SOAL 02 – Document Preprocessing

Tujuan

Tujuan dari tugas ini adalah untuk menerapkan tahapan *document preprocessing* pada kumpulan dokumen teks (korpus mini) agar siap digunakan dalam proses *information retrieval* atau *text mining*. Tahapan preprocessing ini membantu mengubah teks mentah menjadi bentuk yang lebih terstruktur dan mudah diolah oleh komputer.

1. Dataset yang Digunakan

Korpus mini terdiri dari 10 dokumen teks (.txt) yang berisi deskripsi singkat berbagai jenis buku, disimpan di folder data/raw/.
Contoh nama file:

- buku_fantasi.txt
- buku_fiksi_ilmiah.txt
- buku_filsafat.txt
- buku_horor.txt
- buku_komedi.txt
- buku_kriminal.txt
- buku_motivasi.txt
- buku_petualangan.txt
- buku_romansa.txt
- buku_sains.txt

Setiap dokumen berisi 1–2 kalimat pendek yang menggambarkan isi atau tema utama buku.

2. Tahapan Preprocessing yang Diterapkan

Proses *preprocessing* dilakukan melalui empat tahap utama:

a. Case Folding

Semua huruf diubah menjadi huruf kecil untuk menghindari perbedaan makna akibat kapitalisasi.
Contoh:

“Pedang Ajaib di Hutan” → “pedang ajaib di hutan”

b. Tokenisasi

Teks dipecah menjadi satuan kata (token) berdasarkan spasi.
Contoh:

“pedang ajaib di hutan” → [“pedang”, “ajaib”, “di”, “hutan”]

c. Stopword Removal

Kata-kata umum yang tidak memiliki makna penting (seperti “yang”, “di”, “dan”, “adalah”) dihapus menggunakan daftar *stopwords Bahasa Indonesia* dari NLTK.
Contoh:

[“pedang”, “ajaib”, “di”, “hutan”] → [“pedang”, “ajaib”, “hutan”]

d. Stemming

Kata dikembalikan ke bentuk dasarnya menggunakan *PorterStemmer* (atau *Sastrawi Stemmer* untuk Bahasa Indonesia).
Contoh:

[“berlari”, “bermain”, “menemukan”] → [“lari”, “main”, “temu”]

3. Hasil Preprocessing (Before–After)

Berikut adalah perbandingan dua contoh dokumen sebelum dan sesudah dilakukan preprocessing.

Dokumen 1 — buku_fantasi.txt

Before:

“Seorang anak laki-laki bernama Arka menemukan pedang ajaib yang tersembunyi di hutan terlarang. Dengan bantuan penyihir tua, ia harus menyelamatkan kerajaan dari naga hitam.”

After:

anak lakilaki bernama arka menemukan pedang ajaib tersembunyi hutan terlarang bantuan penyihir tua menyelamatkan kerajaan naga hitam

Dokumen 2 — buku_fiksi_ilmiah.txt

Before:

“Di masa depan, manusia hidup berdampingan dengan robot yang memiliki emosi. Namun ketika kecerdasan buatan mulai memberontak, dunia berada di ambang kehancuran.”

After:

manusia hidup berdampingan robot memiliki emosi kecerdasan buatan memberontak dunia ambang kehancuran

Dokumen 3 — buku_filsafat.txt

Before:

“Buku ini membahas pertanyaan tentang makna hidup, kebebasan, dan moralitas. Setiap bab mengajak pembaca untuk merenungkan pandangan hidup mereka.”

After:

buku membaha makna hidup kebebasan moralita bab mengajak pembaca merenungkan pandangan hidup

Dokumen 4 — buku_horor.txt

Before:

“Sebuah rumah tua di tepi kota menyimpan misteri kematian yang tak terungkap. Setiap malam, suara tangisan dan langkah kaki terdengar dari ruang bawah tanah.”

After:

rumah tua tepi kota menyimpan misteri kematian terungkap malam suara tangisan langkah kaki terdengar ruang tanah

Dokumen 5 — buku_komedi.txt

Before:

“Petualangan lucu seorang remaja yang selalu terlibat dalam kejadian konyol di sekolah. Setiap bab menghadirkan tawa sekaligus pelajaran berharga.”

After:

petualangan lucu remaja terlibat kejadian konyol sekolah bab menghadirkan tawa pelajaran berharga

Dokumen 6 — buku_kriminal.txt

Before:

“Seorang detektif muda menyelidiki kasus pembunuhan misterius di kota kecil. Petunjuk demi petunjuk mengarah pada seseorang yang tak pernah ia duga.”

After:

detektif muda menyelidiki pembunuhan misteriu kota petunjuk petunjuk mengarah duga

Dokumen 7 — buku_motivasi.txt

Before:

“Buku ini mengajarkan pentingnya berpikir positif dan pantang menyerah. Kesuksesan dimulai dari kebiasaan kecil yang dilakukan setiap hari.”

