

Akuisisi Data

Deskripsi Modul

| | |
|--------------------------------|---------------------------------|
| Mata Kuliah | Machine Learning |
| Kode Mata Kuliah / SKS | RPL / 3 |
| Semester | 5 |
| Kelas | TRPL |
| Capaian Pembelajaran | |
| Deskripsi Singkat Mata Kuliah | |
| Bahan Kajian Modul | Akuisisi dan Pengolahan Dataset |
| Bentuk dan Metode Pembelajaran | |
| Waktu pembelajaran | 120 Menit |
| Rekomendasi buku teks: | 3. 4. |
| Petunjuk khusus | - |

1. Materi Pembelajaran

1.1 Pendahuluan

Definisi Akuisisi Data

Akuisisi Data atau Data Collecting merupakan proses pengumpulan informasi dari sumber-sumber yang relevan untuk menemukan solusi atas pertanyaan statistik yang diberikan. Pengumpulan Data adalah langkah pertama dan terpenting dalam penyelidikan statistik. Ini adalah langkah penting karena membantu kita membuat keputusan yang tepat, melihat tren, dan mengukur kemajuan.

Berbagai metode pengumpulan data meliputi:

- Wawancara
- Kuesioner
- Pengamatan
- Percobaan
- Sumber yang Diterbitkan dan Sumber yang Tidak Diterbitkan

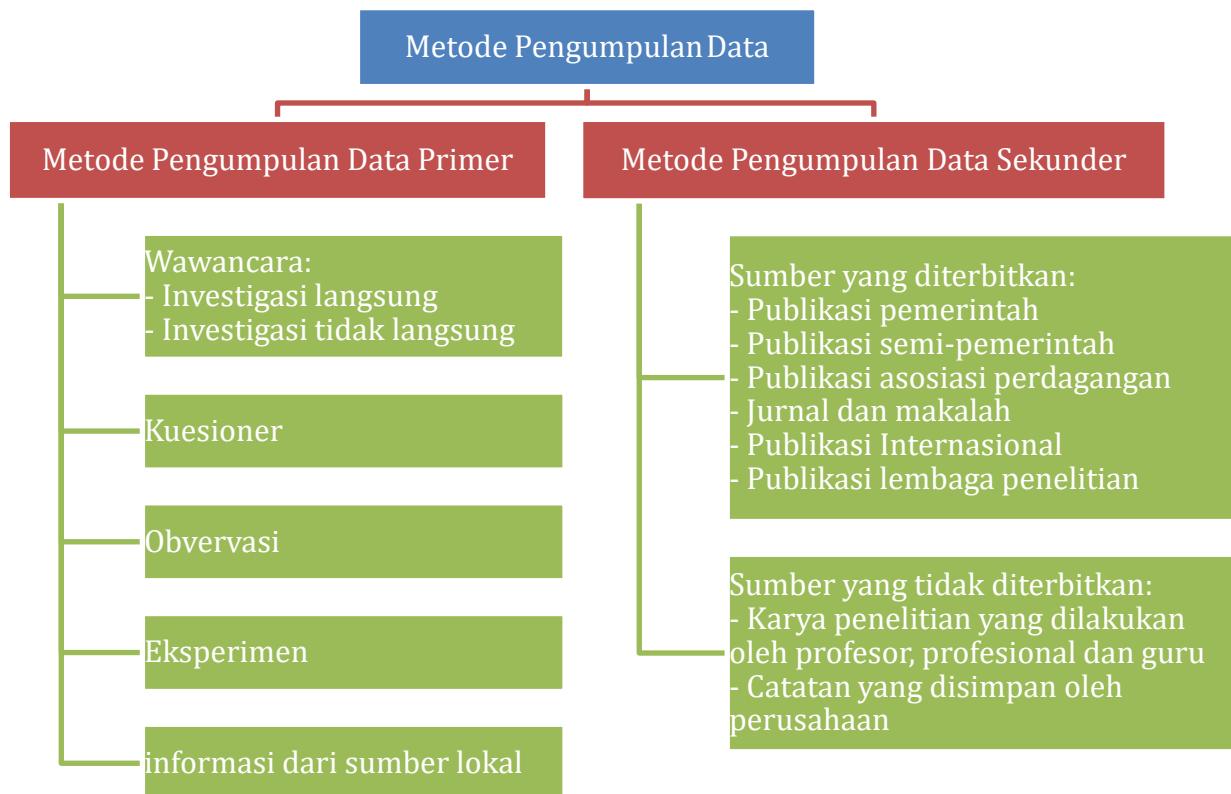
Peran akuisisi data dalam pipeline Machine Learning

Akuisisi data merupakan landasan pembelajaran mesin. Dalam pembelajaran mesin, "akuisisi data" mengacu pada prosedur perolehan dan penyusunan data dari berbagai sumber untuk menguji dan melatih model pembelajaran mesin.

Sumber-sumber data (primer & sekunder)

Data Primer → mengacu pada informasi yang dikumpulkan langsung dari sumber langsung. Data primer tersedia dalam bentuk mentah. Jenis data ini dikumpulkan melalui berbagai metode, termasuk survei, wawancara, eksperimen, observasi, dan kelompok fokus. Salah satu keuntungan utama data primer adalah menyediakan informasi terkini, relevan, dan spesifik yang disesuaikan dengan kebutuhan peneliti, sehingga menawarkan tingkat akurasi dan kontrol yang tinggi terhadap kualitas data.

Data sekunder → mengacu pada informasi yang telah dikumpulkan, diproses, dan dipublikasikan oleh orang lain . Jenis data ini dapat bersumber dari makalah penelitian yang ada, laporan pemerintah, buku, basis data statistik, dan catatan perusahaan. Keuntungan dari data sekunder adalah data tersebut tersedia secara luas dan seringkali gratis atau lebih murah untuk diperoleh dibandingkan dengan data primer. Data sekunder menghemat waktu dan sumber daya karena fase pengumpulan data telah selesai.



1.2 Open-Source Dataset

Open-source dataset adalah kumpulan data yang dapat diakses secara bebas oleh publik dengan batasan penggunaan, modifikasi, dan distribusi yang minimal. Pemerintah, organisasi, atau peneliti perorangan menyumbangkannya sebagai ganti data berbayar. Hal ini memungkinkan akses yang luas untuk mendorong analisis, penelitian, dan inovasi. Data sumber terbuka mendorong transparansi dan kolaborasi. Pengguna dapat melihat, memodifikasi, dan berbagi data tanpa batasan.

Data sumber terbuka memiliki beberapa keuntungan bagi perusahaan dan individu yang mencari wawasan, inovasi, dan pengambilan keputusan yang berharga, diantaranya:

- Aksesibilitas: Tersedia secara gratis untuk integrasi mudah dengan alat analitik dan AI/ML.
- Efektivitas biaya: Mengurangi biaya dengan menghilangkan biaya lisensi.
- Transparansi: Dilengkapi dengan metadata terperinci, memastikan kualitas dan kepatuhan data.
- Fleksibilitas: Dapat disesuaikan untuk alur kerja tertentu dan dapat diadaptasi untuk otomatisasi data.
- Inovasi Berbasis Komunitas: Diperbarui dan ditingkatkan secara berkala oleh kontributor global.

Beberapa contoh open-source dataset:

| | |
|--|--|
| | |
| Kaggle | Pusat komunitas terkenal bagi para ilmuwan data yang menawarkan berbagai macam kumpulan data dalam berbagai topik. Sangat berguna untuk proyek dan kompetisi pembelajaran mesin. https://www.kaggle.com/docs/datasets |
| Data.gov | Menawarkan koleksi besar kumpulan data pemerintah federal AS tentang berbagai topik seperti pertanian, iklim, pemerintahan daerah, dan banyak lagi. https://data.gov/ |
| Google Dataset Search | Memungkinkan Anda mencari kumpulan data yang tersimpan di seluruh web. Ini adalah alat yang ampuh untuk menemukan kumpulan data untuk berbagai keperluan. https://datasetsearch.research.google.com/ |
| FiveThirtyEight | Dikenal karena jurnalisme datanya yang menarik, FiveThirtyEight menerbitkan kumpulan data yang digunakan dalam artikel mereka, ideal untuk analisis dan penceritaan visual. https://data.fivethirtyeight.com/ |
| UCI Machine Learning Repository | Sumber daya populer yang menawarkan kumpulan data yang secara khusus disiapkan untuk pembelajaran mesin. https://archive.ics.uci.edu/ml/index.php |
| AWS Public Datasets | Amazon Web Services menyediakan berbagai kumpulan data yang dapat diintegrasikan langsung ke dalam aplikasi berbasis cloud AWS. https://aws.amazon.com/marketplace/solutions/data-analytics/data-sets |
| GitHub | Menyimpan berbagai kumpulan data real-time dan lainnya yang tersedia untuk penggunaan publik, menjadikannya sumber daya yang berharga untuk proyek ilmu data. https://github.com/ |
| NOAA (National Oceanic and Atmospheric Administration) | Menyediakan arsip lengkap kumpulan data lingkungan, cuaca, dan oseanografi. https://www.noaa.gov/ |

| | |
|-------------------|--|
| Harvard Dataverse | Sebuah repositori untuk berbagai kumpulan data dari para peneliti di seluruh dunia, yang mencakup berbagai bidang studi. Jelajahi Harvard https://dataverse.harvard.edu/ |
| Eurostat | Menawarkan serangkaian kumpulan data statistik yang mencakup Uni Eropa dan negara-negara anggotanya, ideal untuk penelitian ekonomi dan sosial. https://ec.europa.eu/eurostat |

1.3 Web Scraping Dasar

Pengikisan web adalah teknik otomatis yang digunakan untuk mengekstrak data dari situs web. Alih-alih menyalin dan menempel informasi secara manual yang merupakan proses yang lambat dan berulang, teknik ini menggunakan perangkat lunak untuk mengumpulkan sejumlah besar data dengan cepat. Perangkat ini dapat dibuat khusus atau digunakan di beberapa situs. Teknik ini juga membantu individu dan bisnis untuk mengumpulkan data berharga untuk penelitian, pemasaran, dan analisis.

