

PERTEMUAN 2

Representasi Teks dan Text Processing

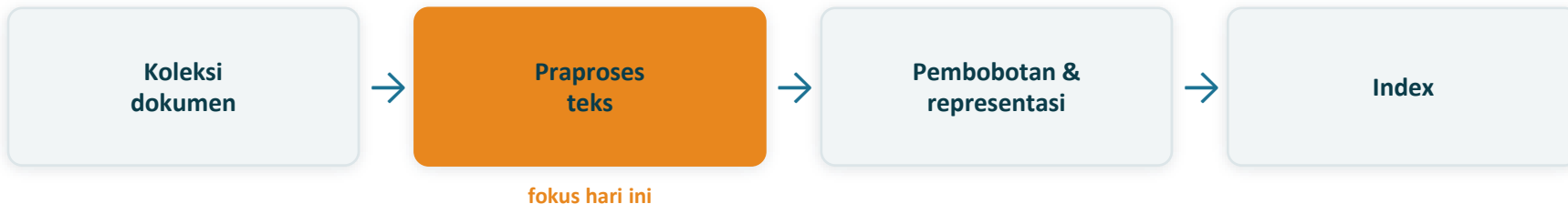
Tokenisasi, normalisasi, stopwords, stemming, dan lemmatization untuk korpus Bahasa Indonesia dan Inggris

token

term

index

Di Mana Kita Berada Sekarang



Mengapa tahap ini menentukan?

Index hanya dapat menemukan apa yang dimasukkan ke dalamnya. Jika kata “membacakan” dan “baca” tersimpan sebagai dua term berbeda, tidak ada model perangkikan secanggih apa pun yang dapat menghubungkan keduanya saat query. Keputusan praproses bersifat permanen, jika kita mengubahnya berarti kita perlu membangun ulang seluruh index.

***Strict rule:** praproses pada dokumen dan pada query harus identik. Ketidakcocokan keduanya menjadi sumber bug paling sering muncul di sistem IR.*

Capaian Pembelajaran Pertemuan Ini

Sub-CPMK-2 — turunan dari CPMK-07.1 dan CPL-TF-07

Sub-CPMK-2

Mahasiswa mampu menerapkan teknik pemrosesan teks pada korpus berbahasa Indonesia dan Inggris.

Ranah: C3 (menerapkan), P3 (presisi)

Indikator penilaian

- 1 Ketepatan memilih tahapan praproses sesuai karakteristik korpus
- 2 Ketepatan implementasi tokenization, stopwords removal, dan stemming
- 3 Kebenaran hasil praproses pada korpus berbahasa Indonesia

Bentuk penilaian: laporan praktikum (formatif). Hasil praktikum ini menjadi modul yang dipakai ulang pada pertemuan 3 dan 4.

Dari *Raw Text* Menjadi Term

Enam tahapan

1

Dokumen mentah

HTML, PDF, DOCX, JSON; masih bercampur markup dan metadata

2

Ekstraksi & encoding

Ambil *plain text*, pastikan memiliki encoding UTF-8, deteksi bahasa dokumen

3

Tokenisasi

Pecah aliran karakter menjadi satuan token yang bermakna

4

Normalisasi

Case folding, penyeragaman ejaan, penanganan angka dan tanda baca

5

Penyaringan stopwords

Buang kata berfrekuensi sangat tinggi yang tidak membedakan dokumen

6

Stemming / lemmatization

Petakan varian bentuk kata ke satu bentuk dasar → hasilnya adalah term

Tidak semua sistem memakai keenam tahap. Pilihan tahap adalah keputusan rancangan, bukan sebuah resep baku.

Empat Istilah yang Sering Tertukar

Karakter, token, tipe, dan term → perbedaannya menentukan cara Anda menghitung

1

Karakter

Satuan terkecil aliran teks. Perhatikan encoding: satu huruf beraksen dapat memakan lebih dari satu byte.

2

Token

Kemunculan satu satuan teks pada posisi tertentu di dokumen. Kata yang muncul lima kali menghasilkan lima token.

3

Tipe

Kelas token yang unik. Lima kemunculan kata yang sama hanya dihitung sebagai satu tipe.

4

Term

Tipe yang lolos seluruh tahap praproses dan benar-benar dicatat dalam index.