After:

buku mengajarkan berpikir positif pantang menyerah kesuksesan kebiasaan

Dokumen 8 — buku_petualangan.txt

Before:

“Tiga sahabat melakukan perjalanan menantang ke puncak gunung tertinggi. Mereka menghadapi badai, jurang, dan ketakutan terbesar demi mencapai tujuan.”

After:

sahabat perjalanan menantang puncak gunung tertinggi menghadapi badai jurang ketakutan terbesar mencapai tujuan

Dokumen 9 — buku_romansa.txt

Before:

“Kisah cinta dua insan yang terhalang oleh jarak dan waktu. Mereka berjuang mempertahankan hubungan di tengah kesibukan dan ambisi masing-masing.”

After:

kisah cinta insan terhalang jarak berjuang mempertahankan hubungan kesibukan ambisi masingmasing

Dokumen 10 — buku_sains.txt

Before:

“Penulis menjelaskan fenomena alam dengan cara yang mudah dipahami. Dari gerhana matahari hingga misteri lubang hitam di luar angkasa.”

After:

penuli fenomena alam mudah dipahami gerhana matahari misteri lubang hitam angkasa

4. Ringkasan Hasil Preprocessing

Berikut ini log ringkas hasil preprocessing untuk seluruh dokumen:

Dokumen 1 — buku_fantasi.txt

10 token paling sering:

[('anak', 1), ('lakilaki', 1), ('bernama', 1), ('arka', 1), ('menemukan', 1), ('pedang', 1), ('ajaib', 1), ('tersembunyi', 1), ('hutan', 1), ('terlarang', 1)]

Jumlah token total: 17

Dokumen 2 — buku_fiksi_ilmiah.txt

10 token paling sering:

[('manusia', 1), ('hidup', 1), ('berdampingan', 1), ('robot', 1), ('memiliki', 1), ('emosi', 1), ('kecerdasan', 1), ('buatan', 1), ('memberontak', 1), ('dunia', 1)]

Jumlah token total: 12

Dokumen 3 — buku_filsafat.txt

10 token paling sering:

[('hidup', 2), ('buku', 1), ('membahas', 1), ('makna', 1), ('kebebasan', 1), ('moralitas', 1), ('bab', 1), ('mengajak', 1), ('pembaca', 1), ('merenungkan', 1)]

Jumlah token total: 12

Dokumen 4 — buku_horor.txt

10 token paling sering:

[('rumah', 1), ('tua', 1), ('tepi', 1), ('kota', 1), ('menyimpan', 1), ('misteri', 1), ('kematian', 1), ('terungkap', 1), ('malam', 1), ('suara', 1)]

Jumlah token total: 16

Dokumen 5 — buku_komedi.txt

10 token paling sering:

[('petualangan', 1), ('lucu', 1), ('remaja', 1), ('terlibat', 1), ('kejadian', 1), ('konyol', 1), ('sekolah', 1), ('bab', 1), ('menghadirkan', 1), ('tawa', 1)]

Jumlah token total: 12

Dokumen 6 — buku_kriminal.txt

10 token paling sering:

[('petunjuk', 2), ('detektif', 1), ('muda', 1), ('menyelidiki', 1), ('pembunuhan', 1), ('misterius', 1), ('kota', 1), ('mengarah', 1), ('duga', 1)]

Jumlah token total: 10

Dokumen 7 — buku_motivasi.txt

10 token paling sering:

[('buku', 1), ('mengajarkan', 1), ('berpikir', 1), ('positif', 1), ('pantang', 1), ('menyerah', 1), ('kesuksesan', 1), ('kebiasaan', 1)]

Jumlah token total: 8

Dokumen 8 — buku_petualangan.txt

10 token paling sering:

[('sahabat', 1), ('perjalanan', 1), ('menantang', 1), ('puncak', 1), ('gunung', 1), ('tertinggi', 1), ('menghadapi', 1), ('badai', 1), ('jurang', 1), ('ketakutan', 1)]

Jumlah token total: 13

Dokumen 9 — buku_romansa.txt

10 token paling sering:

[('kisah', 1), ('cinta', 1), ('insan', 1), ('terhalang', 1), ('jarak', 1), ('berjuang', 1), ('mempertahankan', 1), ('hubungan', 1), ('kesibukan', 1), ('ambisi', 1)]

Jumlah token total: 11

Dokumen 10 — buku_sains.txt

10 token paling sering:

[('penulis', 1), ('fenomena', 1), ('alam', 1), ('mudah', 1), ('dipahami', 1), ('gerhana', 1), ('matahari', 1), ('misteri', 1), ('lubang', 1), ('hitam', 1)]

Jumlah token total: 11