Penggunaan Web Scraping

| | |
|------------------------------|--|
| | |
| Analisis Pasar dan Pesaing | Bisnis mengumpulkan harga produk, ulasan pelanggan, dan penawaran pesaing dari berbagai situs web. Hal ini membantu mereka tetap mengetahui tren pasar dan menyesuaikan strategi mereka agar tetap kompetitif. |
| Pengumpulan Data Keuangan | Investor dan analis mengekstrak harga saham, data historis, dan laporan keuangan secara real-time. Informasi ini mendukung pengambilan keputusan yang lebih baik dan respons yang tepat waktu terhadap perubahan pasar. |
| Pemantauan Media Sosial | Pemasar mengumpulkan data dari platform media sosial untuk melacak topik yang sedang tren, sentimen pelanggan, dan efektivitas kampanye. Ini membantu dalam membentuk strategi pemasaran dan meningkatkan keterlibatan pelanggan. |
| Pelacakan SEO | Perusahaan menggunakan alat scraping untuk memantau peringkat situs web mereka di mesin pencari untuk kata kunci tertentu dari waktu ke waktu. Ini membantu mengoptimalkan konten dan meningkatkan visibilitas online. |
| Riset dan Pembelajaran Mesin | Peneliti dan ilmuwan data mengumpulkan kumpulan data besar dari berbagai situs web untuk melatih model pembelajaran mesin atau melakukan studi berbasis data. Scraping mengotomatiskan pengumpulan data ini dan membantu menghemat waktu dan tenaga. |

Teknik Scraping Web

1. Scraping Manual

Proses ini melibatkan penyalinan (copy) dan penempelan (paste) data secara manual. Proses ini sederhana tetapi lambat, tidak efisien dan tidak praktis untuk data berskala besar atau yang sering diperbarui.

2. Scraping Otomatis

Scraping otomatis menggunakan skrip atau perangkat lunak untuk mengambil dan memproses data dalam skala besar. Scraping otomatis lebih cepat, lebih andal, dan cocok untuk konten dinamis. Metode otomatis yang umum meliputi:

- Parsing HTML: Mengekstrak data dari HTML mentah halaman web statis.
- Parsing DOM: Berinteraksi dengan Model Objek Dokumen (DOM) untuk mengekstrak konten yang dimuat secara dinamis.
- Akses API: Bila tersedia, API menyediakan data terstruktur dan andal secara langsung—sering kali menjadi metode yang lebih disukai dibanding pengikisan.
- Browser Tanpa Kepala seperti Selenium: Ini mensimulasikan interaksi pengguna dalam browser, yang memungkinkan ekstraksi data dari situs web yang banyak menggunakan JavaScript atau interaktif.

Tools Populer untuk Web Scraping

| BeautifulSoup | pustaka Python yang mudah digunakan bagi pemula yang digunakan untuk mengurai dokumen HTML dan XML. Pustaka ini memungkinkan kita untuk menelusuri struktur halaman dan mengekstrak elemen tertentu menggunakan tag dan kelas. |
|----------------------|---|
| Requests (Python) | Permintaan digunakan dengan BeautifulSoup untuk membantu mengirim permintaan HTTP ke situs web dan mengambil konten HTML dari halaman web. |
| Scrapy | Framework Python canggih yang dibuat untuk web scraping. Framework ini mendukung fitur-fitur seperti crawling, penanganan permintaan/respons, pengelolaan alur kerja, dan penyimpanan data hasil scraping secara efisien. |
| Selenium | alat otomatisasi web yang dapat mengendalikan peramban seperti pengguna sungguhan. Alat ini berguna untuk mengikis situs web yang menggunakan JavaScript untuk memuat konten seperti menu gulir tak terbatas atau menu tarik-turun. |

| | |
|--------------------|--|
| Playwright | alternatif baru untuk Selenium, mendukung standar web modern dan menyediakan kinerja yang lebih baik untuk mengikis konten dinamis dengan kontrol browser tanpa kepala. |
| Platform Komersial | <ul style="list-style-type: none"> - Bright Data (sebelumnya Luminati) : Platform berbasis proxy premium dengan fitur pengikisan yang kuat. - Import.io : Memungkinkan pengikisan tanpa pengkodean yang ideal bagi non-programmer. - Webhose.io : Menawarkan umpan data terstruktur untuk berita, blog, dan konten daring. - Dexi.io dan Scrapinghub: Menyediakan layanan scraping berbasis cloud dengan penjadwalan bawaan, penyimpanan, dan dukungan proksi. |

2.4 Akuisisi Data dari Sensor / IoT

Tujuan: Mengenal metode pengambilan data dari sensor secara real-time (jika tersedia).

Materi:

Data dari Arduino/ESP32

Penggunaan MQTT atau Serial

2.5 Data Cleaning dan Pra-Pemrosesan Sederhana

Tujuan: Menyiapkan data hasil akuisisi agar siap dipakai untuk training model.

Materi:

Data cleaning juga dikenal sebagai data cleansing atau data scrubbing. Data cleaning merupakan proses mengidentifikasi, mengoreksi atau menghapus kesalahan, inkonsistensi dan ketidakakuratan dalam dataset, sehingga memastikan data berkualitas untuk tahapan analisa. Data cleaning adalah kegiatan penting dalam pre-pemrosesan data karena menentukan bagaimana data akan digunakan dan diproses dalam tahapan pemodelan.

Tugas-tugas yang paling umum dilakukan saat data cleaning:

1. Menghapus missing values

Missing value merupakan masalah umum yang terdapat dalam dataset. Solusi yang bisa digunakan untuk menangani masalah ini adalah:

- Menghapus record → menghapus baris dengan nilai yang hilang jika jumlahnya relatif sedikit dan tidak signifikan
- Mengganti nilai → mengganti nilai yang hilang dengan nilai estimasi, seperti nilai rata-rata, median atau modus dari kumpulan data

- Menggunakan algoritma → menggunakan teknik seperti regresi atau model machine learning untuk memprediksi dan mengisi nilai yang hilang
2. Menghapus duplikasi

Duplikasi dapat mendistorsi analisis dan menghasilkan hasil yang tidak akurat. Mengidentifikasi dan menghapus data duplikat akan memastikan bahwa setiap titik data unik dan terwakili secara akurat.
 3. Memperbaiki ketidakakuratan

Kesalahan entri data, seperti kesalahan ketik atau nilai yang salah, perlu diidentifikasi dan diperbaiki. Kesalahan ini mencakup referensi silang dengan sumber data lain atau penggunaan aturan validasi untuk memastikan keakuratan data.
 4. Standarisasi format

Data dapat dimasukkan dalam bentuk berbagai format, sehingga sulit dianalisis. Standarisasi format, seperti tanggal, alamat dan nomor telepon, memastikan konsistensi dan memudahkan pengolahan data.
 5. Menangani penyimpangan

Penyimpangan dapat mendistorsi analisis dan menghasilkan hasil yang salah. Mengidentifikasi dan menangani kesalahan, baik dengan menghapusnya maupun mentransformasi data, membantu menjaga integritas dataset.

Langkah-langkah data cleaning:

1. Menilai kualitas data

Langkah pertama dalam data cleaning adalah untuk mengakses kualitas data. Kegiatan ini meliputi pengecekan:

- Missing value (nilai yang hilang)

Mengidentifikasi nilai yang kosong (blank value) atau nilai null (null/nan value) dalam dataste. Missing value bisa dikarenakan berbagai alasan seperti pengumpulan data yang tidak lengkap (scrapping data berhenti di tengah proses scrapping), kesalahan entri data atau kehilangan data selama transmisi

- Incorrect value

Memeriksa nilai yang berada di luar rentang yang diharapkan atau tidak konsisten dengan tipe data. Contohnya, kolom tanggal dengan tanggal yang tidak valid atau kolom numerik dengan karakter non-numerik

- Ketidakkonsistenan format data

Memastikan format data konsisten di seluruh dataset. Contohnya, memastikan tanggal menggunakan format yang sama (contoh TTTT-BB-HH) dan variabel kategorikal memiliki label yang konsisten.

