GETTING INSIGHTS FROM YOUR TEXT DATA

Welcome!

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

MODERATOR

Oktafian Yusuf Prasetya

Training Coordinator, Narasio Data

Webinar Rules

- Diharapkan untuk **menggunakan username sesuai nama pada regis form** dan **mematikan microphone** selama kegiatan berlangsung.
- Kamu bisa bertanya terkait materi dengan **slido**, sedangkan kolom chat zoom hanya digunakan untuk *interactive talk* dengan speaker.
- Pertanyaanmu akan dijawab oleh **asisten kelas** lewat chat atau langsung oleh ***speaker*** pada saat sesi tanya jawab.
- Kamu akan **mendapatkan e-sertifikat** jika kamu **mengikuti webinar hingga akhir** dan **mengisi feedback form**.
- Recording tersedia di **channel youtube Narasio,** link akan dikirimkan bersamaan dengan e-sertifikat H+7 setelah acara.
- Diharapkan untuk **mengaktifkan kamera dan menggunakan VBG yang telah disediakan** pada sesi foto bersama.

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

CONTOH

Syarat dan Ketentuan

1. Dikhususkan untuk peserta Narasio Datafest Workshop “Getting Insights from Your Text Data”.
2. Screenshot webinar ini yang menampilkan speaker dan materi.
3. Berikan kesanmu mengenai webinar ini melalui caption IG story.
4. Jangan lupa screenshot hal-hal yang kamu lakukan sebagai bukti!
5. Share screenshot beserta caption yang sudah kamu buat dan tag **@narasiodata** (pastikan akun IGmu tidak diprivate ya biar kami bisa melihatnya dan jangan lupa follow IG @narasiodata)
6. 3 peserta dengan kesan menarik akan mendapatkan saldo OVO sebesar Rp. 25.000 yang akan diumumkan di IG @narasiodata pada 5 Oktober 2022 (keputusan mutlak ditentukan oleh tim Narasio).

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

Drop your questions here!

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

SPEAKER

Nilam Mufidah

AI Engineer, Widya Wicara

Sept 2018 - **Data Analyst & FB Ads Specialist** at
Mangrove Inspiration

Dec 2018 - **Data Analyst** at
PT Botika Teknologi Indonesia (Botika)

Feb 2021 - **AI Engineer** at
PT. Widya Informasi Nusantara (Widya Wicara)

Github : github.com/nilammufidah
LinkedIn : https://www.linkedin.com/in/nilam-mufidah-8a50a8163/

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM YOUR TEXT DATA

Let's get started!

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

Web Scraping

Apa sih web scraping itu ?

> **Web scraping** adalah metode yang digunakan untuk mendapatkan sejumlah data/mengekstrak data dari situs web.

Web Scraping ⟶ **Data Collection**

Apakah Web Scraping Legal?

Data Protection Protocols

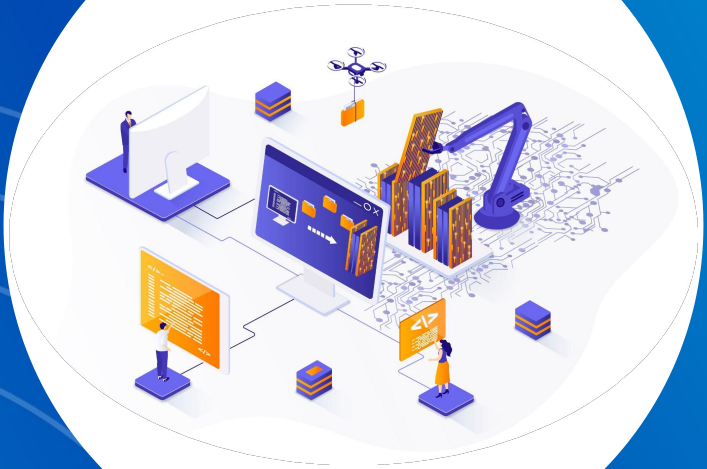
Copyrights

Terms of Service

Penyalahgunaan Trade Secrets

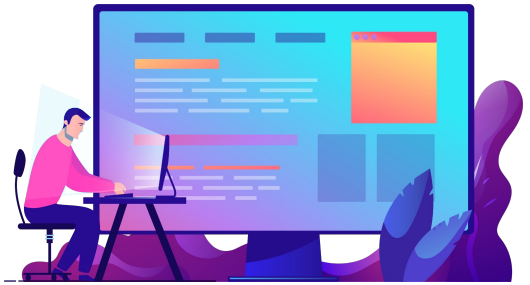
Non-public Information

SCRAPING



=

CRAWLING



?

