UNIVERSITAS MUHAMMADIYAH SIDOARJO
Jl. Mojopahit 666 B, Telp. 031-8945444, Faks. 031-8949333 Sidoarjo - 61215
Email : umsida@umsida.ac.id | www.umsida.ac.id

SOAL UJIAN TENGAH SEMESTER

Mata Kuliah	**: Web & Teks Mining**		
Kode MK	**:** INF23748	**Tahun Akademik**	**: 2025/2026**
Semester	**: 7**	**Waktu Ujian**	**: 1 Pekan**
		Sifat Ujian	**: Terbuka**

Instruksi Umum

- Kerjakan project ini secara **Berkelompok (1 kelompok terdiri dari 4 anggota)**
- Waktu pengerjaan: **1 minggu**
- Kumpulkan dalam bentuk:
  - Laporan (PDF)
  - Source code (Python/R)
  - Dataset yang digunakan
- Presentasi hasil: max 15 menit via Youtube

PROJECT: Analisis Sentimen dan Preprocessing Teks dari Media Sosial/Berita

A. Deskripsi Project (Bobot: 100%)

Anda diminta untuk melakukan project text mining lengkap mulai dari pengambilan data hingga preprocessing. Pilih **salah satu** topik berikut:

1. Sentimen masyarakat terhadap kebijakan pemerintah
2. Review produk e-commerce
3. Komentar berita online
4. Tweet/posting media sosial tentang isu tertentu
5. Ulasan film/restoran

B. Tahapan Project dan Rubrik Penilaian

1. Pengambilan Data (Data Collection) - 25%

Tugas:

- Kumpulkan minimal **500 dokumen teks** dari sumber yang Anda pilih
- Gunakan salah satu metode:
  - Web Scraping (BeautifulSoup, Selenium, Scrapy)
  - API (Twitter API, Reddit API, dll)

Yang harus dijelaskan dalam laporan:

- Sumber data dan justifikasi pemilihan
- Metode pengambilan data (teknik dan tools)
- Kendala yang dihadapi dan solusinya
- Statistik deskriptif data mentah (jumlah dokumen, rata-rata panjang teks, dll)

- Kode yang digunakan dan penjelasannya

Poin penilaian:

- Kesesuaian sumber data dengan topik (5%)
- Kompleksitas teknik pengambilan data (10%)
- Dokumentasi kode dan penjelasan (5%)
- Jumlah dan kualitas data (5%)

---

2. Eksplorasi Data Awal (Initial Data Exploration) - 15%

Tugas:

- Lakukan analisis awal terhadap data mentah
- Identifikasi karakteristik data

Yang harus ditampilkan:

- Contoh 5-10 dokumen teks mentah
- Distribusi panjang teks (histogram)
- Identifikasi masalah data (missing values, duplikasi, noise, dll)
- Word cloud dari data mentah

Poin penilaian:

- Kelengkapan eksplorasi (7%)
- Visualisasi yang informatif (5%)
- Identifikasi masalah data (3%)

---

3. Text Preprocessing - 40%

Tugas: Lakukan preprocessing lengkap dengan tahapan berikut (minimal 6 tahapan):

a) Case Folding (wajib)

- Konversi ke lowercase

b) Cleaning (wajib)

- Hapus URL, mention, hashtag (jika dari media sosial)
- Hapus HTML tags (jika dari web scraping)
- Hapus special characters dan angka
- Hapus extra whitespace

c) Tokenization (wajib)

- Pemisahan teks menjadi token/kata

d) Stopwords Removal (wajib)

- Hapus kata-kata tidak bermakna
- Gunakan stopwords library atau buat custom stopwords
- Jelaskan pemilihan stopwords

e) Normalization (pilih minimal 1)

- Stemming (Sastrawi untuk Bahasa Indonesia / Porter/Lancaster untuk English)
- Lemmatization

- Slang/typo correction

f) Lainnya (opsional, nilai tambah)

- Spelling correction
- Emoji handling
- Emoticon conversion
- Named Entity Recognition
- Part-of-Speech Tagging

Yang harus dijelaskan:

- Alasan pemilihan setiap tahapan preprocessing
- Contoh transformasi teks di setiap tahapan (before-after)
- Perbandingan hasil dengan/tanpa setiap tahapan
- Pengaruh setiap tahapan terhadap data (statistik)
- Kode implementasi dengan komentar

Poin penilaian:

- Kelengkapan tahapan preprocessing (15%)
- Kesesuaian metode dengan jenis data (10%)
- Analisis perbandingan before-after (10%)
- Kualitas dokumentasi kode (5%)

---

4. Hasil Preprocessing dan Analisis - 15%

Tugas:

- Tampilkan hasil akhir preprocessing
- Lakukan analisis terhadap data bersih

Yang harus ditampilkan:

- Contoh 5-10 dokumen setelah preprocessing lengkap
- Perbandingan word cloud (sebelum vs sesudah)
- Top 20 kata paling sering muncul setelah preprocessing
- Statistik perubahan data (jumlah token, vocabulary size, dll)
- Simpan data bersih dalam format yang sesuai (CSV, JSON, dll)

Poin penilaian:

- Kualitas hasil preprocessing (8%)
- Analisis dan visualisasi (5%)
- Dokumentasi hasil (2%)

---

5. Laporan dan Dokumentasi - 5%

Format laporan harus mencakup:

1. Cover
2. Pendahuluan (latar belakang, tujuan, ruang lingkup)
3. Landasan teori (penjelasan setiap teknik yang digunakan)
4. Metodologi
5. Hasil dan pembahasan

6. Kesimpulan dan saran
7. Referensi
8. Lampiran (kode lengkap)

Poin penilaian:

- Struktur dan kelengkapan laporan (2%)
- Penulisan dan tata bahasa (2%)
- Kualitas referensi (1%)