

**BIDANG PENDIDIKAN DAN PENGAJARAN
BERITA ACARA PERKULIAHAN
PERIODE SEMESTER GANJIL 2025/2026**



Dosen:

RIADI MARTA DINATA (0320087704)

MATAKULIAH:

MA1524 - Artificial Intelligence 3 - 3 SKS

LAMPIRAN BERITA ACARA PERKULIAHAN :

1. SK.DEKAN FSTI SEMESTER GANJIL 2025/2026
2. PRESENSI KEHADIRAN DOSEN DAN MATERI AJAR
3. NILAI KOMULATIF; KEHADIRAN,TUGAS, UTS DAN UAS
4. CONTOH HAND OUT MATERI AJAR

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS TERAPAN DAN TEKNOLOGI
INSTITUT SAINS DAN TEKNOLOGI NASIONAL ISTN**



INSTITUT SAINS DAN TEKNOLOGI NASIONAL

Jl. Moch. Kahfi II No.RT.13, RT.13/RW.9, Srengseng Sawah, Kec. Jagakarsa, Kota Jakarta Selatan, DKI Jakarta

Website : www.istn.ac.id / e-Mail : admin@istn.ac.id / Telepon : (021) 7270090

JURNAL PERKULIAHAN MATEMATIKA 2025 GANJIL

MATA KULIAH : Artificial Intelligence 3
NAMA DOSEN : RIADI MARTA DINATA, S.TI., M.Kom.
KREDIT/SKS : 3 SKS
KELAS : A

TATAP MUKA KE	HARI/TANGGAL	MULAI	SELESAI	RUANG	STATUS	RENCANA MATERI	REALISASI MATERI	KEHADIRAN MHS	PENGAJAR	TANDA TANGAN
1	Selasa, 23 September 2025	22:00	23:30	R-C1	Selesai	Pengantar Artificial Intelligence III, ruang lingkup NLP, serta tinjauan ulang dasar AI dan konsep vektor dalam pemrosesan teks. Mahasiswa diperkenalkan pada hubungan antara AI dan NLP, serta bagaimana NLP menjadi subdisiplin penting dalam pemahaman bahasa alami. Diperkenalkan pipeline umum NLP dari input teks mentah hingga output hasil analisis. https://gasal25.istncoffee.my.id/Ai3 (https://gasal25.istncoffee.my.id/Ai3)	Mahasiswa memahami keterkaitan AI, Machine Learning, dan NLP. Dilakukan diskusi terbimbing mengenai representasi teks secara matematis (vektor kata), contoh aplikasi NLP seperti analisis sentimen dan chatbot. Praktik awal dilakukan dengan Python untuk menampilkan frekuensi kata. https://gasal25.istncoffee.my.id/Ai3/sesi1.html (https://gasal25.istncoffee.my.id/Ai3/sesi1.html)	(4 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
2	Selasa, 30 September 2025	22:00	23:30	R-C1	Selesai	Proses pra-pemrosesan teks: tokenisasi, case folding, dan normalisasi. Mahasiswa belajar tentang pentingnya pembersihan data teks sebelum dimasukkan ke model.	Mahasiswa mengimplementasikan tokenisasi dan normalisasi pada dataset teks sederhana (ulasan produk). Diperlihatkan efek perbedaan tokenisasi terhadap jumlah kata unik dan distribusi frekuensi kata. https://gasal25.istncoffee.my.id/Ai3/sesi2.html (https://gasal25.istncoffee.my.id/Ai3/sesi2.html)	(4 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	

3	Selasa, 7 Oktober 2025	22:00	23:30	R-C1	Selesai	Konsep stopword, stemming, dan lemmatization. Mahasiswa memahami bagaimana mengurangi kata tidak penting dan menyatukan bentuk kata untuk efisiensi analisis.	Implementasi stemming bahasa Indonesia dengan Sastrawi dan lemmatization bahasa Inggris dengan spaCy. Diperbandingkan hasil stemming vs lemmatization dan efeknya terhadap hasil klasifikasi sederhana. https://gasal25.istncoffee.my.id/Ai3/sesi3.html (https://gasal25.istncoffee.my.id/Ai3/sesi3.html)	(3 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
4	Selasa, 14 Oktober 2025	22:00	23:30	R-C1	Selesai	Representasi teks menggunakan Bag of Words (BoW) dan TF-IDF. Pembahasan konsep bobot kata, term frequency, dan inverse document frequency.	Mahasiswa membuat vektor TF-IDF dari dataset berita pendek. Analisis dilakukan terhadap kata dengan bobot tertinggi serta visualisasi vektor dengan PCA untuk menunjukkan dimensi fitur. https://gasal25.istncoffee.my.id/Ai3/sesi4.html (https://gasal25.istncoffee.my.id/Ai3/sesi4.html)	(6 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
5	Selasa, 21 Oktober 2025	22:00	23:30	R-C1	Selesai	Cosine Similarity dan penerapannya dalam sistem pencarian teks. Mahasiswa belajar menghitung kemiripan antar dokumen dan membuat sistem pencarian sederhana berbasis TF-IDF.	Mahasiswa membangun sistem pencarian mini yang mampu menerima query dan menampilkan peringkat dokumen terdekat menggunakan cosine similarity. https://gasal25.istncoffee.my.id/Ai3/sesi5.html (https://gasal25.istncoffee.my.id/Ai3/sesi5.html)	(6 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
6	Selasa, 28 Oktober 2025	22:00	23:30	R-C1	Selesai	memahami dan menerapkan K-Nearest Neighbors (KNN) untuk klasifikasi teks di ruang TF-IDF / embedding, memilih k, membandingkan metrik jarak (<i>cosine vs euclidean</i>), dan mengevaluasi dengan metrik klasifikasi.	Learning Outcomes: (1) Menjelaskan prinsip KNN & pengaruh k; (2) Menerapkan KNN untuk teks dengan metrik yang sesuai; (3) Melakukan validasi silang untuk memilih hyperparameter; (4) Membaca confusion matrix & metrik (precision/recall/F1). https://gasal25.istncoffee.my.id/Ai3/sesi6.html (https://gasal25.istncoffee.my.id/Ai3/sesi6.html)	(3 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
7	Selasa, 4 November 2025	22:00	23:30	R-C1	Selesai	Regresi untuk Klasifikasi: Logistik & Linear (One-vs-Rest) Tujuan: membangun baseline kuat untuk klasifikasi teks dengan Regresi Logistik (dan regresi linear via SGDClassifier / LinearSVC preview), menerapkan regularisasi L1/L2, menyesuaikan ambang keputusan, dan melakukan probability calibration.	Learning Outcomes: (1) Memformulasikan logit & loss logistik; (2) Menerapkan regulasi dan pemilihan C/alpha; (3) Tuning ambang berbasis metrik/biaya; (4) Membaca koefisien fitur sebagai pentingnya fitur; (5) Kalibrasi probabilitas (Platt/Isotonic). https://gasal25.istncoffee.my.id/Ai3/sesi7.html (https://gasal25.istncoffee.my.id/Ai3/sesi7.html)	(3 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	