Yang disimpan di dalam index adalah term, bukan token, bukan kata sebagaimana ditulis penulisnya.

Menghitung Token, Tipe, dan Term

Satu kalimat dapat memiliki empat angka yang berbeda

Banjir melanda Surabaya. Banjir itu melanda tiga kecamatan.

Karakter

60

Termasuk spasi dan tanda titik

Token

9

Banjir, melanda, Surabaya, Banjir, itu, melanda, tiga, kecamatan → setiap kemunculan dihitung

Tipe

7

“Banjir” dan “melanda” masing-masing muncul dua kali, jadi dihitung sekali

Term

5

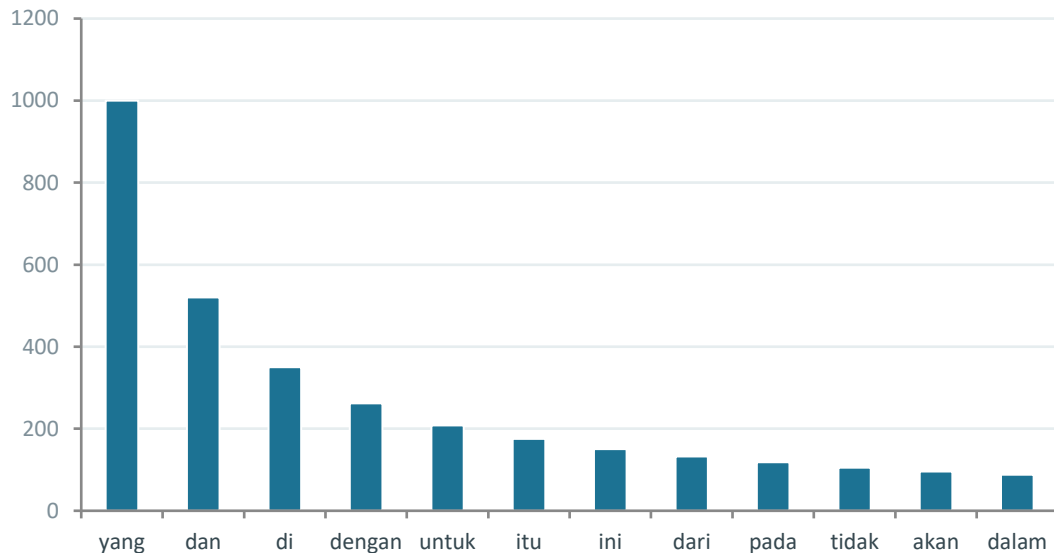
Setelah case folding, pembuangan “itu” sebagai stopword, dan stemming

Hasil akhir → [banjir] [landa] [surabaya] [tiga] [camat]

Hukum Zipf: Sedikit Kata Menguasai Segalanya

Frekuensi sebuah kata kira-kira berbanding terbalik dengan peringkat frekuensinya

Frekuensi 12 kata teratas (ilustrasi)



Grafik bersifat ilustratif untuk memperlihatkan bentuk kurva, bukan hasil pengukuran korpus tertentu.

Yang perlu diingat

- Kurva turun sangat curam: kata peringkat kedua muncul sekitar separuh dari peringkat pertama.
- Sebagian kecil kosakata menyumbang sebagian besar seluruh token dalam korpus.
- Sebaliknya, sebagian besar kata hanya muncul satu atau dua kali di seluruh koleksi.

Tiga Konsekuensi Praktis Hukum Zipf

Bentuk kurva ini menjelaskan hampir seluruh keputusan rancangan pada tahap praproses

1

Reasonable Stopwords

Kata yang paling sering muncul di hampir semua dokumen sehingga hampir tidak membantu membedakan satu dokumen dari lainnya.

2

Long tails are expensive

Kata yang hanya muncul sekali menyumbang sebagian besar ukuran kamus, tetapi jarang dipakai dalam query.