SCRAPING	VS	CRAWLING
Data scraping tools bisa digunakan untuk **mengekstrak data** dari **berbagai source**, termasuk web		Data crawling tools digunakan untuk **mendownload/indeksasi page dari web**
Tidak perlu mengunjungi setiap halaman untuk mendapatkan informasi		**Perlu untuk mengunjungi setiap halaman hingga setiap line** untuk mendapatkan informasi
Ukuran pengambilan data sesuai kebutuhan, **bisa baik besar maupun kecil**		Pengumpulan informasi dalam **jumlah yang besar**

TOP 3 Scraping Framework

1. Beautiful Soup
2. Selenium
3. Scrapy

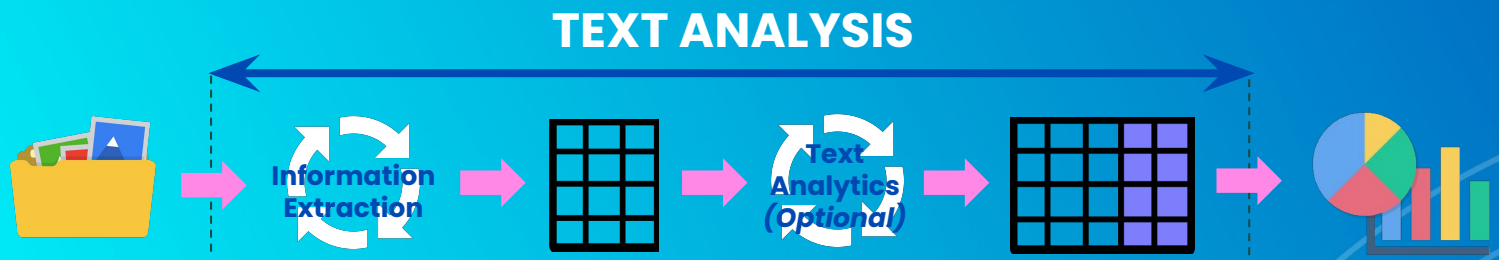
Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

Text Analyst

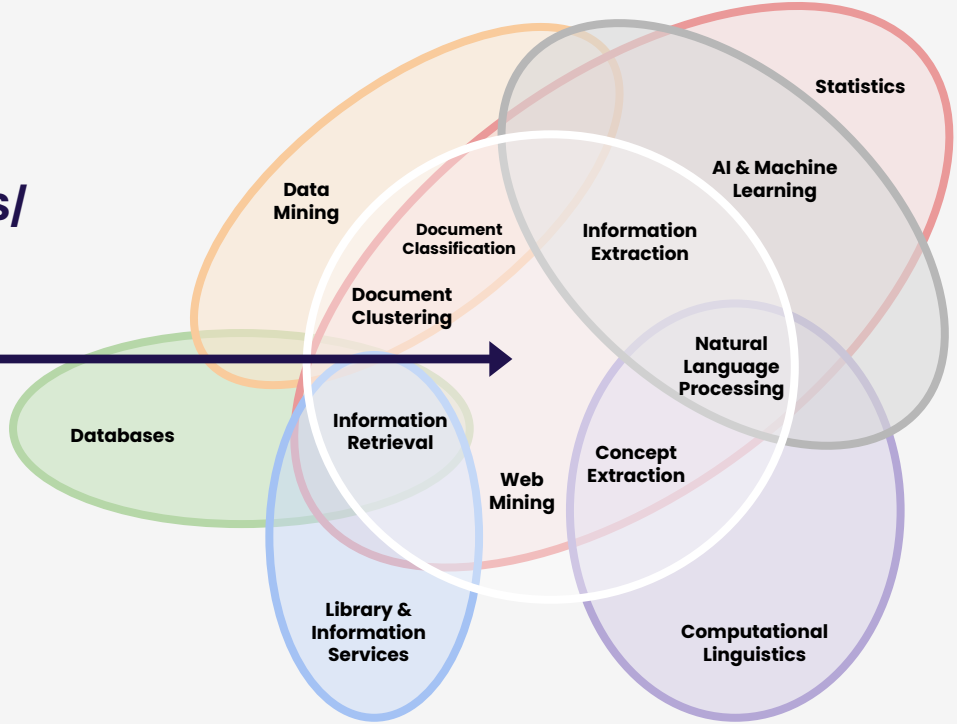
Apa itu Text Analyst?



Text analytics, atau yang sering disebut juga dengan text analysis atau text mining, adalah studi yang mempelajari proses analisis unstructured data untuk mendapatkan suatu informasi yang memiliki nilai, makna, dan pattern.