Contoh kasus:

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|------------|
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| Jane | null | 70 | 2022-01-03 |
| Nan | 35 | 95 | 2022-01-04 |
| Alice | | 85 | 2022-01-05 |
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| | 40 | 100 | 2022-01-06 |

Kesalahan pada DataFrame di atas, yaitu:

- Duplikasi baris: baris 6 dan 7 adalah duplikasi, yang menunjukkan potensi masalah duplikasi
- Missing value: baris 3 memiliki nilai yang hilang di kolom “Usia”, baris 4 memiliki nilai yang hilang di kolom “Nama”, baris 5 memiliki nilai yang hilang di kolom “Usia” dan baris 8 memiliki nilai yang hilang di kolom “Nama”
- Format tanggal yang tidak konsisten: kolom “Tanggal” berisi tanggal dalam format “YYYY-MM-DD” yang konsisten, tetapi penting untuk memastikan konsistensi di semua entri tanggal
- Kemungkinan penyimpangan: skor 100 pada baris 8 dapat dianggap sebagai penyimpangan, tergantung pada konteks data dan sistem penilaian yang digunakan

2. Hapus data yang tidak relevan

Duplikasi data dapat mendistorsi hasil analisis dan menyebabkan kesimpulan yang salah.

Deduplikasi melibatkan:

- Identifikasi entri duplikasi: menggunakan teknik seperti pengurutan, pengelompokan atau pembuatan hash untuk mengidentifikasi data yang duplikasi
- Menghapus catatan duplikasi: setelah duplikasi teridentifikasi, hapus dari dataset untuk memastikan bahwa setiap titik data unik dan terwakili secara akurat
- Identifikasi redundansi (pengulangan): cari catatan duplikasi atau identik yang tidak menambahkan informasi baru
- Menghilangkan informasi yang tidak relevan: hapus variabel atau kolom apa pun yang tidak relevan dengan analisis atau tidak memberikan wawasan yang berguna

Contoh kasus:

DataFrame tidak sempurna

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|------------|
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| Jane | null | 70 | 2022-01-03 |
| Nan | 35 | 95 | 2022-01-04 |
| Alice | | 85 | 2022-01-05 |
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| | 40 | 100 | 2022-01-06 |

DataFrame deduplikasi

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|------------|
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| Jane | null | 70 | 2022-01-03 |
| Nan | 35 | 95 | 2022-01-04 |
| Alice | | 85 | 2022-01-05 |
| | 40 | 100 | 2022-01-06 |

Baris 6 dan 7 yang memiliki nilai duplikasi telah dihapus pada DataFrame deduplikasi

3. Memperbaiki kesalahan struktural

Kesalahan struktural mencakup inkonsistensi dalam format data, konvensi penamaan atau tipe variabel. Standarisasi format, koreksi perbedaan penamaan dan keseragaman representasi data sangat penting untuk analisis yang akurat. Kegiatan ini meliputi:

- Standarisasi format data: memastikan bahwa tanggal, waktu dan tipe data lainnya diformat secara konsisten di seluruh dataset
- Memperbaiki perbedaan penamaan: memeriksa ketidakkonsistenan dalam nama kolom, variabel atau label dan menstandarisasikan
- Memastikan keseragaman dalam representasi data: verifikasi bahwa data direpresentasikan secara konsisten, seperti menggunakan unit yang sama untuk pengukuran atau skala yang sama untuk penilaian

DataFrame deduplikasi

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|------------|
| John | 25 | 90 | 2022-01-01 |
| Mary | 31 | 80 | 2022-01-02 |
| Jane | null | 70 | 2022-01-03 |
| Nan | 35 | 95 | 2022-01-04 |
| Alice | | 85 | 2022-01-05 |
| | 40 | 100 | 2022-01-06 |

DataFrame dengan format tanggal yang distandarisasi

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|---------------------|
| John | 25 | 90 | 2022-01-01 00:00:00 |
| Mary | 31 | 80 | 2022-01-02 00:00:00 |
| Jane | null | 70 | 2022-01-03 00:00:00 |
| Nan | 35 | 95 | 2022-01-04 00:00:00 |
| Alice | | 85 | 2022-01-05 00:00:00 |
| | 40 | 100 | 2022-01-06 00:00:00 |

Kolom “Tanggal” telah distandarisasi ke format “T1T2T-BB-HH” di semua entri. Hal ini memastikan konsistensi dalam format tanggal.

4. Menangani missing value

Beberapa strategi untuk menangani missing value:

- Mengimputasi missing value: menggunakan metode statistik seperti rata-rata, median atau modus untuk mengisi yang hilang
- Menghapus catatan missing value: jika nilai yang hilang sangat banyak atau tidak dapat diperhitungkan secara akurat, hapus catatan dengan nilai yang hilang
- Memanfaatkan teknik imputasi lanjutan: menggunakan teknik seperti imputasi regresi, K-NN atau desicion tree untuk mengimputasikan nilai yang hilang

Contoh kasus:

DataFrame dengan format tanggal yang distandarisasi

| Nama | Usia | Nilai | Tanggal |
|-------|------|-------|---------------------|
| John | 25 | 90 | 2022-01-01 00:00:00 |
| Mary | 31 | 80 | 2022-01-02 00:00:00 |
| Jane | null | 70 | 2022-01-03 00:00:00 |
| Nan | 35 | 95 | 2022-01-04 00:00:00 |
| Alice | | 85 | 2022-01-05 00:00:00 |
| | 40 | 100 | 2022-01-06 00:00:00 |

DataFrame dengan penanganan missing value

| Nama | Usia | Nilai | Tanggal |
|---------|------|-------|---------------------|
| John | 25 | 90 | 2022-01-01 00:00:00 |
| Mary | 31 | 80 | 2022-01-02 00:00:00 |
| Unknown | 35 | 95 | 2022-01-04 00:00:00 |
| Unknown | 40 | 100 | 2022-01-06 00:00:00 |

Nilai yang hilang pada kolom “Usia” (baris 3 dan 5) dihapus, nilai yang hilang pada kolom “Nama” (baris 4 dan 6) diganti dengan “Unknown” untuk menandakan bahwa nama tidak diketahui atau tidak tersedia.

5. Normalisasi Data

Normalisasi data melibatkan pengorganisasian data untuk mengurangi redundansi dan meningkatkan efisiensi penyimpangan. Kegiatan ini meliputi:

- Membagi data ke dalam beberapa tabel: membagi data ke dalam tabel terpisah, masing-masing menyimpan jenis informasi tertentu
- Memastikan konsistensi data: verifikasi bahwa data terstruktur sehingga memudahkan query dan analisis yang efisien

Normalized data (Informasi siswa)

| Nama | Usia | Tanggal |
|---------|------|---------------------|
| John | 25 | 2022-01-01 00:00:00 |
| Mary | 31 | 2022-01-02 00:00:00 |
| Unknown | 35 | 2022-01-04 00:00:00 |
| Unknown | 40 | 2022-01-06 00:00:00 |

Normalized data (Nilai)

| Nama | Nilai |
|---------|-------|
| John | 90 |
| Mary | 80 |
| Unknown | 95 |
| Unknown | 100 |

6. Identifikasi dan kelola penyimpangan (Outlier)

Outlier adalah titik data yang menyimpang secara signifikan dari norma dan dapat mendistorsi hasil analisis. Tergantung pada konteksnya, dapat memilih kegiatan berikut:

- Hapus outlier: jika outlier disebabkan oleh kesalahan entri data atau tidak mewakili populasi, hapus outlier tersebut dari kumpulan data
- Transformasi outlier: jika outlier valid tetapi ekstrim, transformasikan untuk meminimalkan dampaknya pada analisis

Contoh kasus:

Normalized data (Nilai)

| Nama | Nilai |
|---------|-------|
| John | 90 |
| Mary | 80 |
| Unknown | 95 |
| Unknown | 100 |

DataFrame dengan outlier termanajemen

| Nama | Nilai |
|------|-------|
| John | 90 |
| Mary | 80 |

Alat dan Teknik untuk Data Cleaning

Alat Perangkat Lunak:

1. Microsoft excel: menawarkan fungsi permbersihan data dasar seperti menghapus duplikasi, menangani missing value dan menstandarisasi format
2. OpenRefine: alat sumber terbuka yang dirancang khusus untuk pembersihan dan transformasi data
3. Library Python: library seperti Pandas dan Numpy menyediakan fungsi hebat untuk pembersihan dan manipulasi data
4. R: bahasa pemrograman R menawarkan paket untuk data cleaning, seperti dplyr dan tidyR

Teknik:

1. Ekspresi reguler: berguna untuk pencocokan pola dan manipulasi teks
2. Profil data: melibatkan pemeriksaan data untuk memahami struktur, konten dan kualitasnya
3. Audit data: memeriksa data secara sistematis untuk menemukan kesalahan dan ketidakkonsistenan

2. Praktikum

2.1. Web Scraping Dasar

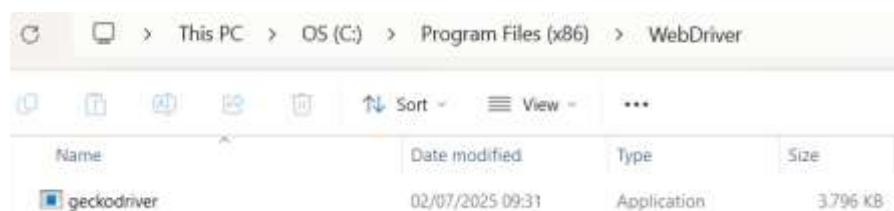
Scraping review film How To Train Your Dragon dari website IMDb.