3

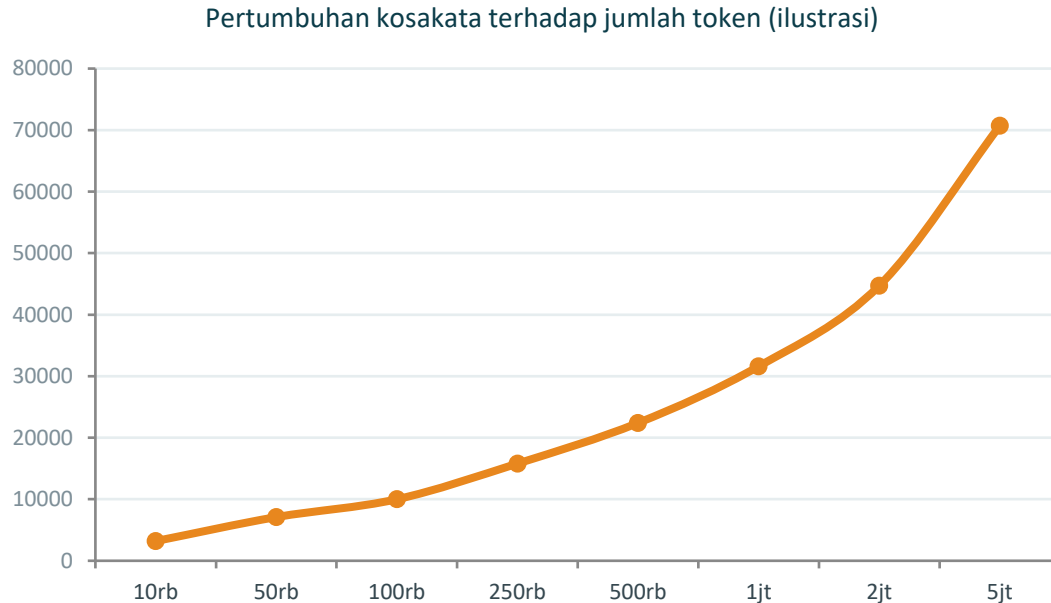
Very effective compression

Distribusi yang sangat timpang membuat postings list dapat dikompresi secara agresif → dibahas di pertemuan berikutnya.

Pertanyaan renungan: jika kata “yang” dibuang, bagaimana nasib query “yang muda yang bercerita”?

Hukum Heaps: Kosakata Terus Bertambah

Ukuran kamus tumbuh mengikuti $V = k \times N^b$, dengan b berkisar sekitar 0,5



Sumbu horizontal menunjukkan jumlah token yang telah diproses; nilai bersifat ilustratif.

Implikasi

- Menambah dokumen selalu memunculkan kata baru: nama orang, istilah teknis, salah ketik.
- Kamus tidak dapat diasumsikan selalu berukuran tetap saat merancang struktur index.
- Stemming dan normalisasi menekan laju pertumbuhan ini secara signifikan.

Tokenisasi Tidak Sesederhana Memecah pada Spasi

Setiap keputusan di bawah ini akan menentukan apakah query pengguna nanti dapat menemukan dokumen

Tanda hubung

state-of-the-art

Satu token, tiga token, atau keduanya sekaligus?

Apostrof

O'Brien, Ma'aruf, isn't

Nama orang dan bentuk singkat memerlukan aturan berbeda

Angka & tanggal

3/12/2026, Rp1.500.000

Format lokal berbeda; angka sering justru penting dalam query

URL & email

sinta.kemdiktisaintek.go.id

Dipecah pada titik akan menghancurkan maknanya

Nama majemuk

New York, Jawa Timur

Dua token yang bermakna hanya jika berdampingan

Bahasa tanpa spasi

Tionghoa, Jepang, Thai

Memerlukan segmentasi kata berbasis kamus atau model

Tidak ada jawaban yang benar secara universal, yang penting adalah aturan yang sama diterapkan pada dokumen dan pada query.

Tantangan Khas Bahasa Indonesia

Empat pola yang tidak ditemui pada korpus berbahasa Inggris

1

Kata ulang

anak-anak, berlari-lari

Tanda hubung menandai reduplikasi, bukan pemisah kata. Memecahnya menghasilkan token “anak” ganda yang menyesatkan frekuensi.

2

Imbuhan produktif

mempertanggungjawabkan

Satu kata dasar dapat melahirkan puluhan bentuk turunan. Tanpa stemming, semuanya menjadi term terpisah.

3

Kata majemuk

rumah sakit, kereta api

Dua token yang maknanya hilang jika dipisah. Perlu daftar frasa atau indexing posisi.