Source:
https://www.ibmbigdatahub.com/blog/why-analyzing-text-so-hard

TEXT ANALYSIS/ TEXT MINING

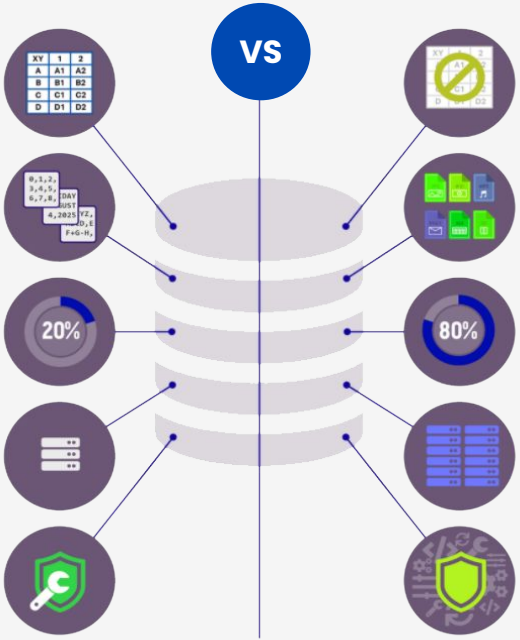


Source: PURE SPEECH TECHNOLOGY

Tipe Data

STRUCTURED DATA

- **Can** be displayed in rows, columns, and relational databases
- Numbers, dates, strings
- Estimated **20%** of enterprise data *(Gartner)*
- Requires **less** storage
- **Easier** to manage and protect with legacy solutions



UNSTRUCTURED DATA

- **Cannot** be displayed in rows, columns, and relational databases
- Images, audio, video, word processing files, e-mails, spreadsheets
- Estimated **80%** of enterprise data *(Gartner)*
- Requires **more** storage
- **More difficult** to manage and protect with legacy solutions

APA YANG MEREKA PIKIRKAN?

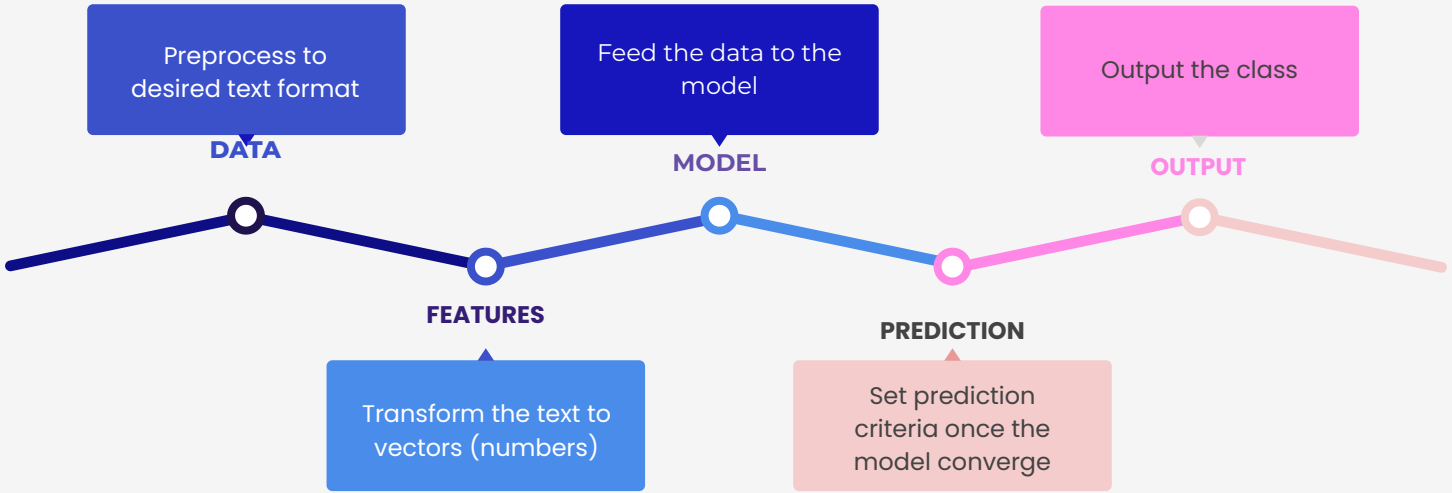
SENTIMENT ANALYSIS

Source:
https://www.surveysensum.com/customer-experience/sentiment-analysis/

TANTANGAN DALAM TEXT ANALYTICS

- Mendapatkan kesimpulan dari format yang tidak tentu dan sulit untuk dianalisis dengan cara biasa.
- Mendapatkan pattern dari text data yang dimiliki.
- Perbedaan case study mempengaruhi preprocessing data.
- Terdapat slang, sarkasme, dan pengaruh budaya yang mempengaruhi tata bahasa.
- Keterbatasan corpus dalam beberapa bahasa.