Setup geckodriver

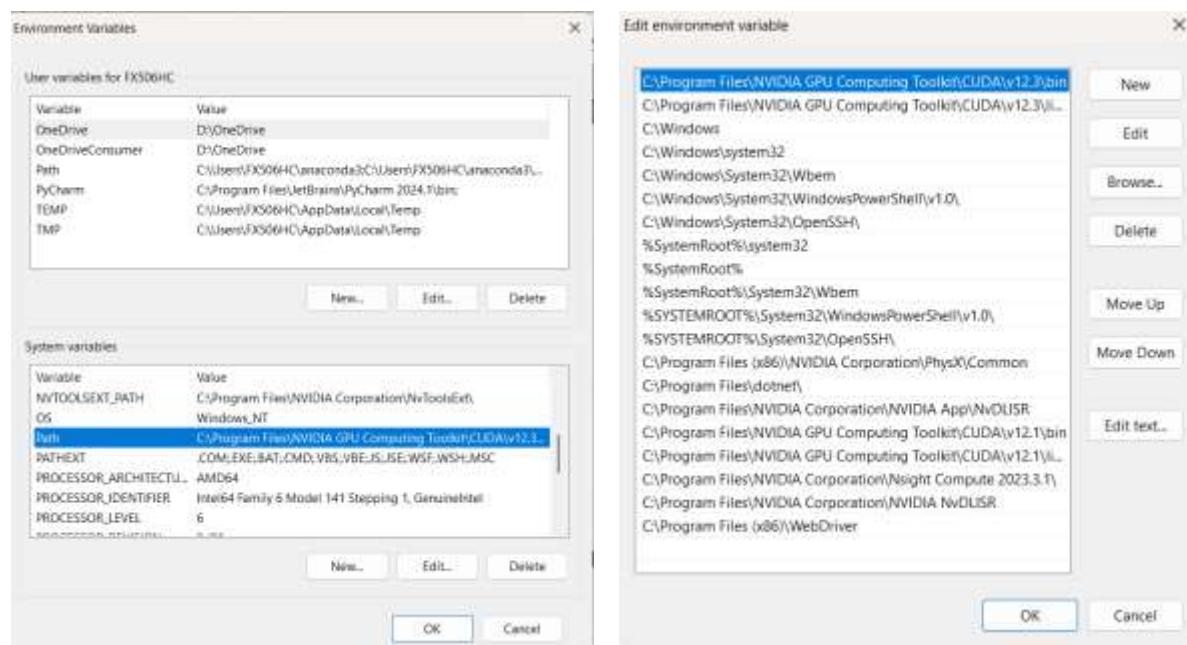
1. Download aplikasi geckodriver <https://github.com/mozilla/geckodriver/releases>



2. Extract geckodriver ke folder C driver



3. Tambahkan geckodriver ke path system komputer



Instalasi Library

1. Terdapat beberapa library yang dibutuhkan untuk melakukan scraping data dari website. Library pertama yang dibutuhkan adalah selenium.

```
Pip install selenium
```

2. Library kedua yang dibutuhkan adalah web-driver

```
Pip install web-driver
```

Pembuatan jupyter notebook

1. Kode pertama yang dijalankan adalah kode untuk **mengatur opsi browser Firefox** sebelum menjalankan Selenium WebDriver

```
from selenium.webdriver.firefox.options import Options  
#from selenium.webdriver.firefox.service import Service  
  
# SETUP FIREFOX  
options = Options()  
#options.headless = True # Jalankan tanpa tampilan GUI  
#service = Service() # geckodriver otomatis ditemukan jika sudah di PATH
```

2. Kemudian webDriver Firefox dijalankan

```
from selenium import webdriver  
  
# Jalankan Firefox driver  
driver = webdriver.Firefox(options=options)
```

3. Langkah ketiga adalah mengakses halaman film yang ingin di scrap dengan waktu tunggu `time.sleep(10)`. Artinya program membutuhkan waktu 10 detik untuk menampilkan halaman web secara keseluruhan

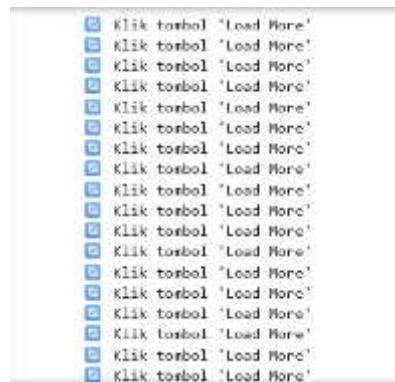
```
import time  
  
# Ganti URL dengan hotel yang ingin di-scrape  
url = "https://www.imdb.com/title/tt26743210/reviews"  
driver.get(url)  
time.sleep(10) # Tunggu render halaman
```

4. Langkah berikutnya adalah mencari tombol “load more” untuk mengakses keseluruhan review yang terdapat dalam halaman film

```
from selenium.common.exceptions import NoSuchElementException

# Scroll dan klik "Load More" jika ada
while True:
    try:
        svg = driver.find_element(By.CSS_SELECTOR, 'svg.ipc-icon--expand-more')
        button = svg.find_element(By.XPATH, './ancestor::button')
        driver.execute_script("arguments[0].click()", button)
        print("✅ Klik tombol 'Load More'")
        time.sleep(2)
    except NoSuchElementException:
        print("✅ Semua review sudah dimuat.")
        break
```

Hasil dari kode di atas:



The screenshot shows a terminal window with a light gray background and white text. It contains 20 identical lines of text, each starting with a small blue square icon followed by the text 'Klik tombol 'Load More''. This indicates that the script successfully clicked the 'Load More' button 20 times.

5. Langkah berikutnya adalah mengatur item atau elemen apa saja yang ingin di scrap (diambil) datanya. Pada praktikum ini, data yang ingin diambil adalah elemen nama reviewer, tanggal review, rating dan komentar.

```
from selenium.webdriver.common.by import By
import re

# Ambil elemen review
review_blocks = driver.find_elements(By.CSS_SELECTOR, 'article[class*="user-review-item"]')
print(f"✓ Total review ditemukan: {len(review_blocks)}")

data = []
for review in review_blocks:
    try:
        title = review.find_element(By.CSS_SELECTOR, 'h2[data-testid="subtitle"]').text.strip()
    except:
        title = '-'

    try:
        user = review.find_element(By.CSS_SELECTOR, 'a[data-testid="author-link"]').text.strip()
    except:
        user = '-'

    try:
        tanggal = review.find_element(By.CSS_SELECTOR, 'li.review-date').text.strip()
    except:
        tanggal = '-'

    try:
        rating = review.find_element(By.CSS_SELECTOR, 'span.ipc-rating-star--rating').text.strip()
    except:
        rating = '-'

    try:
        comment = review.find_element(By.CSS_SELECTOR, 'div[data-testid="review-content"]').text.strip()
    except:
        try:
            comment = review.find_element(By.CSS_SELECTOR, 'div.ipc-html-content-inner-div').text.strip()
        except:
            comment = '-'

    data.append({
        'Nama Reviewer': user,
        'Tanggal': tanggal,
        'Judul Review': title,
        'Rating': rating,
        'Komentar': comment
    })

# Tutup browser
driver.quit()
```

Akses ke dalam css seperti ini `By.CSS_SELECTOR, 'div[data-testid="review-content"]` tergantung dari css yang digunakan oleh halaman website. Sehingga jika menggunakan praktikum berikutnya menggunakan website yang berbeda, maka akses elemen css juga akan berbeda.

Hasil dari kode ini adalah:

Total review ditemukan: 25

Hasil scrap adalah jumlah review yang ditampilkan dalam satu halaman sebelum menekan tombol “load more”. Pada kasus ini, halaman review IMDb hanya menampilkan 25 review dalam satu halaman. Dan menampilkan 25 review berikutnya ketika tombol “load more” ditekan.

