Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

PERTEMUAN KE: 12

Pengolahan Bahasa Alami (NLP)

KULIAH

OLEH: NURJOKO

[12-09-2023] | [KODE MK: SSD23201 | MATAKULIAH: PENGANTAR SAINS DATA | SKS: 2 | VERSI: 01

Learning Objectives

- Mampu mengembangkan model dan algoritma untuk memahami bahasa manusia, termasuk struktur kalimat, makna kata, dan konteks linguistik.
- Mampu membangun teknik untuk membagi teks menjadi unit-unit yang lebih kecil (tokenisasi) dan menghilangkan elemen yang kurang penting seperti stop words.
- Mampu membangun teknik untuk membagi teks menjadi unit-unit yang lebih kecil (tokenisasi) dan menghilangkan elemen yang kurang penting seperti stop words.
- Mampu mengembangkan model yang memungkinkan komputer memahami dan menginterpretasikan makna dalam konteks bahasa manusia.
- Mampu menerapkan teknik untuk menentukan sentimen atau perasaan yang terkandung dalam teks, seperti positif, negatif, atau netral.
- Mampu membangun sistem yang mampu menerjemahkan teks dari satu bahasa ke bahasa lain dengan akurasi tinggi.

Konsep Dasar NLP

- Natural Language Processing (NLP) adalah cabang dari artificial intelligent (AI) dan komputasional linguistik yang mengfokuskan pada interaksi antara komputer dan bahasa alami manusia (Kao dan Poteet, 2007).
- Sebuah natural language system harus memperhatikan pengetahuan terhadap bahasa itu sendiri, baik dari segi kata yang digunakan, bagaimana kata-kata tersebut digabung untuk menghasilkan suatu kalimat dan sebagainya.
- Pengolahan Bahasa alami (NLP) adalah pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.

Konsep Dasar NLP - Lanjut

Menurut Rich dan Knight (2006) secara singkat, pengolahan bahasa alami mengenal beberapa tingkat pengolahan, yaitu :

1. Fonetik dan fonologi, berhubungan dengan suara yang menghasilkan kata yang dapat dikenali. Bidang ini menjadi penting dalam proses aplikasi yang memaknai metode speech based system
2. Morfologi, merupakan pengetahuan tentang kata dan bentuknya dimanfaatkan untuk membedakan satu kata dengan kata lainnya. Pada tingkat ini juga dapat dipisahkan antara kata dan elemen lain seperti tanda baca.
3. Sintaksis, merupakan pemahaman tentang urutan kata dalam pembentukan kalimat dan hubungan antar kata tersebut dalam proses perubahan bentuk dari kalimat menjadi bentuk yang sistematis. Meliputi proses pengaturan tata letak suatu kata dalam kalimat akan membentuk kalimat yang dapat dikenali.
4. Semantik, merupakan pemetaan bentuk struktur sintaks dengan memanfaatkan tiap kata ke dalam bentuk yang lebih mendasar dan tidak tergantung struktur kalimat. Semantik mempelajari arti suatu kata, dan bagaimana dari arti kata-arti kata tersebut membentuk suatu arti kalimat yang utuh. Dalam tingkatan ini belum tercakup konteks dari kalimat tersebut.
5. Pragmatik, Pengetahuan pada tingkatan pragmatik berkaitan dengan masing-masing konteks yang berbeda tergantung pada situasi dan tujuan pembuatan sistem
6. Discourge knowledge, melakukan pengenalan apakah suatu kalimat yang sudah dibaca dan dikenali sebelumnya dapat mempengaruhi arti dari kalimat selanjutnya
7. Word knowledge, mencakup arti sebuah kata secara umum dan apakah ada arti khusus bagi suatu kata dalam suatu percakapan dalam konteks tertentu.

Pembagian NLP

Masalah pemrosesan bahasa alami dibagi menjadi dua bagian besar, yaitu :

1. **Pemrosesan Naskah Tertulis**

   menggunkan pengetahuan tentang leksikal, sintax, dan semantik

2. **Pemrosesan Bahasa Lisan**

   menggunakan semua pengetahuan dari pemrosesan naskah tertulis ditambah pengetahuan tentang phonology.