8	Selasa, 18 November 2025	22:00	23:30	R-C1	Selesai	UTS & Review Materi (Sessi 1–7) UTS menguji pemahaman dan keterampilan praktik dari praproses hingga baseline klasifikasi. Format campuran: teori singkat + praktik notebook.	Learning Outcomes: (1) Menjelaskan konsep kunci (tokenisasi/normalisasi, stopword/stemming/lemma, BoW/TF–IDF, cosine, KNN, LogReg); (2) Menyusun pipeline end-to-end pada dataset baru; (3) Menafsirkan metrik evaluasi dan koefisien fitur. https://gasal25.istncoffee.my.id/Ai3/sessi8.html (https://gasal25.istncoffee.my.id/Ai3/sessi8.html)	(3 / 6)	RIADI MARTA DINATA, S.TI., M.Kom.	
---	--------------------------	-------	-------	------	---------	---	--	---------	-----------------------------------	--



INSTITUT SAINS DAN TEKNOLOGI NASIONAL

Jl. Moch. Kahfi II No.RT.13, RT.13/RW.9, Srengseng Sawah, Kec. Jagakarsa, Kota Jakarta Selatan, DKI Jakarta

Website : www.istn.ac.id / e-Mail : admin@istn.ac.id / Telepon : (021) 7270090

LAPORAN DAFTAR NILAI MAHASISWA

Program Studi S1 Matematika

Periode 2025 Ganjil

Mata kuliah : Artificial Intelligence 3
Pengajar : RIADI MARTA DINATA, S.TI., M.Kom.

Nama Kelas : A
Sistem Kuliah :

No	NIM	NAMA	NILAI	NILAI ANGKA	NILAI HURUF	KET.
1	23310001	ANNISA WIDYADHANA	90.00	4.00	A	
2	23310002	SYAUQI MUHAMMAD	85.00	4.00	A	
3	23310003	SRI PUJIATI	80.00	4.00	A	
4	23310004	MUHAMMAD KEIVAN ANANAMI	80.00	4.00	A	
5	23310005	LINDA NURHALIZA	90.00	4.00	A	
6	23310006	NASRANI NEHE	84.00	4.00	A	



INSTITUT SAINS DAN TEKNOLOGI NASIONAL

Jl. Moch. Kahfi II No.RT.13, RT.13/RW.9, Srengseng Sawah, Kec. Jagakarsa, Kota Jakarta Selatan, DKI Jakarta

Website : www.istn.ac.id / e-Mail : admin@istn.ac.id / Telepon : (021) 7270090

LAPORAN NILAI PERKULIAHAN MAHASISWA

Program Studi S1 Matematika

Periode 2025 Ganjil

Mata kuliah : Artificial Intelligence 3

Nama Kelas : A

Kelas / Kelompok :

Kode Mata kuliah : MA1524

SKS : 3

No	NIM	Nama Mahasiswa	TUGAS INDIVIDU (20,00%)	UTS (40,00%)	UAS (40,00%)	Nilai	Grade	Lulus	Sunting KRS?	Info
1	23310001	ANNISA WIDYADHANA	90.00	90.00	90.00	90.00	A	✓		
2	23310002	SYAUQI MUHAMMAD	85.00	85.00	85.00	85.00	A	✓		
3	23310003	SRI PUJIATI	80.00	80.00	80.00	80.00	A	✓		
4	23310004	MUHAMMAD KEIVAN ANANAMI	80.00	80.00	80.00	80.00	A	✓		
5	23310005	LINDA NURHALIZA	90.00	90.00	90.00	90.00	A	✓		
6	23310006	NASRANI NEHE	80.00	80.00	90.00	84.00	A	✓		
Rata-rata nilai kelas			84.17	84.17	85.83	84.83	A			

Pengisian nilai untuk kelas ini ditutup pada **Selasa, 10 Februari 2026** oleh **202401465**

Tanggal Cetak : Jumat, 13 Maret 2026, 17:03:24

Paraf Dosen :

RIADI MARTA DINATA, S.Ti., M.Kom.

Prof. Dr. Bambang Soegijono, M.Si.



INSTITUT SAINS DAN TEKNOLOGI NASIONAL

Jl. Moch. Kahfi II No.RT.13, RT.13/RW.9, Srengseng Sawah, Kec. Jagakarsa, Kota Jakarta Selatan, DKI Jakarta

Website : www.istn.ac.id / e-Mail : admin@istn.ac.id / Telepon : (021) 7270090

LAPORAN PERSENTASE PRESENSI MAHASISWA

MATEMATIKA

2025 GANJIL

Mata kuliah : Artificial Intelligence 3

Nama Kelas : A

Dosen Pengajar : RIADI MARTA DINATA, S.TI., M.Kom.

No	NIM	Nama	Kelas / Kelompok	Pertemuan	Alfa	Hadir	Ijin	Sakit	Persentase
Peserta Reguler									
1	23310001	ANNISA WIDYADHANA		16	2	13	1		81.25
2	23310002	SYAUQI MUHAMMAD		16	3	12	1		75
3	23310003	SRI PUJIATI		16	3	8	5		50
4	23310004	MUHAMMAD KEIVAN ANANAMI		16	3	13			81.25
5	23310005	LINDA NURHALIZA		16	3	13			81.25
6	23310006	NASRANI NEHE		16	4	12			75

Jakarta, 13 Maret 2026

Ketua Prodi Matematika

Drs. KURNIAWAN ATMADJA, M.Si.