6. Langkah yang terakhir adalah menyimpan hasil scraping ke dalam file excel, agar memudahkan untuk mengolah data ke tahapan pre-processing

```
import pandas as pd

# Simpan ke Excel
df = pd.DataFrame(data)
df.to_excel("review_toothless.xlsx", index=False)

print("✅ Data review berhasil disimpan ke file Excel!")
```

Hasil dari kode ini adalah:

Data review berhasil disimpan ke file Excel!

7. Kemudian kembali ke folder root project untuk melihat apakah file excel hasil scraping sudah tersimpan

| Name | Date modified | Type | Size |
|--------------------|------------------|----------------------|----------|
| .ipynb_checkpoints | 03/07/2025 03:40 | File folder | |
| live_chat | 01/07/2025 11:18 | Microsoft Excel C... | 2.783 KB |
| review_toothless | 03/07/2025 04:05 | Microsoft Excel W... | 290 KB |

8. Buka file excel untuk melihat hasil scraping

| nama Review | Tanggal | adult Review | Rating | Komentar |
|-------------|--------------|--------------|--------|--|
| bahae19 | Jun 7, 2025 | 9 | 9 | Honesty... I didn't expect to feel the same way I did back in 2010, but this film brought it all back. The remake went far beyond my expectations. I went with low expectations because the original was already too good. And not to my disappointment! It was truly one of the best movies I've seen in a long time. |
| daentjeva | Jun 7, 2025 | 9 | 9 | This movie runs mostly as the original animation. Could be said, one of the most stunning live-action remake. Graphic cool, CGI cool. |
| wayangyo | Jun 8, 2025 | 9 | 9 | This is my first time reviewing a movie! English is not my native language, so excuse me my sentences! I go to the cinema with my brother and we always watch the original version. This movie is a great surprise! |
| trihouwini | Jun 7, 2025 | 10 | 10 | The long-awaited live-action adaptation of the beloved animated classic How To Train Your Dragon exceeds all expectations. Direct |
| donalstrevi | Jun 8, 2025 | 10 | 10 | I think one of the problems with this current live action era is how they treat the original material and go: "Yes we are not doing that" an |
| lauran-09 | Jun 8, 2025 | 10 | 10 | Wow! This film was nothing short of breathtaking. How To Train Your Dragon [2025] completely exceeded my expectations - and the |
| spiderman | Jun 7, 2025 | 10 | 10 | Even though the movie is probably the best live-action adaptation it doesn't come close to the original. There are many shots missing |
| panosvoul | Jun 21, 2025 | 6 | 6 | Kept true to the story. No unneeded changes. Tiny differences made it even better. THIS IS HOW YOU DO IT. The actors were perfect |
| manicaku | Jun 9, 2025 | 10 | 10 | Great movie not sure about the plot. It's much like the original cartoon but I guess excellent entertainment. I enjoy the ci |
| UnguquePat | Jun 23, 2025 | 7 | 7 | I just got out of the early screening and I feel so happy. This live-action adaptation is honestly EVERYTHING a fan could hope for - tr |
| imbotben | Jun 7, 2025 | 10 | 10 | For my personal taste, I'd rate it a 7 out of 10 mainly because I was hoping for a more adult tone, especially since it's a live-action a |
| khammou | Jun 12, 2025 | 7 | 7 | You could tell by the face, the voice, the posture, the presence that Gerard Butler really loves his character. There is just so mi |
| morkoff | Jun 11, 2025 | 5 | 5 | *spoiler free review* It's basically a copy and paste of the original. Will kids enjoy it? Yes. But the acting is terrible. Almost none of i |
| afloes_em | Jun 15, 2025 | 5 | 5 | As someone who grew up a fan of the original HTTYD films, watching the live-action adaptation was an emotional experience. It tra |
| ammonoxde | Jun 8, 2025 | 10 | 10 | Just Got Out of a Screening of How To Train Your Dragon.I never saw the original film, so I don't know if it's a shot for shot remake. |
| DoNotCox | Jun 10, 2025 | 10 | 10 | Adrenalin, a very beautiful atmosphere, very good acting, but Astrid is not what I expected in animated films. But I liked it, and I rec |
| dissfz | Jun 7, 2025 | 10 | 10 | The trend of transitioning animated movies to live-action is on the rise these years, with Lilo & Stitch and other movies. This time, w |
| Mysterygo | Jun 12, 2025 | 8 | 8 | When I heard they were making a 'live action' re-make of 'How to Train Your Dragon', my first thought was 'Why?' The original carto |
| comps_78 | Jun 15, 2025 | 6 | 6 | How to Train Your Dragon is an incredibly faithful remake where every change it makes is subtle and avoids having disastrous conse |
| masonsai | Jun 10, 2025 | 8 | 8 | Hiccup is wholesome, just like in the original. Astrid is awesome, just like in the original. Toothless is adorable, just like in the original. S |
| cjph-8112 | Jun 19, 2025 | 6 | 6 | Absolutely epic. When this movie was first announced and as the trailer rolled around I was relatively conservative about how good |
| steverem | Jun 9, 2025 | 10 | 10 | While this was not a bad effort the animated trilogy was by far the better effort while this live action remake was kinda boring in tw |
| theroman | Jun 17, 2025 | 6 | 6 | If I hadn't watched the original, I would have given this movie an 8/10. But I have. So I had to ask: WHAT'S THE POINT????This movie ha |
| dritybyshe | Jun 19, 2025 | 5 | 5 | The film is a decent attempt at the original HTTYD but it lacks the heart and soul of the original. I enjoyed it but it wasn't as good |
| Reynald | Jun 19, 2025 | 7 | 7 | Overall, I would give this movie a 7/10. It's a good movie but it's not as good as the original. I enjoyed it but it wasn't as good |

2.2. Akuisisi Data dari Sensor / IoT

2.3. Akuisisi Live Stream Data

Download dataset dari Youtube

Membaca data dari Youtube API v3 (menggunakan consol google cloud)

Pembuatan Youtube API Key

- Daftar akun di <https://console.cloud.google.com>

- Pilih menu API & Services, pilih enable API & Services

3. Scroll ke bawah pilih menu Youtube Data API v3

The screenshot shows the Google Cloud API Library interface. On the left, there's a sidebar with various services like Storage, Monitoring, Media, YouTube, Google Workspace, Security Command Center Services, Firebase, and Media and Entertainment. In the main area, there are three cards: 'YouTube Data API v3' (selected and highlighted with a red box), 'YouTube Analytics API', and 'YouTube Reporting API'. Each card has a brief description and a 'View all (3)' link.

4. Jika sudah pernah membuat API sebelumnya, maka akan muncul tombol manage.

The screenshot shows the 'Product details' page for the YouTube Data API v3. It features a large play button icon and the text 'YouTube Data API v3' and 'Google'. Below this, a description states: 'The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...'. At the bottom, there are three buttons: 'Manage' (highlighted with a red box), 'Try this API', and a green 'API Enabled' indicator.

Namun jika sebelumnya belum pernah membuat API sebelumnya, maka akan muncul tombol enable. Pilih tombol enable atau manage.

The screenshot shows the 'Product details' page for the YouTube Data API v3. It features a large play button icon and the text 'YouTube Data API v3' and 'Google'. Below this, a description states: 'The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...'. At the bottom, there are two buttons: 'Enable' (highlighted with a red box) and 'Try this API'.

5. Pilih create credential

The screenshot shows the 'API/Service Details' page for the YouTube Data API v3. At the top, there's a note: 'To use this API, you may need credentials.' A 'Create credentials' button is highlighted with a red box. Below this, there's a brief description: 'The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists, and channels.' The service is listed as 'YouTube Data API v3' with a play button icon. It's provided by Google. The status is 'Enabled'. There are links for 'Documentation' (with a 'Learn more' link), 'Explore' (with a 'Try in API Explorer' link), and 'Support' (with a 'Maintenance & Support' link).

6. Kemudian pilih public data pada menu berikut

This screenshot shows the 'Create credentials' wizard, step 1: 'Credential Type'. It asks 'Which API are you using?' and shows 'YouTube Data API v3' selected in a dropdown. It then asks 'What data will you be accessing?' and shows 'Public data' selected. A note says: 'Google data that is publicly available, like public Maps data showing restaurant information. This will create an API key.' A 'Next' button is at the bottom.

Setelah itu klik next dan tunggu hingga loading selesai

7. Kemudian akan muncul API key yang akan di copy ke jupyter notebook

This screenshot shows the 'Your Credentials' page. It shows a generated API key: 'AlzaSyC5kRPyZc3L6pq56dzsaV-FXUkm8bQJtg'. A note says: 'We recommend restricting this key before using it in production.' A 'Restrict key' button is shown. Below the key, it says: 'Here is your API key. This is always available for you on the [credentials page](#)'. There are 'Done' and 'Cancel' buttons at the bottom.