GENERAL PROCESS

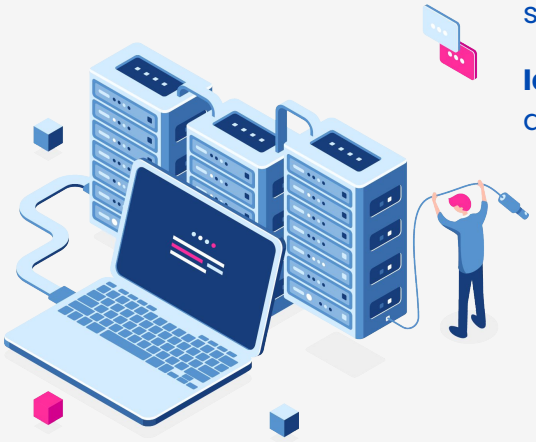


Source: https://medium.com/@raxsuite/text-classification-69f60e0e2ce5

TIPE TEKNIK PROCESSING

- Lower/Upper Casing
- Menghilangkan Karakter yang Tidak Diharapkan
- Tokenization
- Stemming
- Lemmatization
- Stopword Removal

Tf-Idf



Tf adalah term frequency dimana menunjukkan frekuensi kata "*t*" di dalam suatu dokumen "*d*".

Idf adalah inverse document frequency menunjukkan perhitungan suatu kata didistribusikan dalam corpus yang digunakan.

TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen.

Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus.

LIBRARY POPULER UNTUK TEXT ANALYTICS

- NLTK
- TextBlob
- spaCy
- Sastrawi
- Gensim
- Regex

Time to **Practice**

github.com/nilammufidah/WorkshopNLPNarasio

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

Drop your questions here!

RAIH KARIR TERBAIKMU DENGAN DATA BERSAMA NARASIO DATA

100% JOB GUARANTEE

- International Certification
- Career Support
- 1 on 1 Mentoring
- Internship Program

Narasio DATA

Programs

CERTIFIED
JUNIOR DATA SCIENTIST

CERTIFIED
DATA ENGINEER

CERTIFIED
SENIOR DATA SCIENTIST

DATA ANALYTICS
FOR ACCOUNTING

CERTIFIED
DATA ANALYST

DATA SCIENCE
FOR ACCOUNTING

Narasio DATA

CERTIFIED

SENIOR DATA SCIENTIST

Program Certified Senior Data Scientist akan memperkenalkan kamu pada ilmu data dan pemodelan untuk Machine Learning melalui bahasa pemrograman Python. Program ini dirancang untuk kamu yang telah menguasai ilmu dasar data science dan machine learning.

Kamu akan belajar:

- Konsep aljabar linear dan PCA
- Teknik web scraping
- Teknik forecasting data terstruktur
- Proses data cleaning dan pengolahan data teks (tidak terstruktur)
- Teknik web scraping

CERTIFIED
SENIOR DATA SCIENTIST

Narasio DATA

Tools yang digunakan:

Python

Poin - Poin Pembelajaran:

1 Aljabar Linear dan PCA

- ★ **Konsep Aljabar Linier**
  - Matrix properties
  - Vector properties
  - Operation matrix and vector
- ★ **Konsep Principal Component Analysis (PCA)**
  - Kriteria eigenvalues dan eigenvectors
  - Teknik mereduksi variabel atau banyak features

2 Web Scraping

- ★ Dasar-dasar web scraping
- ★ Scraping dengan BeautifulSoup
- ★ Memahami pemilihan HTML
- ★ Studi kasus scraping dengan Python

3 Time Series Forecasting

- ★ Konsep dasar time series
- ★ Konsep exponential smoothing dan hands-on
- ★ Konsep stasioner dan autokorelasi
- ★ Konsep ARIMA dan hands-on
- ★ Studi kasus time series dengan Python

4 Sistem Rekomendasi

- ★ Konsep dasar sistem rekomendasi dan aplikasinya
- ★ Konsep sparsity dan similarity matrix serta hands-on
- ★ Implementasi sistem rekomendasi dengan beberapa jenis pendekatan menggunakan Python

5 NLP

- ★ Konsep pre-processing teks
- ★ Embedding dan modeling data teks
- ★ Studi kasus dengan Python

6 Deploy Machine Learning

- ★ Deploy dengan flask
- ★ Penggunaan HTML untuk deployment
- ★ Integrasi dan otomisasi model machine learning

Prasyarat: Lulus kelas Junior Data Scientist / Lulus Assessment

Register Now

https://bit.ly/RegisICDP2022

Further Information

CP: +62 813-1437-9570 (WA)

course.narasiodata.com/icdp/

Metode Pembayaran

Cash via Digital Payment

Credit Card

E-Wallet

Bank Transfer
(Virtual Account)

Others

Cicilan

Danacita

http://danacita.co.id/s/narasio-daftar

6-12 Bulan

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

Drop your questions here!

Narasio DATA

NARASIO DATAFEST
WORKSHOP

GETTING INSIGHTS FROM
YOUR TEXT DATA

GETTING INSIGHTS FROM YOUR TEXT DATA

https://bit.ly/NarasioDatafestFF11

Narasio Data

@narasiodata