Pengantar Data Sains

#4 Meeting

Import & Cleansing Data

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 19900520202406100 1

Import Data

Setiap proyek data sains dimulai dengan langkah fundamental yang sama: memuat data ke dalam lingkungan kerja kita (R atau Python). Data bisa datang dalam berbagai format, dan setiap format memiliki fungsi spesifik untuk membacanya.

Format Data	Fungsi di R (library)	Fungsi di Python (library)
Dasar & Teks		
CSV (.csv)	read.csv() (base)	pd.read_csv() (pandas)
Text Delimited (.txt, .tsv)	read.delim() / read.table() (base)	pd.read_csv(sep='\t') (pandas)
Excel (.xlsx, .xls)	read_excel() (readxl)	pd.read_excel() (pandas, openpyxl)
Semi-Terstruktur		
JSON (.json)	fromJSON() (jsonlite)	pd.read_json() (pandas)
XML (.xml)	xmlToDataFrame() (XML)	pd.read_xml() (pandas, lxml)
Sumber Web & Database		
Tabel HTML (dari URL)	html_table() (rvest)	pd.read_html() (pandas, lxml)
SQL Database	dbReadTable() / dbGetQuery() (DBI + driver, misal: RPostgres)	pd.read_sql() / pd.read_sql_query() (pandas + SQLAlchemy + driver, misal: psycopg2)

Cleansing Data - String Matching

- Masalah di Dunia Nyata: Mengapa VLOOKUP/Join Standar Sering Gagal?
- Konsep Inti: Apa Itu Fuzzy Matching?

Cleansing Data - String Matching

- Mengapa Join Standar Sering Gagal?
- Data di dunia nyata seringkali diinput oleh manusia dan berasal dari sistem yang berbeda, sehingga menghasilkan inkonsistensi.

Tabel A (Master)	Tabel B (Transaksi)	Hasil VLOOKUP
Cynthia Dewi Lestari	Chintia Lestari	#N/A
Roti Sobek Cokelat Keju	Roti Sobek Coklat	#N/A
Weksi Budiaji	Weksi B.	#N/A

- Problem: Kesalahan ketik, singkatan, dan format yang berbeda membuat pencocokan eksak (exact match) tidak mungkin dilakukan.
- "Komputer hanya melihat string yang 100% identik."

Fuzzy Matching

- Fuzzy Matching (juga dikenal sebagai Approximate String Matching atau Record Linkage) adalah teknik untuk menemukan kecocokan antara dua string teks yang mirip, tetapi tidak identik.
- Cara Kerja Manusia: Kita bisa langsung tahu "Chintia Lestari" merujuk pada "Cynthia Dewi Lestari".
- Cara Kerja Fuzzy Matching: Meniru intuisi manusia dengan mengukur kemiripan secara matematis.
- Hasil dari Fuzzy Matching bukanlah "Cocok / Tidak Cocok", melainkan sebuah Skor Kemiripan (Similarity Score), biasanya dalam rentang 0 hingga 1 (atau 0-100).

Fuzzy Matching

- Cara Kerja Fuzzy Matching
- Prosesnya melibatkan dua komponen utama:

1. Metrik Jarak/Kemiripan (String Distance Metric)

- Algoritma yang menghitung seberapa "jauh" atau "mirip" dua string.
- Setiap algoritma punya kelebihan dan kekurangan masing-masing.
- Contoh: Levenshtein, Jaro-Winkler, Jaccard.

2. Ambang Batas (Threshold)

- Nilai skor (misalnya 0.8 atau 80%) yang kita tentukan untuk memutuskan apakah dua string dianggap "cukup mirip" untuk dicocokkan.
- Threshold yang tinggi = Lebih akurat, tapi berisiko kehilangan beberapa pasangan.
- Threshold yang rendah = Menangkap lebih banyak pasangan, tapi berisiko salah cocok.

Fuzzy Matching : Levenshtein

- Mengukur jarak antara dua string dengan menghitung jumlah minimum "edit" yang diperlukan untuk mengubah satu string menjadi string lainnya.
- "Edit" yang dihitung ada tiga:

1. Insertion (Sisip): Menambah satu karakter.
2. Deletion (Hapus): Menghapus satu karakter.
3. Substitution (Ganti): Mengganti satu karakter.

- Contoh: Untuk mengubah "SINTA" menjadi "CINTA":
- Ganti 'S' dengan 'C'.
- Jarak Levenshtein = 1.
- Untuk mengubah "ROTI" menjadi "RODA":
- Ganti 'T' dengan 'D'.
- Ganti 'I' dengan 'A'.
- Jarak Levenshtein = 2.
- Sangat baik untuk menangani kesalahan ketik (typo).

Fuzzy Matching : Levenshtein

- Contoh Jarak Levenshtein = 2 (Membutuhkan 2 "Edit")
- Contoh 1: Ganti + Ganti
- Untuk mengubah "PASAR" menjadi "BASAH":
- Ganti 'P' dengan 'B'. (String sementara: BASAR)
- Ganti 'R' dengan 'H'. (String akhir: BASAH)
- Dua langkah penggantian. Jarak Levenshtein = 2.

- Contoh 2: Sisip + Ganti
- Untuk mengubah "KATA" menjadi "KARTU":
- Sisipkan 'R' setelah 'A' pertama. (String sementara: KARTA)
- Ganti 'A' terakhir dengan 'U'. (String akhir: KARTU)
- Satu langkah sisip dan satu langkah ganti. Jarak Levenshtein = 2.