Pembuatan jupyter notebook

2. Install library google-api-python-client terlebih dahulu

```

Anaconda Prompt - devlivelit
C:\Users\fxb00hc>pip install google-api-python-client
Collecting google-api-python-client
  Downloading google-api-python-client-2.174.0-py3-none-any.whl.metadata (7.0 kB)
Collecting httplib>=1.0.0,<19.0 (from google-api-python-client)
  Downloading httplib-1.22.0-py3-none-any.whl.metadata (2.6 kB)
Requirement already satisfied: google-auth<2.24.0,!=2.28.0,<3.0.0,>=1.32.0 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-api-python-client<2.40.2)
Collecting google-auth-httplib2<1.8.0,>=0.1.0 (from google-api-python-client)
  Downloading google_auth_httplib2-0.2.2-py2.py3-none-any.whl.metadata (2.1 kB)
Collecting google-api-core<2.35.1-py3-none-any.whl.metadata (3.0 kB)
  Downloading google_api_core-2.35.1-py3-none-any.whl.metadata (2.6 kB)
Collecting uritemplate<4.2.8-py3-none-any.whl.metadata (2.6 kB)
  Downloading uritemplate-4.2.8-py3-none-any.whl.metadata (2.8 kB)
Collecting googleapis-common-protos<2.8.0,>=1.56.2 (from google-api-core<2.8.0,!=2.1.*,>=2.2.*,>=2.3.0,<3.0.0,>=1.31.5>google-api-python-client)
  Downloading googleapis_common_protos-1.70.0-py3-none-any.whl.metadata (9.3 kB)
Requirement already satisfied: protobuf<3.20.0,!=3.20.1,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,!=4.21.6 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-api-core<2.8.0,!=2.1.*,>=2.2.*,>=2.3.0,<3.0.0,>=1.31.5>google-api-python-client)
Collecting proto-plus<2.0.0,>=1.22.3 (from google-api-core<2.8.0,!=2.1.*,>=2.2.*,>=2.3.0,<3.0.0,>=1.31.5>google-api-python-client)
  Downloading proto_plus-1.26.1-py3-none-any.whl.metadata (3.2 kB)
Requirement already satisfied: requests<3.0.0,>=2.18.8 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-api-core<2.8.0,!=2.1.*,>=2.2.*,>=2.3.0,<3.0.0,>=1.31.5>google-api-python-client) (2.32.33)
Requirement already satisfied: cachetools<6.8,>=2.8.0 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-auth<2.24.0,!=2.25.0,<3.0.0,>=1.32.0>google-api-python-client) (5.5.1)
Requirement already satisfied: pyasn1-modules<0.2.1 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-auth<2.24.0,!=2.25.0,<3.0.0,>=1.32.0>google-api-python-client) (0.4.2)
Requirement already satisfied: rsa<5,>=3.1 in c:\users\fxb00hc\anaconda3\envs\tensor_env\lib\site-packages (from google-auth<2.24.0,!=2.25.0,<3.0.0,>=1.32.0>google-api-python-client) (4.9.1)

```

- Kemudian buka jupyter notebook dan buat sebuah notebook baru. Import semua library yang dibutuhkan dan set-up Youtube API Key dan Youtube Video ID yang akan di scrap

```

from googleapiclient.discovery import build
import time
import csv

API_KEY = 'AIzaSyBEm5SqUP_h7pbpTs02_4IUxlt4A4SKjBc'
VIDEO_ID = 'V90rh1M_gDo' # didapat dari halaman video Youtube yang sedang live

youtube = build('youtube', 'v3', developerKey=API_KEY)

```

Video_ID didapat dari halaman video Youtube yang sedang live. **Video yang akan di scrap harus sedang live dan kolom chat live stream harus dibuka** (jika kolom chat di disable maka halaman video tidak bisa di scrap).



- Kemudian buat fungsi untuk masuk ke Video_ID live

```

# Step 1: Get Live Chat ID
def get_live_chat_id(video_id):
    response = youtube.videos().list(
        part='liveStreamingDetails',
        id=video_id
    ).execute()
    live_details = response['items'][0]['liveStreamingDetails']
    return live_details['activeLiveChatId']

```

5. Scrap chat live

```

# Step 2: Fetch Chat Messages
def get_chat_messages(live_chat_id):
    response = youtube.liveChatMessages().list(
        liveChatId=live_chat_id,
        part='snippet,authorDetails'
    ).execute()
    messages = []
    for item in response['items']:
        msg = item['snippet']['displayMessage']
        user = item['authorDetails']['displayName']
        messages.append((user, msg))
    return messages

```

6. Stream chat live dan simpan ke dalam file CSV

```

# Step 3: Stream and Save to CSV
def stream_chat(video_id):
    live_chat_id = get_live_chat_id(video_id)
    with open('live_chat.csv', 'w', newline='', encoding='utf-8') as f:
        writer = csv.writer(f)
        writer.writerow(['User', 'Message'])
        print("Streaming started... Press Ctrl+C to stop.")
        try:
            while True:
                messages = get_chat_messages(live_chat_id)
                for user, msg in messages:
                    print(f"{user}: {msg}")
                    writer.writerow([user, msg])
                time.sleep(5)
        except KeyboardInterrupt:
            print("Streaming stopped.")

# Run
stream_chat(VIDEO_ID)

```

7. Kemudian running semua cell notebook dan akan keluar hasil stream data seperti berikut

```

[*] # Run
stream_chat(VD506_HC)

Streaming started... Press Ctrl+C to stop.
Endul Susanti: Bisamillahramaastrahim
Endul: Tolong dong kameru drone nya 4k atau 8k gitu... maha acara kerigasaa kameru 1880p.. ah eloh.
Endul: Kalah sama drone gantuan ah.
Gomoy : BOY Surya cu bukal jadi TERSANGKA 🚫⚠️⚠️⚠️
Firenda Enggar: ini kali yang pertama smangat polisi
Suaranya Suaranya: TEMPAT ISOLI HUT BHAYANGKARA KE-79 , HARAPAN ANAK, PARUSA KEDUA SIA SIA , JILOH DAW AIEL.
Kawan tolong PK Prabowo... sekarangnya ganti kapoldri... selama tiggit listio menjahat jadi Kapoldri bryoli kasus yang tidak tantas...tolong PK Prabowo denga rkaan suara rakyat
endi aewli: sukses selalu POLRI...
DOKUM SAMA RENDAH BERDIRI SAMA TINGGI : assalamualaikum warahmatullahi warahmatulah THL dan polri selamat, seja, juga peristruum walailikumalaik warahmatull ahdi wabarakatuh.
Adji Official: IAMA BAGAY ESTIMENYA 🌟
Adji Official: IAMA BAGAY ESTIMENYA 🌟
DOKUM SAMA RENDAH BERDIRI SAMA TINGGI : semoga bermartabat menjadikan bangsa Indonesia menjadi bangsa yang besar dan di hormati bangsa lain.
Tolong ajaya: ma relli Harjo Susanto Sutiyono Ray Suryo gak ada salah
endi aewli: Presiden Prabowo seharusnya selalu amanin.
Tio Setiawati: Kang dedi Mulyadi manis nihhhh
Pragaming#*: @Endul: koko lo yang smot minimal bayarin tut drone jangan caca kuman
Endi Magor: digataku bayangkara ke 79
DOKUM SAMA RENDAH BERDIRI SAMA TINGGI : mari bersatu HNL - jaya sentosa Indonesia, semuanya merdeka tetapi merdeka walailikumalaik warahmatullahi warahmatul ah... amda yarabbel aamdin.
Etsuritt: Maay Allah Alhadhillah Allahu Akbar Selamat Bekerja Presiden Pakap Jenderal Prabowo Subianto
Firenda Enggar: siangnya error binkin masalah
Putri Palpus: amp2
Halizi Limi: Bismillah penenun bangsa ini kemen sekali
Suferto Wijaya: BRAVO BAPAK PRABOWO SUBIANTO , PRESIDEN INDONESIA KOTA RAKYAT MASYARAKAT INDONESIA. SETELAH BAPAK LEMBES DEJANJUTKAN KADER BAPAK yg TERRABE Y
AZTU KAHN HEDD MULYADIS 🌟🌟🌟🌟🌟
Hengky Herwan Subiyanto: BRAVO POLRI
Kira Kartika: emangnya trut anak2 Indonesia
Mitra Ajaya: polri klu empang panje manekuk batikin
DOKUM SAMA RENDAH BERDIRI SAMA TINGGI : Ganteng tv selalu bertutu-huti - jaya semoga nikah
Endul Susanti: Bisamillahramaastrahim
Endul: Tolong dong kameru drone nya 4k atau 8k gitu... maha acara kerigasaa kameru 1880p.. ah eloh.
Endul: Kalah sama drone gantuan ah.
Gomoy : BOY Surya cu bukal jadi TERSANGKA 🚫⚠️⚠️⚠️

```

8. Hasil stream data juga telah tersimpan otomatis ke dalam file csv di folder project

| Name | Date modified | Type | Size |
|----------------------|------------------|-----------------------|--------|
| .ipynb_checkpoints | 01/07/2025 09:45 | File folder | |
| live_chat.ipynb | 01/07/2025 10:16 | Microsoft Excel Co... | 557 KB |
| twitter_stream.ipynb | 01/07/2025 00:31 | IPYNB File | 10 KB |
| youtube_stream.ipynb | 01/07/2025 09:57 | IPYNB File | 50 KB |

2.4. Pra-Pemrosesan Sederhana

Pra_pemrosesan (Pre-processing) data disesuaikan dengan kebutuhan data. Pada praktikum ini data yang diakuisisi adalah data teks atau bahasa dan data citra, sehingga tahapan-tahapan pre-processing data akan menyesuaikan dengan tahapan Natural Language Processing (NLP) dan tahapan Computer Vision Ketika menggunakan data citra, audio atau video, maka tahapan pre-processing akan berbeda. Seperti jika pada data citra, tahapan pre-processing yang umum adalah cropping citra, menghapus background objek, konversi citra rgb menjadi citra biner (jika fitur yang digunakan adalah fitur bentuk, namun jika fitur yang digunakan adalah fitur warna, maka citra tetap rgb).