NIP. 200905-001



INSTITUT SAINS DAN TEKNOLOGI NASIONAL

Jl. Moch. Kahfi II No.RT.13, RT.13/RW.9, Srengseng Sawah, Kec. Jagakarsa, Kota Jakarta Selatan, DKI

Jakarta

Website : www.istn.ac.id / e-Mail : admin@istn.ac.id / Telepon : (021) 7270090

REKAP PRESENSI DOSEN

Periode : 2025 Ganjil Prodi : Matematika
Mata Kuliah : MA1524 - Artificial Intelligence 3 Nama Kelas : A
Tgl : 22 September 2025 sd 10 Januari 2026 SKS : 3
Perkuliahan

No	NIP	Nama	Alfa	Hadir	%
1	202103-002	RIADI MARTA DINATA, S.TI., M.Kom.	0	15	93.75 %

Keterangan

A : Alfa

H : Hadir

Jakarta, 13 Maret 2026

Ketua Prodi Matematika

Drs. KURNIAWAN ATMADJA, M.Si.

NIP. 200905-001

Artificial Intelligence III — MA1524

Fokus Natural Language Processing

Mata Kuliah ini membahas lanjutan konsep AI dengan penekanan pada penerapan **Machine Learning dan Deep Learning** dalam konteks *Natural Language Processing (NLP)*. Pembelajaran berlangsung selama 16 sesi dengan pendekatan berbasis praktik melalui Google Colab dan studi kasus dunia nyata.

Informasi Umum

- **Program Studi:** S1 Matematika — Fakultas Sains Terapan dan Teknologi (FSTT)
- **Dosen Pengampu:** Riadi Marta Dinata
- **Semester:** Ganjil 2025/2026
- **Bobot:** 3 SKS
- **Kode MK:** MA1524

Setiap sesi terdiri dari paparan teori singkat, studi kasus, serta implementasi mandiri di Google Colab.

Daftar Sesi Materi (Klik untuk Akses)

Sesi 1	Sesi 2	Sesi 3	Sesi 4
Sesi 5	Sesi 6	Sesi 7	Sesi 8
Sesi 9	Sesi 10	Sesi 11	Sesi 12
Sesi 13	Sesi 14	Sesi 15	Sesi 16

Tujuan Pembelajaran

- Memahami konsep lanjutan AI yang relevan dengan pemrosesan bahasa alami.
- Menerapkan metode Machine Learning (KNN, SVM, Regresi) pada data teks.
- Mengembangkan model deep learning sederhana untuk NLP (LSTM, CNN teks).
- Menganalisis hasil model dengan metrik evaluasi dan audit etika (fairness, explainability).
- Menyusun proyek akhir terintegrasi berbasis studi kasus dunia nyata.

Sesi 1 — Pengantar AI3 & Fokus NLP

Tujuan sesi: memahami posisi NLP dalam AI, meninjau matematik dasar (vektor & probabilitas kata), dan melihat alur proyek NLP dari data mentah hingga keluaran model.

Learning Outcomes: (1) Menjelaskan peran NLP dalam AI; (2) Menjabarkan pipeline NLP; (3) Memahami konsep ruang vektor teks dan frekuensi kata; (4) Menjalankan eksplorasi awal korpus.

1) Definisi & Gambaran Umum

Artificial Intelligence (AI) adalah studi dan rekayasa sistem yang mampu melakukan tugas yang biasanya membutuhkan kecerdasan manusia. **Natural Language Processing (NLP)** adalah cabang AI yang berfokus pada pemrosesan dan pemahaman teks/bahasa manusia. Dalam kursus ini, kita menekankan pendekatan *statistical & vector space* untuk merepresentasikan teks, sehingga dapat diolah oleh algoritma pembelajaran mesin.

Pemodelan Teks sebagai Vektor

Kalimat/dokumen diubah menjadi vektor fitur (mis. frekuensi kata). Setiap dimensi mewakili istilah tertentu; besarnya nilai menunjukkan pentingnya istilah dalam dokumen. Hal ini memungkinkan perhitungan *jarak* atau *kesamaan* antar dokumen.

Pipeline NLP (ringkas): Kumpulan data → Pra-proses (tokenisasi, normalisasi, stopword, stemming/lemmatization) → Representasi (BoW/TF-IDF/Embedding) → Model (KNN/SVM/Regresi) → Evaluasi (confusion matrix, F1) → Analisis kesalahan & iterasi.

Ringkasan Matematika

- Vektor & Norma:** $|\vec{z}| = \sqrt{\sum_i z_i^2}$
- Cosine Similarity:** $\cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|}$
- Term Frequency (TF):** proporsi kemunculan istilah pada dokumen.
- Inverse Document Frequency (IDF):** bobot yang menurunkan pengaruh istilah yang terlalu umum.

2) Studi Kasus Nyata — Eksplorasi Korpus Campuran (ID+EN)

Kita mulai dari korpus multi-topik tanpa label untuk eksplorasi awal. Target: memahami distribusi kata, menemukan istilah dominan, dan menyiapkan data untuk sesi-sesi berikutnya (TF-IDF, similarity, klasifikasi).