Konsep Dasar NLP - Lanjut

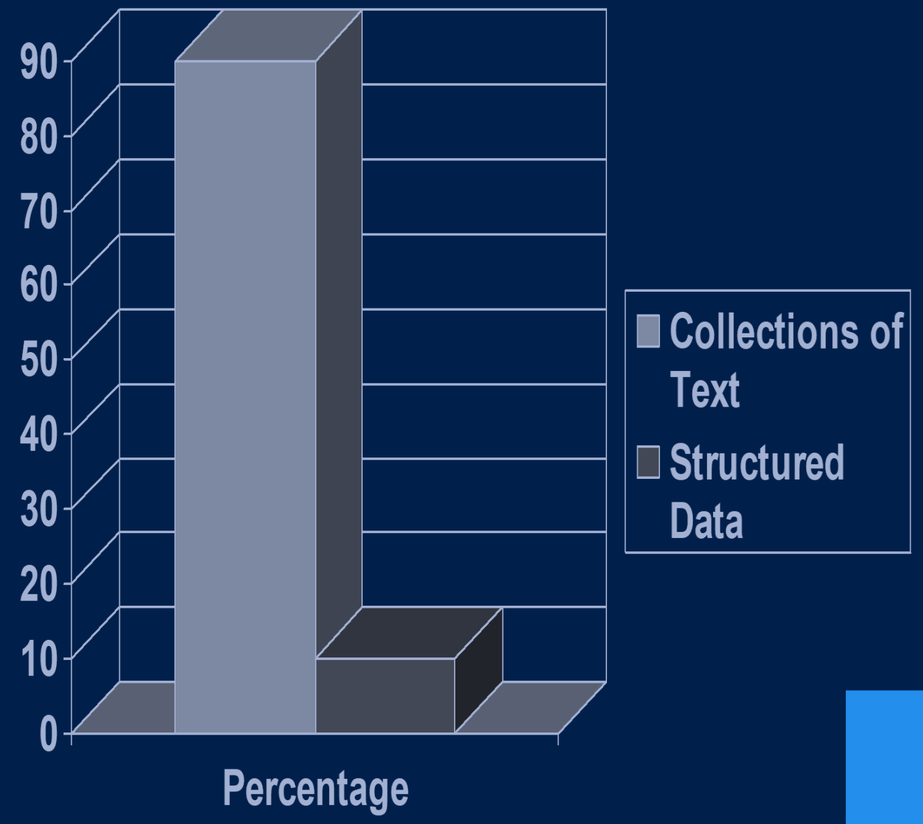
Bidang-bidang utama pekerjaan NLP antara lain :

1. Machine Translation
2. name entity Recognition (NER)
3. Optical Character Recognition
4. part-of-speech Tagging (POS-Tag)
5. sentiment Analysis
6. speech recognition
7. information retrieval

Text Pre-Processing

Latar Belakang

- Dokumen-dokumen yang ada kebanyakan **tidak memiliki struktur yang pasti** sehingga informasi di dalamnya tidak bisa diekstrak secara langsung.
- Tidak semua kata **mencerminkan** makna/isi yang terkandung dalam sebuah dokumen.