2.4.1. Pre-Processing Data Teks (Natural Language Processing)

Untuk kegiatan praktikum berikutnya, data yang digunakan adalah data review film dari website IMDb. Berikut tahapan pre-processing data teks atau bahasa:

1. Loading Data

Langkah pertama dalam membersihkan data adalah dengan menambahkan data ke dalam jupyter notebook. Berikut kode loading data excel:

```

import pandas as pd

# 1. Baca file Excel
df = pd.read_excel("review_toothless.xlsx")
df.head() #menampilkan 5 data teratas

```

Hasil dari kode diatas adalah:

| [1]: | Nama Reviewer | Tanggal | Judul Review | Rating | Komentar |
|------|---------------------|-------------|--------------|--------|--|
| 0 | bahae19 | Jun 7, 2025 | | - | 9 Honestly... I didn't expect to feel the same w... |
| 1 | daantjevanwijk | Jun 7, 2025 | | - | 9 I went with low expectations because the origi... |
| 2 | wayanpyott | Jun 8, 2025 | | - | 9 This movie runs mostly as the original animati... |
| 3 | trihouzinho | Jun 7, 2025 | | - | 10 This is my first time reviewing a movie!\n\nEn... |
| 4 | danieltrevino-73758 | Jun 8, 2025 | | - | 10 The long-awaited live-action adaptation of the... |

- Menyiapkan kolom baru untuk menampung hasil pre-processing

```

# 2. Siapkan kolom kosong untuk hasil
df['Komentar Bersih'] = ''

```

- Menyiapkan kamus stopword menggunakan library nltk

```

import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer

nltk.download('punkt')
nltk.download('stopwords')

# 3. Siapkan alat NLP
stop_words = set(stopwords.words('english'))
stemmer = PorterStemmer()

```

```

[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\FX506HC\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\FX506HC\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

```

- Membersihkan data berdasarkan tahapan pre-processing Natural Language Processing

| Tahapan | Fungsi |
|-------------------------|---|
| 1. Lowercasing | Menstandarkan semua huruf menjadi huruf kecil |
| 2. Tokenization | Memecah teks menjadi unit kata atau kalimat |
| 3. Removing Punctuation | Menghapus tanda baca |
| 4. Stopwords Removal | Menghapus kata-kata umum (dan, yang, dengan, dll) yang tidak bermakna kontekstual |

| | |
|--------------------------------|--|
| 5. Stemming/Lemmatization | Mengubah kata ke bentuk dasar (makan → makan, memakan → makan) |
| 6. Removing Numbers | Menghapus angka (jika dianggap tidak relevan) |
| 7. Removing Whitespace | Menghapus spasi ekstra |
| 8. Removing Special Characters | Menghapus karakter seperti @, #, \$, dsb |

```

import string
from nltk.tokenize import word_tokenize

# 4. Preprocessing satu per satu baris (tanpa fungsi)
for i in range(len(df)):
    #Ambil teks
    text = str(df.loc[i, 'Komentar'])

    #Lowercase
    text = text.lower()

    #Hapus tanda baca
    text = text.translate(str.maketrans(' ', ' ', string.punctuation))

    #Tokenisasi
    tokens = word_tokenize(text)

    #Hapus stopword
    tokens = [word for word in tokens if word not in stop_words]

    #Stemming
    tokens = [stemmer.stem(word) for word in tokens]

    #Gabungkan kembali
    clean_text = ' '.join(tokens)

    #Simpan ke kolom baru
    df.loc[i, 'Komentar Bersih'] = clean_text

#tampilkan 5 data teratas
print(df[['Komentar', 'Komentar Bersih']].head())

```

| | Komentar \ |
|---|---|
| 0 | Honestly... I didn't expect to feel the same w... |
| 1 | I went with low expectations because the origi... |
| 2 | This movie runs mostly as the original animati... |
| 3 | This is my first time reviewing a movie!\n\nEn... |
| 4 | The long-awaited live-action adaptation of the... |

| | Komentar Bersih |
|---|---|
| 0 | honestli didnt expect feel way back 2010 film ... |
| 1 | went low expect origin alreadi good disappoint... |
| 2 | movi run mostli origin anim could said one stu... |
| 3 | first time review movi english nativ languag e... |
| 4 | longawait liveact adapt belov anim classic tra... |

5. Simpan hasil pre-processing ke dalam file excel

```
# 5. Simpan hasil
df.to_excel("data_toothless_clean.xlsx", index=False)

print("✅ Pre-processing selesai, hasil disimpan ke
'data_toothless_clean.xlsx'")
```

✅ Pre-processing selesai, hasil disimpan ke 'data_toothless_clean.xlsx'

2.4.2. Pre-Processing Data Citra (Computer Vision)

Seperti yang dijelaskan sebelumnya, tahapan pre-processing (pengolahan) data berbeda tergantung dari jenis datanya. Pada praktikum ini, data citra akan digunakan untuk ranah computer vision (bukan Pengolahan Citra Digital).

Perbedaan Computer Vision dan Pengolahan Citra Digital

Ranah computer vision dan pengolahan citra digital memiliki beberapa perbedaan, yaitu:

| Aspek | Pengolahan Citra Digital (DIP) | Computer Vision (CV) |
|------------------------|--|--|
| Tujuan utama | Memperbaiki atau memodifikasi citra | Memahami dan mengekstrak informasi dari citra |
| Fokus kerja | Transformasi pixel, filter, konversi warna, segmentasi | Deteksi objek, klasifikasi, pengenalan wajah, pelacakan |
| Tingkat kecerdasan | Tidak membutuhkan kecerdasan buatan (AI) | Biasanya menggunakan AI, ML, atau Deep Learning |
| Contoh teknik | Filtering (Gaussian, median), thresholding, histogram equalization | CNN, YOLO, SVM untuk klasifikasi, deteksi objek |
| Input & output | Input: gambar → Output: gambar baru | Input: gambar → Output: informasi (label, posisi objek) |
| Contoh hasil akhir | Gambar lebih tajam, lebih kontras, bebas noise | Gambar + label: "Ini anjing", "Objek ada di koordinat X,Y" |
| Contoh tools/libraries | OpenCV, PIL, MATLAB, skimage | OpenCV, TensorFlow, PyTorch, Detectron, Keras |

Contoh ilustrasi:

| Masalah | Pengolahan Citra Digital | Computer Vision |
|---------------------|---|---|
| Gambar buram | Diperjelas dengan sharpen filter | Tidak fokus; CV bisa gagal mengenali objek |
| Warna terlalu gelap | Perbaikan kontras dengan histogram equalization | Deteksi objek mungkin gagal tanpa preprocessing |

| | | |
|--------------------|------------------|--|
| Foto wajah | Crop & grayscale | CV akan mendeteksi wajah, bahkan mengenali siapa |
| Kamera lalu lintas | Perbaiki noise | CV akan hitung jumlah kendaraan, klasifikasi mobil/motor |

Persamaan Computer Vision dan Pengolahan Citra Digital terletak pada tahapan pre-processing data. Computer Vision sering menggunakan Pengolahan Citra Digital sebagai tahap awal (pre-processing) sebelum melakukan pemahaman lebih dalam terhadap gambar.

Untuk data citra tahapan pengolahannya memiliki beberapa tahapan, yaitu:

| Tahap | Tujuan |
|------------------------------|--|
| 1. Resize | Menyesuaikan ukuran gambar (misalnya 64×64 px) |
| 2. Normalization | Menstandarkan nilai piksel (0–255 menjadi 0–1 atau -1 sampai 1) |
| 3. Augmentation (opsional) | Menambah variasi data (flip, rotate, zoom, dsb) untuk mencegah overfitting |
| 4. Grayscale (opsional) | Mengurangi saluran warna jika hanya perlu intensitas abu-abu |
| 5. Noise Removal (opsional) | Mengurangi gangguan seperti blur, shadow, dsb |
| 6. Color Space Conversion | Misalnya dari BGR ke RGB (karena OpenCV defaultnya BGR) |
| 7. Flattening (jika KNN/SVM) | Mengubah gambar 2D menjadi vektor 1D (array 1 dimensi) |

Pre-Processing untuk Data 1 Citra

Untuk menjadi contoh, maka kegiatan praktikum dimulai dengan mengolah data 1 buah citra.