Contoh Korpus (≥ 30 Dokumen)

Korpus ini mencakup ulasan produk, berita singkat, opini singkat, dan deskripsi layanan (campuran Indonesia/Inggris). Anda dapat memperluas atau mengganti dengan data kelas Anda.

```
corpus = [
    "Pengiriman cepat, kualitas barang sangat baik dan sesuai deskripsi.",
    "Battery life is impressive; however, the screen glare is noticeable outdoors.",
    "Layanan pelanggan responsif, tetapi proses retur memakan waktu.",
    "The movie had stunning visuals but a weak storyline.",
    "Harga terjangkau, performa mantap untuk kebutuhan harian.",
    "Network latency spikes during peak hours affected our dashboard.",
    "Rasa kopi kuat dan aromanya enak, kemasan juga rapi.",
    "App update fixed the crash but introduced login delays.",
    "Fitur kamera malam bagus, namun stabilisasi video kurang.",
    "Documentation is thorough, examples help onboarding.",
    "Pengalaman belanja menyenangkan, voucher potongan harga membantu sekali.",
    "Server downtime last night impacted API integrations.",
    "Tekstur kue lembut, rasa tidak terlalu manis sesuai selera keluarga.",
    "Delivery was late; packaging was dented but item worked fine.",
    "Antarmuka aplikasi intuitif, navigasi mudah dipahami pemula.",
    "Customer support promised a callback but never followed up.",
    "Kualitas kain adem, ukuran sesuai chart, jahitan rapi.",
    "Dataset imbalance requires stratified sampling for evaluation.",
    "Suara speaker jernih, bass cukup, volume maksimal tidak pecah.",
    "Checkout process failed at payment gateway intermittently.",
    "Konektivitas Bluetooth stabil; jarak efektif sekitar 8 meter.",
    "The hotel staff were friendly, but the room smelled of smoke.",
    "Keyboard travel nyaman, backlight konsisten, keycaps anti-slip.",
    "Performa menurun setelah update; cache clear membantu sementara.",
    "Shipping fee too high for small accessories.",
    "Resep mudah diikuti, waktu memasak sesuai estimasi.",
    "Price-to-performance is excellent for students.",
    "Layar AMOLED tajam, warna hidup, tingkat kecerahan tinggi.",
    "Refund diproses cepat setelah komplain diajukan.",
    "Weather data ingestion requires timezone normalization."
]
```

Korpus di atas bersifat *unlabeled* (tanpa label). Pada sesi berikutnya Anda akan menambahkan pipeline praproses, representasi TF-IDF/embedding, dan model.

3) Praktik di Google Colab — Eksplorasi Awal

Salin blok kode berikut ke Google Colab. Setiap sel mandiri dan berurutan agar mudah diikuti mahasiswa.

A. Setup & Data

```
!pip -q install pandas numpy scikit-learn nltk matplotlib Sastrawi

import re, math, collections
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# ----- DATASET (≥30 dokumen campuran ID+EN) -----
corpus = [
    # (gunakan daftar dari blok sebelumnya atau gabungkan di sini)
]

# Jika ingin menambah dokumen sendiri:
# corpus.extend(["tambah kalimat di sini", ...])

print(f"Total dokumen: {len(corpus)}")

# Simpan ke CSV untuk dokumentasi eksperimen
pd.DataFrame({"text": corpus}).to_csv("corpus_sesi1.csv", index=False)
print("Tersimpan: corpus_sesi1.csv")
```

B. Tokenisasi & Normalisasi Sederhana

```
import nltk
nltk.download('punkt', quiet=True)

# Tokenizer sederhana (huruf, angka, tanda hubung)
TOKEN_RE = re.compile(r"[A-Za-zÀ-Öø-ÿ0-9_\\-']+")

def tokenize(text):
    return [t.lower() for t in TOKEN_RE.findall(text)]

TOKENS = [tokenize(doc) for doc in corpus]
TOKENS[:2]
```

C. Frekuensi Kata & Visualisasi Top Terms

```
# Hitung frekuensi global
from collections import Counter
freq = Counter([tok for doc in TOKENS for tok in doc])

# 25 istilah teratas (tanpa stopwords formal dulu)
TOP = pd.DataFrame(freq.most_common(25), columns=["term", "freq"])
print(TOP)

plt.figure(figsize=(10,5))
plt.bar(TOP["term"], TOP["freq"])
plt.xticks(rotation=60, ha='right')
plt.title('Top 25 Frekuensi Istilah (pra-stopword)')
plt.tight_layout(); plt.show()
```

D. TF, DF, dan Skor TF-IDF Manual (Intuisi)

```
# Document Frequency (DF)
from math import log

DF = Counter()
for doc in TOKENS:
    for t in set(doc):
        DF[t] += 1

N = len(TOKENS)

def tf(doc):
    c = Counter(doc)
    L = len(doc) if len(doc)>0 else 1
    return {t: c[t]/L for t in c}

# Hitung TF-IDF sederhana
VOCAB = sorted(DF.keys())
IDX = {t:i for i,t in enumerate(VOCAB)}

import numpy as np
X = np.zeros((N, len(VOCAB)), dtype=float)
for i,doc in enumerate(TOKENS):
    tfd = tf(doc)
    for t,v in tfd.items():
        idf = math.log((N+1)/(DF[t]+1)) + 1 # smoothed IDF
        X[i, IDX[t]] = v * idf

print("Matriks TF-IDF:", X.shape)
```

E. Cosine Similarity Dokumen vs Query

```
from numpy.linalg import norm

def cosine(a,b):
    na, nb = norm(a)+1e-12, norm(b)+1e-12
    return float(np.dot(a,b)/(na*nb))

# Query: ketik apa pun yang relevan
query = "pengiriman cepat dan kualitas bagus"
q_tokens = tokenize(query)
q_vec = np.zeros((len(VOCAB),), float)
qt = Counter(q_tokens)
L = sum(qt.values()) or 1
for t,cnt in qt.items():
    if t in IDX:
        idf = math.log((N+1)/(DF[t]+1)) + 1
        q_vec[IDX[t]] = (cnt/L) * idf

sims = [cosine(X[i], q_vec) for i in range(N)]
rank = np.argsort(sims)[::-1][:5]

print("Query:", query)
for r in rank:
    print(f"sims={sims[r]:.3f}", corpus[r])
```

F. Catatan Lanjut

- Di Sesi 2–3, kita tambahkan *stopword*, *stemming*, dan *lemmatization*.
- Di Sesi 4–5, kita gunakan **TF-IDF** dari `scikit-learn` dan membangun *search* sederhana.
- Di Sesi 6–11, kita masuk ke model klasifikasi (KNN, Regresi, SVM) dan *sentiment analysis*.
- Embedding seperti **word2vec** dan pendekatan tanpa label awal akan diperkenalkan menjelang Sesi 10 & 15.