Latar Belakang

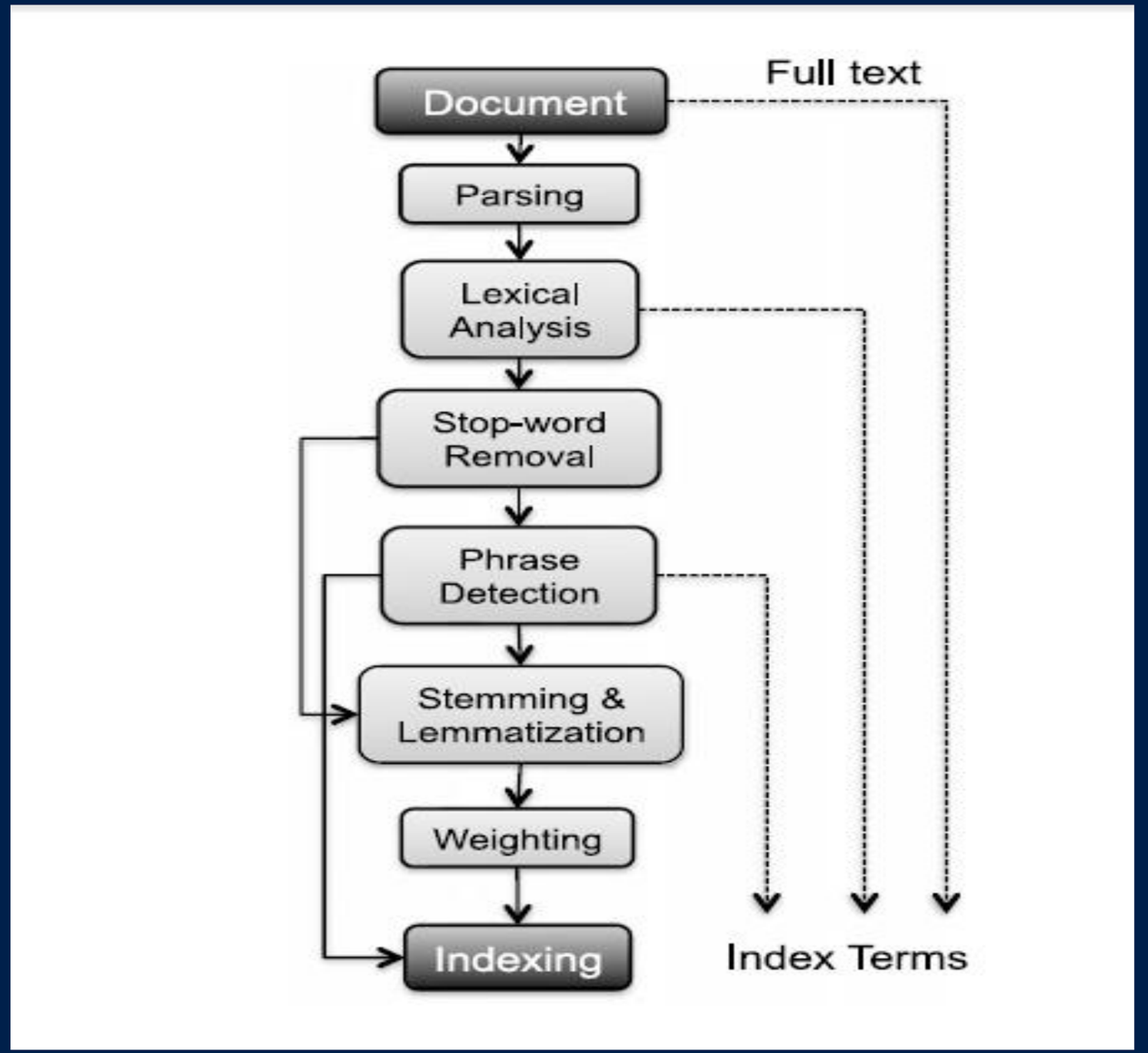
- Preprocessing diperlukan untuk memilih kata yang akan digunakan sebagai **indeks**
- **Indeks** ini adalah kata-kata yang **mewakili dokumen** yang nantinya digunakan untuk membuat pemodelan untuk Information Retrieval maupun aplikasi teks mining lain.

Definisi

- Definisi Pemrosesan Teks (Text Preprocessing) adalah suatu proses **pengubahan** bentuk data yang **belum terstruktur** menjadi data yang **terstruktur** sesuai dengan kebutuhan, untuk proses mining yang lebih lanjut (sentiment analysis, peringkasan, clustering dokumen, etc.).
- **Preprocessing** adalah **merubah** teks menjadi term index
- **Tujuan:** menghasilkan sebuah set **term index** yang bisa **mewakili** dokumen

Langkah-langkah Text Pre-processing

Langkah-langkah umum dalam Text Pre-processing



Langkah 2. Parsing

- ***Parsing Dokumen*** berurusan dengan **pengenalan dan "pemecahan"** struktur dokumen menjadi komponen-komponen terpisah. Pada langkah preprocessing ini, kita menentukan mana yang dijadikan **satu unit dokumen**;
- Contoh. Buku dengan **100 halaman** bisa dipisah menjadi **100 dokumen**; masing-masing halaman menjadi 1 dokumen
- **Satu *tweet*** bisa dijadikan sebagai **1 dokumen**.
- Begitu juga dengan sebuah komentar pada forum atau review produk.