4

Campur kode

meeting-nya di-cancel

Akar bahasa Inggris dengan imbuhan Indonesia, stemmer baku sering gagal menanganinya.

Konsekuensi: pustaka tokenisasi berbahasa Inggris tidak dapat dipakai apa adanya untuk korpus Indonesia.

Case Folding dan Normalisasi

Menyeragamkan bentuk penulisan, dengan konsekuensi yang harus disadari

Yang diuntungkan

- Query “Banjir” menemukan dokumen “banjir”
- Ukuran kamus menyusut cukup besar
- Pengguna tidak dituntut mengetik dengan rapi

Yang dikorbankan

- “Padang” nama kota melebur dengan “padang” rumput
- Akronim melebur dengan kata biasa: “BUMN”, “Ada”
- Sinyal nama diri untuk information extraction hilang

Bentuk normalisasi lain yang lazim

Tanda baca

Dibuang, kecuali yang bermakna

Diakritik

café → cafe

Angka

Distandarkan atau dibuang

Spasi ganda

Dirapatkan menjadi satu

Praktik yang disarankan: simpan bentuk asli dokumen secara terpisah agar cuplikan hasil pencarian tetap terbaca oleh pengguna.

Penyaringan Stopword

Membuang kata berfrekuensi tinggi yang tidak membedakan dokumen, namun tidak selalu aman

Contoh stopwords Indonesia: yang dan di ke dari untuk pada dengan ini itu adalah akan

Alasan membuang

- Menghemat ukuran postings secara berarti
- Mempercepat pemrosesan query
- Nilai pembeda antar-dokumen mendekati nol

Alasan mempertahankan

- Query frasa: “yang muda yang bercerita”
- Negasi penting: “tidak” pada analisis sentimen
- Sistem modern mampu menangannya lewat pembobotan

Kecenderungan mutakhir

Sistem IR modern cenderung memakai daftar stopwords yang sangat pendek atau tidak sama sekali. Pembobotan TF-IDF dan BM25 sudah otomatis memberi bobot nyaris nol pada kata yang muncul di hampir semua dokumen → dibahas pada pertemuan 4 dan 5.

Stemming atau Lemmatization?

Dua cara menyatukan varian bentuk kata menjadi satu term

Aspek	Stemming	Lemmatization
Cara kerja	Memotong imbuhan berdasarkan aturan	Menganalisis morfologi dan kelas kata dengan kamus
Hasil	Potongan kata, belum tentu kata yang sah	Selalu berupa kata dasar yang sah
Contoh Inggris	studies → studi	studies → study
Contoh Indonesia	kecamatan → camat	kecamatan → kecamatan
Kecepatan	Sangat cepat	Jauh lebih lambat
Kebutuhan sumber daya	Cukup seperangkat aturan	Kamus dan penanda kelas kata
Lazim dipakai untuk	Sistem IR skala besar	Tugas NLP yang menuntut ketelitian makna

Untuk IR, stemming hampir selalu menjadi pilihan pertama: kesalahan kecil dapat ditoleransi selama diterapkan konsisten pada dokumen dan query.

Porter Stemmer untuk Bahasa Inggris

Algoritma berbasis aturan yang telah menjadi rujukan sejak 1980

Aturan diterapkan bertahap, dari akhiran paling panjang ke paling pendek

sses → ss	caresses → caress
ies → i	ponies → poni
s → ∅	cats → cat
ational → ate	relational → relate

Sifat penting

Tidak memakai kamus sama sekali. Cepat, ringan, dan mudah diterapkan ulang (tetapi buta terhadap makna).

Contoh penyatuan varian

connect

connected

connecting

connection

connections



connect

Lima term berbeda menyusut menjadi satu → recall naik tanpa menambah beban penyimpanan.