4) Rangkuman & Checklist

- Paham posisi NLP dalam AI dan gambaran pipeline end-to-end.
- Mampu memuat (≥30) dokumen teks mentah untuk eksplorasi.
- Menghitung frekuensi kata dan membuat visualisasi sederhana.
- Mengerti intuisi TF, DF, dan skor TF-IDF serta penggunaan cosine similarity untuk pencarian.

← Sebelumnya

Berikutnya →

Sessi 2 — Tokenisasi & Normalisasi (Pra-Pemrosesan Teks)

Tujuan: merancang pipeline pra-proses yang *reproducible*, mengurangi *noise* (URL, emoji, angka, elongasi), dan menyiapkan data untuk TF-IDF/embedding pada sesi berikutnya.

- Learning Outcomes:** (1) Menjelaskan peran tokenisasi & normalisasi; (2) Mengimplementasikan aturan normalisasi yang konsisten; (3) Mengukur dampak pra-proses terhadap ukuran kosakata & TTR; (4) Menyimpan korpus bersih untuk sesi 3–5.

1) Definisi Singkat

Tokenisasi memecah teks menjadi unit (token) seperti kata, angka, atau simbol.
Normalisasi menyeragamkan bentuk token: huruf kecil (*case folding*), penghapusan URL, penyamaan *elongation* ("baaaagus" → "bagus"), penggantian angka/emoji dengan placeholder (<num> , <emoji>), dan standardisasi ejaan umum ("nggak"→"tidak").

Prinsip Desain

- Task-aware*: aturan mengikuti tugas (sentimen vs topik).
- Deterministik & terdokumentasi: hasil sama untuk input yang sama.
- Non-destruktif bila ragu: simpan versi asli berdampingan.

Metrix Dampak Pra-Proses

- |V|**: ukuran kosakata
- TTR**: type-token ratio
- Rasio angka/URL/emoji → berapa yang dinormalisasi
- Distribusi panjang token

2) Dataset (≥ 30 Dokumen Campuran ID+EN)

Pakai korpus multi-domain yang lebih *noisy* agar aturan normalisasi berdampak nyata. Anda boleh menambah dokumen kelas sendiri.

```
corpus = [
    "Pengiriman 2x lebih cepat dr ekspetasi!!! link: https://toko.id/a/123 🛒",
    "Battery-life ~10hrs; screen-glare outdoors :( #annoying",
    "Pelayanan CS oke bgt, tp refund > 7 hari ya? email: cs@brand.co",
    "This movie is sooo goooood!!! Must watch!!!",
    "Harga 1.299.000 jd 1.099.000 pas promo 11.11 🎉",
    "Server down @ 02:13 UTC; logs @ s3://bucket/app/logs",
    "Rasa kopi mantapppp!! ☕ kemasan rapi, ekspedisi cepat.",
    "Update v2.1 fixed crash; but login delay ~3-5s remains.",
    "UI/UX oke, on-boarding jelas; dark-mode 🔥",
    "Dokumennya lengkap bgt → examples jelas, thanks!",
    "Packing kurang rapih; item aman sih. Kurir telat 1 hari.",
    "Support promised callback @Mon, never happened...",
    "Keyboard travel enakkkk; backlight ok; keycaps anti-slip.",
    "Kualitas kain adem bgt; ukuran pas; jahitan rapiii-",
    "Bluetooth stable ~8m; beyond that stutter.",
    "Staff ramah; kamar wangi? hmm bau rokok dikit :/",
    "Checkout gagal (PG timeouts). Retry 3x baru bisa.",
    "Harga murah; performa ok buat mahasiswa.",
    "Stabilisasi video kurang; night-mode mantap.",
    "Delivery was late; packaging dented; item OK tho.",
    "Resep gampang; 20' jadi. Anak2 suka.",
    "Top-ups via e-wallet failed 2x; SMS OTP telat.",
    "Refund diproses <24h setelah komplain. mantap!",
    "Weather ingestion needs tz normalization (Asia/Jakarta).",
    "Docs thorough; examples accelerate onboarding.",
    "Price/perf = excellent; would rec to classmates.",
    "App cache clear helps; after update still slow :(",
    "Kopi pait? menurutku balanced kok ☺",
    "Layar AMOLED tajam; nits tinggi; glare minim.",
    "Shipping fee too high for small items..."
]
```

3) Praktik Google Colab — Pipeline Pra-Proses

Salin per blok ke Colab. Setiap bagian independen dan bisa dieksekusi berurutan.

A. Setup

```
!pip -q install pandas numpy scikit-learn nltk regex emoji Unidecode

import re, regex as re2, math, unicodedata
import numpy as np, pandas as pd
import matplotlib.pyplot as plt
from collections import Counter
from emoji import replace_emoji
from unidecode import unidecode
import nltk
nltk.download('punkt', quiet=True)

pd.set_option('display.max_colwidth', 120)

# Muat korpus
# (Anda dapat impor dari CSV: pd.read_csv('corpus_sessi1.csv')['text'].tolist())
```

B. Aturan Normalisasi yang Dapat Dikonfigurasi

```
# Placeholder standar
PLACEHOLDERS = {"URL":"","EMAIL":"","USER":"","NUM":"","EMOJI":""}

URL_RE   = re2.compile(r"https?:\/\/S+[s3:\/\S+", re2.I)
EMAIL_RE = re2.compile(r"[\w.-]+@[ \w.-]+\.\w.-]+", re2.I)
USER_RE  = re2.compile(r"@ \w+")
NUM_RE   = re2.compile(r"(\d+)
```

C. Tokenisasi & Perbandingan Varian

```
def tokenize_regex(text):
    return [m.group(0) for m in TOKEN_RE.finditer(text)]

# NLTK (umum); untuk ID bisa tetap pakai regex custom
from nltk.tokenize import word_tokenize

toks_raw = [tokenize_regex(t) for t in corpus]

toks_norm_regex = [tokenize_regex(t) for t in norm]

toks_norm_nltk = [word_tokenize(t) for t in norm]

print("Contoh token RAW :", toks_raw[0][:12])
print("Contoh token NORM R", toks_norm_regex[0][:12])
print("Contoh token NORM N", toks_norm_nltk[0][:12])
```