Contoh 3: Ganti + Hapus
Untuk mengubah "SELASA" menjadi "KELAS":
Ganti 'S' pertama dengan 'K'. (String sementara: KELASA)
Hapus 'A' terakhir. (String akhir: KELAS)
Satu langkah ganti dan satu langkah hapus. Jarak Levenshtein = 2.

Fuzzy Matching : Jaro-Winkler

- Metode yang lebih kompleks, dirancang khusus untuk string pendek seperti nama orang atau perusahaan.
- Mempertimbangkan karakter yang cocok dan jumlah transposisi (karakter yang urutannya terbalik, misal 'RD' vs 'DR').
- Memberikan bobot ekstra pada karakter yang cocok di awal string (awalan).

Contoh:

- MARTHA vs MARHTA -> Skor Jaro-Winkler akan sangat tinggi karena hanya ada transposisi (1 transposisi T vs H).
- FERDIAN vs FERDIYAN -> Skor akan tinggi karena mayoritas karakter (7 dari 8 sudah ada), terutama di awal, tidak ada transposisi (Urutan sempurna).
- Sangat direkomendasikan untuk mencocokkan nama orang dan alamat.

Fuzzy Matching : Jaccard Similarity

- Bekerja dengan memecah setiap string menjadi set potongan-potongan kecil (disebut n-grams) lalu mengukur kemiripan berdasarkan proporsi potongan yang sama.

Contoh (n-gram = 2 karakter):

- "MALAM" -> Set: {"MA", "AL", "LA", "AM"}
- "LAMPU" -> Set: {"LA", "AM", "MP", "PU"}

- Potongan yang sama (Intersection): {"LA", "AM"} (ada 2) Total potongan unik (Union): {"MA", "AL", "LA", "AM", "MP", "PU"} (ada 6)
- Jaccard Similarity = (Intersection / Union) = 2 / 6 = 0.33
- Baik untuk teks yang lebih panjang dan tidak terpengaruh oleh urutan kata.

Study Case : Data Penjualan Produk

1. Download Raw Data Excel pada Link Berikut :

Sample Data

2. Buka File Excel tersebut kemudian Klik “Enable Editing” untuk memastikan file ini dapat digunakan dengan baik.

3. Simpan/ Pindahkan ke file path yang diinginkan :
Misal : C:\Users\user\NAMA FILE EXCEL YANG AKAN DIPAKAI.

4. Di dalam file excel tersebut akan terdapat 3 Sheet :
   - Sheet : profil_pelanggan
   - Sheet : transaksi_pembelian
   - Sheet : master_produk

Study Case : Data Penjualan Produk

Profil Pelanggan

id_profil	nama_lengkap	alamat_domisili
P001	Ferdian Bangkit Wijaya	Jl. Merpati Putih 5 Blok C-21, Taman Cilegon
P002	Cynthia Dewi Lestari	Komp. Mawar Melati Kavling 8, Sukajadi, Bandung
P003	Ahmad Zaelani	Apartemen Cempaka Tower B Lantai 11 No. 1108, Jakarta Pusat
P004	Weksi Budiaji	Perumahan Griya Serang Indah Blok E4 Nomor 12

Master Produk

id_produk	nama_produk_resmi	harga
PROD-A	Kopi Susu Gula Aren 1L	55000
PROD-B	Roti Sobek Cokelat Keju	25000
PROD-C	Donat Gula Klasik Isi 6	45000
PROD-D	Teh Melati Segar 500ml	15000

Transaksi Pembelian

id_transaksi	nama_pelanggan	alamat_pengiriman	nama_barang	jumlah
TRX01	Ferdian B.	Jl. Merpati Putih V Blok C21, Cilegon	Kopi Gula Aren 1 Liter	1
TRX02	Weksi B.	Griya Serang Indah E4 No. 12	Roti Sobek Coklat	2
TRX03	Chintia Lestari	Komp. Mawar Melati Kav 8 Sukajadi	Donat Klasik 6pcs	1
TRX04	Ahmad Zaelany	Apt. Cempaka Tower B/11/1108 Jakpus	Kopi Susu Aren 1L	1
TRX05	Weksi Budiaji	Perum. GSI Blok E4 No. 12, Serang	Teh Melati 500ml	3

Study Case : Data Penjualan Produk

Alur Kerja Pembersihan & Penggabungan Data

Kita akan menggunakan pendekatan dua tahap untuk memastikan akurasi.

1. Tahap 1: Validasi Pelanggan
   - Fuzzy join transaksi dengan profil berdasarkan nama.
   - Hitung skor kemiripan alamat.
   - Filter hasil hanya jika kedua skor (nama & alamat) di atas threshold.
2. Tahap 2: Validasi Produk
   - Fuzzy join hasil Tahap 1 dengan produk berdasarkan nama barang.
   - Filter berdasarkan skor produk.
3. Hasil: Sebuah tabel transaksi yang bersih dan diperkaya dengan ID yang valid.

Study Case : Data Penjualan Produk

Running di R Studio

Study Case : Data Penjualan Produk

Running di Python

UNIVERSITAS
SULTAN AGENG
TIRTAYASA

UNTIRTA
Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

SEE YOU NEXT WEEK !

Ferdian Bangkit Wijaya, S.Stat., M.Si

NIP. 19900520202406100 1

ferdian.bangkit@untirta.ac.id