Stemming Bahasa Indonesia: Nazief-Adriani

Algoritma yang menjadi dasar pustaka Sastrawi (berbasis kamus dan urutan pelepasan imbuhan)

1	Cek kamus	Bila kata sudah ada dalam kamus kata dasar, proses berhenti
2	Buang partikel	-lah, -kah, -tah, -pun
3	Buang kata ganti milik	-ku, -mu, -nya
4	Buang sufiks derivasi	-i, -an, -kan
5	Buang prefiks derivasi	di-, ke-, se-, te-, be-, me-, pe- beserta aturan peluluhan
6	Recoding & fallback	Kembalikan huruf yang luluh; bila gagal, kembalikan kata semula

Contoh penelusuran

```
mengalahkan
↓ buang -kan
mengalah
↓ buang meng-
alah
↓ recoding: k luluh
kalah ✓
```

*Tanpa aturan recoding, hasilnya berhenti di “alah”
→ kata dasar yang keliru.*

```
pip install Sastrawi
```

Urutan pelepasan tidak boleh dibalik: membuang prefiks lebih dulu akan merusak kata seperti “beruang” dan “beruap”.

Ketika Korpusnya Adalah Media Sosial

Empat masalah yang membuat pipeline standar gagal total

Slang & singkatan

gpp, bgt, yg, dgn, sy

Perlu kamus normalisasi kata tidak baku

Perpanjangan huruf

baguuuusss, mantaaap

Rapatkan huruf berulang sebelum stemming

Emoji & emotikon

😄 :) xD

Buang, atau petakan menjadi token bermakna

Tagar & mention

#banjirsurabaya @akun

Pisahkan tanda, pertimbangkan pemecahan tagar

Praproses bukan resep universal, tetapi harus dirancang mengikuti karakteristik korpus → inti indikator penilaian pertama hari ini.

Setiap Pilihan Punya Konsekuensi

Praproses agresif menaikkan recall, praproses konservatif menjaga precision

Praproses agresif

Stemming kuat, stopword banyak dibuang, seluruh huruf dikecilkan.

Recall naik: lebih banyak dokumen relevan tertangkap.

Precision turun: dokumen tidak relevan ikut terjaring.

Praproses konservatif

Tanpa stemming, stopword dipertahankan, huruf besar dijaga.

Precision naik: hasil lebih tepat sasaran.

Recall turun: varian bentuk kata tidak dapat ditemukan.

Contoh: stemming menyatukan “pendidikan” dan “didik” sehingga query yang satu menemukan dokumen yang lain. Tetapi ia juga menyatukan “kecamatan” dengan “camat”, dua hal yang jelas berbeda bagi pengguna.

Pilihan praproses tidak boleh asal tebak, tapi perlu diukur (cara pengukuran dibahas pada pertemuan 6)

Praktikum: Pipeline Praproses Korpus Berita

Dikerjakan individu, dikumpulkan melalui Ilmu2 dan Notebook Jupyter Google Colab Anda

1

Siapkan korpus

Kumpulkan minimal 50 artikel berita berbahasa Indonesia dalam berkas CSV atau JSON

2

Bangun fungsi praproses

Satu fungsi yang menerima teks mentah dan mengembalikan daftar term

3

Terapkan tiap tahap

Case folding, tokenisasi, pembuangan stopword, stemming dengan Sastrawi

4

Ukur dampaknya

Catat jumlah token dan tipe sebelum dan sesudah setiap tahap

5

Buat grafik Zipf

Plot frekuensi terhadap peringkat untuk korpus Anda sendiri

6

Tulis analisis singkat

Tahap mana yang paling banyak memangkas kosakata, dan apa risikonya?

```
import nltk, re
```

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

Rangkuman & Persiapan Pertemuan Berikutnya

- Yang tersimpan dalam index adalah term yang merupakan hasil olahan, bukan kata sebagaimana ditulis penulis dokumen.
- Hukum Zipf menjelaskan mengapa stopword ada; hukum Heaps menjelaskan mengapa ukuran kamus tidak pernah berhenti tumbuh.
- Bahasa Indonesia menuntut penanganan khusus: kata ulang, imbuhan produktif, kata majemuk, dan campur kode.
- Praproses pada dokumen dan pada query wajib identik, dan pilihannya harus diukur, bukan ditebak.

Persiapan Pertemuan 3

Baca Manning dkk. Bab 1 dan 3. Bawa hasil praktikum hari ini. Daftar term Anda akan langsung dipakai untuk membangun inverted index.

Pustaka pertemuan ini

- [1] Manning, Raghavan & Schütze (2008), Bab 2
- [7] Zhai & Massung (2016), Bab 3
- [12] Budiharto, Knowledge & Information Retrieval