D. Mengukur Dampak Pra-Proses

```
def vocab_size(tokens_list):
    return len(set(tok for doc in tokens_list for tok in doc))

def ttr(tokens_list):
    total = sum(len(doc) for doc in tokens_list)
    return vocab_size(tokens_list) / max(total,1)

raw_V, norm_V = vocab_size(toks_raw), vocab_size(toks_norm_regex)
raw_TTR, norm_TTR = ttr(toks_raw), ttr(toks_norm_regex)

print({"raw_vocab": raw_V, "norm_vocab": norm_V, "raw_TTR": raw_TTR, "norm_TTR": norm_TTR})

# Top 25 kata sesudah normalisasi
from collections import Counter
freq_norm = Counter([t for d in toks_norm_regex for t in d])
TOP = pd.DataFrame(freq_norm.most_common(25), columns=["term","freq"])
print(TOP)

plt.figure(figsize=(10,5))
plt.bar(TOP["term"], TOP["freq"])
plt.xticks(rotation=60, ha='right')
plt.title('Top 25 Istilah setelah Normalisasi')
plt.tight_layout(); plt.show()
```

E. Tabel Transformasi Contoh (Sebelum → Sesudah)

```
sample_ix = [0,1,3,4,10,14,18,22]
view = pd.DataFrame({
    "original": [corpus[i] for i in sample_ix],
    "normalized": [norm[i] for i in sample_ix]
})
print(view)
```

F. Ekspor untuk Sessi 3–5

```
pd.DataFrame({"text": norm}).to_csv("corpus_sessi2_normalized.csv", index=False)
print("Tersimpan: corpus_sessi2_normalized.csv")
```

G. Opsional — Kebijakan Angka & Emoji

Jika tugas Anda butuh angka real (mis. kuantifikasi), matikan mask_num=True.
Jika sentimen bergantung pada emoji, pertahankan EMOJI sebagai token, bukan placeholder.

4) Studi Kasus & Diskusi

- Kasus e-commerce**: normalisasi menurunkan variasi ejaan ("rapih/rapi/rapid?") sehingga TF-IDF lebih stabil.
- Kasus log sistem**: URL, timestamp, dan path `s3:///` lebih baik dimask agar tidak mendominasi kosakata.
- Sentimen**: keputusan mempertahankan emoji bisa meningkatkan sinyal emosi positif/negatif.

5) Tugas Mini (Dinilai)

- Implementasikan dua kebijakan berbeda: (A) mask angka & emoji; (B) pertahankan angka & emoji. Bandingkan |V|, TTR, dan 25 term teratas.
- Buat *README* berisi daftar aturan pra-proses yang Anda terapkan (dan alasan pemilihan aturan).
- Ekspor berkas `corpus_sessi2_normalized.csv` untuk dipakai di Sessi 3–5.

← Sessi 1

Sessi 3 →

Sessi 1	Sessi 2	Sessi 3	Sessi 4	Sessi 5	Sessi 6	Sessi 7	Sessi 8
Sessi 9	Sessi 10	Sessi 11	Sessi 12	Sessi 13	Sessi 14	Sessi 15	Sessi 16

Sessi 3 — Stopword, Stemming, dan Lemmatization

Tujuan: memahami trade-off antara penghapusan stopwords, stemming (akar kata morfologis), dan lemmatization (bentuk kamus), serta mengukur dampaknya pada representasi ($[V]$, sparsity) dan performa tugas awal (kemiripan & sentimen berlabel lemah).

Learning Outcomes: (1) Menjelaskan perbedaan stopwords, stemming, lemmatization; (2) Menerapkan Sastrawi (ID) & spaCy (EN); (3) Mengevaluasi dampak ke TF-IDF & cosine similarity; (4) Menyiapkan korpus siap klasifikasi untuk sesi 4–7.

1) Definisi & Intuisi

- Stopword:** kata sangat umum ("dan", "yang", "the") yang sering tidak informatif untuk topik; bisa berguna untuk sentimen ("tidak").
- Stemming:** memotong afiks untuk mendapatkan *stem* ("berlari" → "lari"); cepat namun bisa agresif.
- Lemmatization:** memetakan ke *lemma* kamus berdasarkan analisis morfologis (EN: spaCy). Lebih mahal, cenderung akurat.

Prinsip: pilih teknik sesuai tugas. Untuk topik stopwords removal + stemming sering membantu. Untuk sentimen: hati-hati menghapus negasi ("tidak").

Checklist Kebijakan

- Stopword ID kustom + EN dari scikit-learn
- Aman untuk negasi ("tidak", "bukan") → *jangan* dihapus
- Bandingkan: Tanpa stemming vs Sastrawi stemming vs Lemmatization EN
- Ukur: $[V]$, TTR, sparsity TF-IDF, perubahan kemiripan

2) Praktik Google Colab — Pipeline Stopword/Stemming/Lemma

Gunakan korpus dari sesi 2 (`corpus_sessi2_normalized.csv`) atau langsung gunakan daftar di bawah (≥30 dokumen).

