib membuat:
    1. **Kode sistem (Python, CLI-based)**.
    2. **Laporan UTS (PDF)**.
    3. **Readme (Dokumentasi)**.

**Instruksi Teknis**

1. Sistem dibangun dalam bahasa Python.
2. Antarmuka berupa **CLI interaktif**, dengan minimal menu:

```
=== INFORMATION RETRIEVAL SYSTEM ===
[1] Load & Index Dataset
[2] Search Query
[3] Exit
====================================
```

3. Gunakan `Whoosh` untuk indexing dan searching.
4. Gunakan `CountVectorizer` dan `cosine_similarity` untuk perhitungan kemiripan dokumen.
5. Lakukan **text preprocessing** minimal:
   - Case folding
   - Tokenization
   - Stopword removal
   - (Opsional: Stemming/Lemmatization)
6. Sistem harus mampu menangani query yang dimasukkan pengguna dan menampilkan 5 dokumen teratas dengan skor tertinggi.

**Dataset**

Setiap tim harus menggabungkan kelima dataset dalam satu folder:

```
datasets/
 |- etd-usk/
 |- etd-ugm/
 |- kompas/
 |- tempo/
 \- mojok/
```

**Kriteria Penilaian**

- Text Preprocessing dan Tokenisasi (15%)
- Representasi Dokumen (Bag of Words) (15%)
- Implementasi Indexing dengan Whoosh (25%)
- Pencarian dan Ranking menggunakan Cosine Similarity (25%)
- Laporan Proyek dan Dokumentasi (15%)
- Kerapian dan Struktur Kode (5%)

**Struktur Laporan**

Setiap tim wajib menyusun laporan dengan struktur sebagai berikut:

1. Cover
2. Pendahuluan
3. Desain Sistem dan Arsitektur
4. Implementasi (Kode & Penjelasan)
5. Pengujian Query dan Analisis Hasil
6. Kesimpulan

**Ketentuan Tambahan**

- Sistem hanya boleh menggunakan library Python standar + `scikit-learn`, `Whoosh`, dan `pandas`.
- Dilarang keras menyalin milik tim lain. Pelanggaran akan berakibat **nilai 0**.
- File akhir:
  - Kode Program
  - Laporan dalam format pdf
  - Readme untuk dokumentasi
- File dikumpul dalam format zip: `NoKelompok_UTS_Praktikum_PI.zip`.
  - Contoh: `01_UTS_Praktikum_PI.zip`.