1. Loading Dataset

```
import cv2
import matplotlib.pyplot as plt

# Load gambar
img = cv2.imread("C:/Users/FX506HC/DEMO/flask-backend/dataset/bukan_teh/daun-sukun.jpeg") # defaultnya BGR

plt.imshow(img)
plt.axis('off')
plt.title("Gambar dari OpenCV")
plt.show()
```

Hasil dari kode di atas:



Selain menggunakan library cv, library lain yang bisa digunakan untuk loading dataset citra adalah library PIL dengan format kode:

```
from PIL import Image

img = Image.open("C:/Users/FX506HC/DEMO/flask-backend/dataset/bukan_teh/daun-sukun.jpeg")
img.show()
```

2. Resize Citra

```
# 1. Resize
img_resized = cv2.resize(img, (64, 64))
```

`cv2.resize(img, (64, 64))` memperkecil ukuran citra menjadi 64×64 , kalau mau mengganti dengan ukuran lain maka ubah di `(64, 64)`.

3. Konversi warna citra dari BGR menjadi RGB

```
# 2. Convert BGR ke RGB
img_rgb = cv2.cvtColor(img_resized, cv2.COLOR_BGR2RGB)
```

Library cv memiliki channel warna default BGR (Blue Green Red) dan harus dikonversi menjadi RGB (Red Green Blue) agar warna objek dalam citra tidak kemerahan atau kebiruan.

4. Mengubah rentang nilai citra menjadi 0-1

```
# 3. Normalisasi (0-1)
img_normalized = img_rgb / 255.0
```

Citra yang diolah oleh komputer aslinya ada kumpulan angka yang tersimpan dalam struktur data array. Angka yang terdapat dalam citra tergantung dari channel warna yang digunakan. Contoh jika channel warna yang digunakan adalah RGB, channel Red memiliki rentang nilai 0-254 (total nilai 255), sehingga untuk menormalkan nilai dengan rentang yang luas dibagi dengan 255.

5. Konversi citra RGB menjadi citra grayscale

```
# 4. (Opsional) Ubah ke grayscale  
img_gray = cv2.cvtColor(img_resized, cv2.COLOR_BGR2GRAY)
```

Konversi menjadi citra grayscale adalah langkah opsional, jika metode nantinya menggunakan fitur bentuk, tahap grayscale bisa digunakan. Tapi jika fitur yang digunakan adalah fitur warna, maka citra akan tetap dalam bentuk RGB.

6. Konversi array 2D menjadi array 1D

```
# 5. (Opsional) Flatten jika pakai KNN  
img_flatten = img_gray.flatten() # dari (64,64) → (50176,)
```

Jika metode yang digunakan adalah KNN atau SVM maka data citra yang awalnya adalah array 2D harus diubah menjadi array 1D.

7. Tampilkan hasil pengolahan citra

```
# Tampilkan hasil resize & normalisasi  
plt.imshow(img_rgb)  
plt.title("Gambar setelah Pre-processing")  
plt.axis('off')  
plt.show()
```



Pre-Processing untuk Data Klasifikasi

1. Labeling citra

Tahapan labeling citra dilakukan pertama kali karena proses labeling pada data citra dilakukan setelah mengambil citra. Labeling pada data citra adalah dengan mengelompokkan citra ke dalam folder dengan nama folder yang merupakan label data.

| Name | Date modified | Type |
|------------------|------------------|-------------|
| belum_siap_petik | 04/07/2025 11:52 | File folder |
| siap_petik | 04/07/2025 11:52 | File folder |

Kemudian setelah labeling folder, tahapan label berlanjut ke penamaan nama file. Labeling file citra disesuaikan dengan label folder. Contoh label belum_siap_petik:



2. Cropping citra

Cropping Citra dilakukan secara manual, dimana citra dipotong satu per satu. Langkah ini digunakan karena citra diambil secara manual. Citra yang diambil secara manual seringkali memiliki layout yang berbeda (portrait dan landscape). Sehingga untuk menjaga objek penelitian tetap aman dan tidak terpotong, maka tahapan cropping dilakukan secara manual.



3. Background Removal

Sama dengan tahapan cropping, tahapan background removal dilakukan secara manual. Hal ini dilakukan agar foreground dengan background bisa terpisahkan sesuai kebutuhan developer.



4. Resize Citra

```
import os

#Loading Dataset

data = []
labels = []

folder_path = 'C:/Users/FX506HC/project teh baru/dataset_teh/'
class_names = os.listdir(folder_path)
```

```
import cv2

for class_label in class_names:
    class_dir = os.path.join(folder_path, class_label)
    for filename in os.listdir(class_dir):
        if filename.lower().endswith('.jpg', '.jpeg', '.png'):
            img_path = os.path.join(class_dir, filename)
            img = cv2.imread(img_path)
            if img is not None:
                img = cv2.resize(img, (64, 64)) # Resize ke 64x64
                img = img / 255.0                 # Normalisasi 0-1
                data.append(img)
                labels.append(class_label)
```

5. Konversi citra menjadi nilai angka

```
import numpy as np

# Convert ke array numpy
X = np.array(data)
y = np.array(labels)

print(f"Jumlah data: {len(X)}, Bentuk fitur: {X.shape}, Label unik: {set(y)}")
```

Jumlah data: 1280, Bentuk fitur: (1280, 64, 64, 3), Label unik: {'belum_siap_petik', 'siap_petik'}

3. Rangkuman

Akuisisi citra merupakan tahapan awal atau pertama dari alur atau siklus machine learning. Tahapan ini sangat penting karena tahapan adalah tahapan pengambilan data. Proses akuisisi data tergantung dari jenis data yang digunakan. Untuk data dengan jenis teks, cara akuisisi adalah dengan scraping dari website, stream data menggunakan API dan dari sensor atau IoT. Untuk data citra bisa diakuisisi secara manual menggunakan kamera atau mengunduh opendataset yang telah disediakan oleh banyak website warehouse. Setelah akuisisi (pengambilan), tahapan selanjutnya adalah pre-processing (pengolahan) dataset. Tahapan ini dibutuhkan agar data yang mentah yang kotor bisa dibersihkan sebelum masuk ke tahapan training. Bersih atau tidak bersihnya data yang dimasukkan ke dalam metode akan berpengaruh terhadap nilai akurasi metode.

4. Tes Formatif

5. Daftar Pustaka

- [1] T. Geeksforgeeks, "Introduction to Web Scraping," Geeksforgeeks, 07 06 2025. [Online]. Available: <https://www.geeksforgeeks.org/web-scraping/introduction-to-web-scraping/>. [Accessed 05 07 2025].
- [2] T. Geeksforgeeks, "Free Public Data Sets For Analysis," Geeksforgeeks, 30 05 2024. [Online]. Available: <https://www.geeksforgeeks.org/data-analysis/free-public-data-sets-for-analysis/>. [Accessed 05 07 2025].
- [3] T. Geeksforgeeks, "Sources of Data Collection | Primary and Secondary Sources," Geeksforgeeks, 31 05 2024. [Online]. Available: <https://www.geeksforgeeks.org/sources-of-data-collection-primary-and-secondary-sources/>. [Accessed 05 07 2025].
- [4] T. Geeksforgeeks, "Data Collection Methods | Primary and Secondary Data," Geeksforgeeks, 15 04 2025. [Online]. Available: <https://www.geeksforgeeks.org/data-analysis/methods-of-data-collection/>. [Accessed 05 07 2025].
- [5] T. Geeksforgeeks, "Machine Learning Lifecycle," Geeksforgeeks, 17 01 2025. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/machine-learning-lifecycle/>. [Accessed 05 07 2025].
- [6] T. Geeksforgeeks, "Top Machine Learning Dataset: Find Open Datasets," Geeksforgeeks, 16 04 2024. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/top-machine-learning-dataset-find-open-datasets/>. [Accessed 05 07 2025].
- [7] T. Geeksforgeeks, "What is Data Acquisition in Machine Learning?," Geeksforgeeks, 13 05 2024. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/what-is-data-acquisition-in-machine-learning/>. [Accessed 05 07 2025].
- [8] T. Geeksforgeeks, "Open Source Data: Your Guide to the Future of Free Data Analysis and Visualization," Acceldata, 20 10 2024. [Online]. Available: https://www-acceldata-io.translate.goog/blog/open-source-data-your-guide-to-the-future-of-free-data-analysis-and-visualization?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=sge#:~:text=Open%20Source%20Data%3A%20Your%20Guide%20to%20the%20Future%20of%20Free. [Accessed 05 07 2025].