A. Setup

```
!pip -q install pandas numpy scikit-learn nltk Sastrawi spacy emoji Unidecode
# Model bahasa Inggris untuk lemmatization
!python -m spacy download en_core_web_sm > /dev/null

import re, math
import numpy as np, pandas as pd
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer
from numpy.linalg import norm
import nltk
nltk.download('punkt', quiet=True)

from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
import spacy
nlp_en = spacy.load('en_core_web_sm')

try:
    df = pd.read_csv('corpus_sessi2_normalized.csv')
    corpus = df['text'].dropna().astype(str).tolist()
    print('loaded corpus_sessi2_normalized.csv:', len(corpus), 'dokumen')
except Exception as e:
    print('Tidak menemukan CSV, gunakan korpus default.')
    corpus = [
        "pengiriman cepat kualitas barang baik sesuai ",
        "battery life is impressive however the screen glare is noticeable outdoors",
        "Layanan pelanggan responsif tetapi proses retur memakan waktu",
        "the movie had stunning visuals but a weak storyline",
        "harga terjangkau performa mantap untuk kebutuhan harian",
        "network latency spikes during peak hours affected our dashboard",
        "rasa kopi kuat aroma enak kemasan rapi",
        "app update fixed the crash but introduced login delays",
        "fitur kamera malam bagus namun stabilisasi video kurang",
        "documentation is thorough examples help onboarding",
        "pengalaman belanja menyenangkan voucher potongan harga membantu",
        "server downtime impacted api integrations",
        "tekstur kue lembut rasa tidak terlalu manis",
        "delivery was late packaging dented item worked fine",
        "antarmuka aplikasi intuitif navigasi mudah dipahami pemula",
        "customer support promised a callback but never followed up",
        "kualitas kain adem ukuran sesuai chart jahitan rapi",
        "dataset imbalance requires stratified sampling for evaluation",
        "suara speaker jernih bass cukup volume maksimal tidak pecah",
        "checkout process failed at payment gateway intermittently",
        "konektivitas bluetooth stabil jarak efektif sekitar meter",
        "the hotel staff were friendly but the room smelled of smoke",
        "keyboard travel nyaman backlight konsisten keycaps anti slip",
        "performa menurun setelah update cache clear membantu sementara",
        "shipping fee too high for small accessories",
        "resep mudah diikuti waktu memasak sesuai estimasi",
        "price to performance is excellent for students",
        "layar AMOLED tajam warna hidup tingkat kecerahan tinggi",
        "refund diproses cepat setelah komplain diajukan",
        "weather data ingestion requires timezone normalization"
    ]

print('Total dokumen:', len(corpus))
```

B. Stopword (ID & EN)

```
# Stopword EN bawaan scikit-learn + stopwords ID dari Sastrawi
from sklearn.feature_extraction import text as sktext
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

stop_en = sktext.ENGLISH_STOP_WORDS
stop_id_default = set(StopWordRemoverFactory().get_stop_words())

# Amankan negasi penting
SAFE_NEG = {"tidak","bukan","no","not","never"}
stop_id = {w for w in stop_id_default if w not in SAFE_NEG}

print(list(sorted(list(stop_id))[:15]), '... (ID)')
print(list(sorted(list(stop_en))[:15]), '... (EN)')
```

C. Stemming (Sastrawi) & Lemma (spaCy EN)

```
factory = StemmerFactory()
stemmer = factory.create_stemmer()

def stem_id(text):
    return ' '.join(stemmer.stem(t) for t in text.split())

def lemma_en(text):
    doc = nlp_en(text)
    return ' '.join(tok.lemma_ for tok in doc)

# Pipeline kebijakan
# v0: tanpa stopwords, tanpa stemming, tanpa lemma
# v1: stopwords EN+ID
# v2: v1 + stemming ID (Sastrawi)
# v3: v1 + lemma EN (spaCy)

def pipe_variant(texts):
    v0 = [t for t in texts]
    v1 = []
    v2 = []
    v3 = []
    for t in texts:
        toks = t.split()
        toks_v1 = [w for w in toks if w not in stop_id and w not in stop_en]
        v1t = ' '.join(toks_v1)
        v1.append(v1t)
        # Stemming untuk token ID (heuristik: token all-lower latin & tanpa anglisise khusus)
        v2.append(stem_id(v1t))
        # Lemma EN (jalankan di versi v1 agar stopwords EN sudah dihapus)
        v3.append(lemma_en(v1t))
    return v0, v1, v2, v3

v0, v1, v2, v3 = pipe_variant(corpus)

pd.DataFrame({
    'original': corpus[:6],
    'v0_raw': v0[:6],
    'v1_stop': v1[:6],
    'v2_stop_stemID': v2[:6],
    'v3_stop_lemmaEN': v3[:6]
})
```

D. TF-IDF & Sparsity

```
def tfidf_stats(texts, ngram=(1,1)):
    vec = TfidfVectorizer(ngram_range=ngram, min_df=1)
    X = vec.fit_transform(texts)
    vocab = vec.get_feature_names_out()
    density = X.nnz/(X.shape[0]*X.shape[1])
    return X, vocab, density

for name, variant in [('v0_raw', v0), ('v1_stop', v1), ('v2_stop_stemID', v2), ('v3_stop_lemmaEN', v3)]:
    X, vocab, dens = tfidf_stats(variant)
    print(name, 'shape=', X.shape, 'vocab=', len(vocab), 'sparsity=', 1-dens)
```

E. Dampak pada Kemiripan (Cosine)

```
from numpy.linalg import norm
import numpy as np

def cosine(a,b):
    return float(a.dot(b.T).toarray()[0,0] / (norm(a.toarray())*norm(b.toarray()) + 1e-12))

# Pilih dokumen contoh & query
query = "pengiriman cepat kualitas bagus"

vec = TfidfVectorizer()
for name, variant in [('v0_raw', v0), ('v1_stop', v1), ('v2_stop_stemID', v2), ('v3_stop_lemmaEN', v3)]:
    X = vec.fit_transform(variant)
    q = vec.transform([query])
    sims = (X @ q.T).toarray().ravel() / (np.linalg.norm(X.toarray(), axis=1)*np.linalg.norm(q.toarray()) + 1e-12)
    top = np.argsort(sims)[::-1][:5]
    print('\n==', name, '==')
    for r in top:
        print(f'sim={sims[r]:.3f}', variant[r][:90])
```

F. Label Lemah (Distant Supervision) untuk Sentimen

```
# Label lemah: berdasarkan leksikon kecil (ID+EN)
POS = ["bagus","mantap","menyenangkan","cepat","baik","excellent","impressive","friendly","clean","tajam"]
NEG = ["buruk","lambat","telat","downtime","lemah","weak","late","dented","smelled","failed"]

def weak_label(text):
    toks = set(text.split())
    pos = len(toks & POS)
    neg = len(toks & NEG)
    if pos>neg: return 1
    if neg>pos: return 0
    return None # tidak jelas

labels = [weak_label(t) for t in v1] # gunakan v1 (stopword removed)
X, vocab, _ = tfidf_stats(v1)

# Pilih hanya dokumen yang terlabel
idx = [i for i,l in enumerate(labels) if l is not None]
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score

if len(idx) >= 10:
    y = np.array([labels[i] for i in idx])
    Xsub = X[idx]
    clf = LogisticRegression(max_iter=200)
    scores = cross_val_score(clf, Xsub, y, cv=5, scoring='f1')
    print('F1 (weak labels, v1 stopwords):', scores.mean().round(3), '+/-' , scores.std().round(3))
else:
    print('Sedikit dokumen yang terlabel lemah; tambahkan korpus atau leksikon.')
```

G. Ekspor Korpus Terproses

```
pd.DataFrame({
    'v0_raw': v0,
    'v1_stop': v1,
    'v2_stop_stemID': v2,
    'v3_stop_lemmaEN': v3
}).to_csv('corpus_sessi3_variants.csv', index=False)
print('Tersimpan: corpus_sessi3_variants.csv')
```

3) Studi Kasus & Analisis

Kasus	Harapan	Catatan
Klasifikasi topik berita	Stopword removal + stemming menurunkan $[V]$ dan menaikkan stabilitas fitur	Hati-hati nama entitas; jangan dihapus
Sentimen ulasan	Jangan hapus negasi ("tidak", "not"); stemming aman	Emoji dapat dipertahankan sebagai fitur
Similarity pencarian	Stemming membantu <i>recall</i> query berimbuhan	Lemma EN menjaga keakuratan bentuk kata

4) Tugas Mini (Dinilai)

- Bandingkan empat varian (v0–v3) pada metrik: $[V]$, sparsity, akurasi *retrieval*@5 untuk 5 query yang Anda desain sendiri.
- Uji label lemah pada varian v1 dan v2; laporkan F1 (CV=5).
- Tulis ringkasan kapan memilih stopwords saja vs +stemming vs +lemma.



YAYASAN PERGURUAN CIKINI
INSTITUT SAINS DAN TEKNOLOGI NASIONAL
Jl. Moh. Kahfi II, Bhumi Srengseng Indah, Jagakarsa, Jakarta Selatan 12640
Telp. 021-7270090 (hunting), Fax 021-7866955, hp: 081291030024
Email: humas@istn.ac.id Website: www.istn.ac.id

SURAT PENUGASAN TENAGA PENDIDIK
Nomor : 039-XI/03.1-F/IX/2025
SEMESTER GANJIL TAHUN AKADEMIK 2025/2026

Nama	: Riadi Marta Dinata, S.TI., M.Kom.	Status Pegawai	: Edukatif Tetap
NIK/ NIDN/ NIDK	: 202103-002/-/0320087704	Program Studi	: Teknik Informatika
Jabatan Akademik	: Tenaga Pendidik		

Bidang	Perincian Kegiatan	Tempat	Jam	Kredit (SKS)	Hari
I. PENDIDIKAN & PENGAJARAN	1. Pengajaran di kelas termasuk laboratorium				
	1. Pemograman (Python, Matlab) (IN Kelas A)	R-A3	08:00 - 11:00	2	Senin
	2. Sistem Digital dan Arsitektur Komputer (IN Kelas B)	Lab Komp	07:00 - 11:00	2	Sabtu
	3. Artificial Intelligence 3 (Math Kelas A)	R-C1	22:00 - 23:30	1,5	Selasa
	4. Big Data dan Data Sains (Fi Kelas A)	R-A6	15:00 - 18:00	4	Rabu
	5. Strategic Foresight & Transformasi Digital Berbasis AI (Si Kelas A)	R-A4	08:00 - 10:15	1,5	Kamis
	6. Kecerdasan Buatan (FI Kelas A)	R-A5	21:00 - 23:20	3	Kamis
	7. Pengantar Data Sains (IN Kelas A)	R-A2	09:00 - 10:30	1	Jumat
	8. Sistem Otonom dan Kontrol Edge (IF Kelas B)	Lab Komp	18:00 - 20:10	3	Jumat
	2. Pembimbing				
	1. Seminar				
	2. Kerja Praktek				
	3. Tugas Akhir/Tesis				
	4. Pembimbing Akademik			1	
	3. Penguji				
	1. Tugas Akhir/Tesis				
	2. Kerja Praktek				
	4. Tugas Tambahan				
	1. Menjabat Sebagai Ka. Lab Komputer Tif dan Si			3	
II. PENELITIAN	1. Penelitian Ilmiah			1	
	2. Penulisan Karya Ilmiah				
	3. Penulisan Diktat Kuliah				
	4. Menerjemahkan Buku Kuliah				
	5. Pengembangan Program Kuliah Kurikulum				
	6. Pengembangan Bahan Ajar				
III. PENGABDIAN PADA MASYARAKAT	1. Menduduki jabatan di Pemerintahan				
	2. Pengembangan Hasil Pendidikan dan Penelitian				
	3. Memberikan penyuluhan/pelatihan/penataran/ceramah			1	
	4. Memberikan Pelayanan Kepada Masyarakat				
	5. Menulis karya Pengmas yang tidak dipublikasikan			1	
	6. Pengelolaan Jurnal Ilmiah				
IV. PENUNJANG	1. Menjadi anggota/panitia pada badan/lembaga suatu PT				
	2. Menjadi anggota Badan Lembaga Pemerintah				
	3. Menjadi anggota organisasi profesi				
	4. Mewakili PT/lembaga pemerintah, duduk dalam panitia antar lembaga				
	5. Menjadi anggota delegasi nasional ke pertemuan internasional				
	6. Berperan Serta Aktif dalam pertemuan ilmiah/seminar				
	7. Anggota dalam tim layanan pendidikan				
Jumlah Total				25	

Kepada yang bersangkutan akan diberikan gaji/honorarium sesuai dengan peraturan penggajian yang berlaku di Institut Sains dan Teknologi Nasional. Penugasan ini berlaku dari tanggal 01 September 2025 sampai dengan 28 Februari 2026

Tembusan :

1. Wakil Rektor 1 - ISTN
2. Wakil Rektor 2 - ISTN
3. Ka. Biro Sumber Daya Manusia - ISTN
4. Arsip



Jakarta, 01 September 2025
Dekan FSTT

Dr. Ir. Kun Wardana Abyoto, M.T.
NIDN. 0311086903