Langkah 2 : Lexical Analysis

- Lebih populer disebut Lexing atau **Tokenization / Tokenisasi**
- Tokenisasi adalah proses **pemotongan** string input berdasarkan tiap kata penyusunnya.
- Pada prinsipnya proses ini adalah **memisahkan** setiap kata yang menyusun suatu dokumen.

Langkah 2 : Lexical Analysis

- Pada proses ini dilakukan **penghilangan** angka, tanda baca dan karakter selain huruf alfabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (delimiter) dan tidak memiliki pengaruh terhadap pemrosesan teks.
- Pada tahapan ini juga dilakukan proses ***case folding***, dimana semua huruf diubah menjadi huruf kecil.
- Cleaning adalah proses **membersihkan** dokumen dari komponen-komponen yang tidak memiliki hubungan dengan informasi yang ada pada dokumen, seperti tag html, link, dan script, dsb.

Tokens, Types, and Terms

- **Text**: “apakah singo dan boyo bermain bola di depan rumah boyo?”
- **Token** adalah kata-kata yang dipisah-pisah dari teks aslinya tanpa mempertimbangkan adanya duplikasi
- **Token**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”

Tokens, Types, and Terms

- **Text:** “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Type** adalah token yang memperhatikan adanya duplikasi kata. Ketika ada duplikasi hanya dituliskan sekali saja.
- **Type**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”

Tokens, Types, and Terms

- **Text:** “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Term** adalah type yang sudah dinormalisasi (dilakukan stemming, filtering, dsb)
- **Term** : “culo”, “boyo”, “main”, “bola”, “depan”, “rumah”

1. **Text:** “apakah culo dan boyo bermain bola di depan rumah boyo?”
2. **Token**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”
3. **Type**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”
4. **Term**: “culo”, “boyo”, “main”, “bola”, “depan”, “rumah”

Contoh Tokenisasi

Teks Bahasa	Namanya adalah Santiago. Santiago sudah memutuskan untuk mencari sang alkemis.
Tokens	namanya
	adalah
	santiago
	santiago
	sudah
	memutuskan
	untuk
	mencari
	sang
	alkemis

Langkah 3 : Stopword Removal

- Disebut juga **Filtering**
- **Filtering** adalah tahap **pemilihan** kata-kata penting dari hasil token, yaitu kata-kata apa saja yang akan digunakan untuk mewakili dokumen.
- **Metode:** menggunakan algoritma stopword. **Stoplist** atau **stopword** adalah **kata-kata yang tidak deskriptif (tidak *penting*)** yang dapat dibuang dengan pendekatan *bag-of-words*.
- **Contoh:** stopwords adalah ada, adalah adanya, adapun, agak agaknya, agar, dll

Stopword Removal : Metode

- Algoritma **wordlist**
- **Wordlist** adalah **kata-kata yang deskriptif (*penting*)** yang harus disimpan dan tidak dibuang dengan pendekatan *bag-of-words*.
- Kita memiliki database kumpulan **kata-kata yang deskriptif (*penting*)**, kemudian kalau hasil tokenisasi itu ada yang merupakan kata penting dalam database tersebut, maka hasil tokenisasi itu disimpan.

Text:

Namanya adalah Santiago. Santiago sudah memutuskan untuk mencari sang alkemis.

Contoh wordlist adalah:

santiago, namanya, mencari, memutuskan, alkemis, dst.

Hasil Token	Hasil Filtering
namanya	namanya
adalah	-
santiago	santiago
Santiago	santiago
sudah	-
memutuskan	memutuskan
untuk	-
mencari	mencari
sang	-
alkemis	alkemis

Using Stop Words or Not?

- Kebanyakan aplikasi text mining ataupun IR **bisa ditingkatkan** performanya dengan penghilangan stopword.
- Akan tetapi, secara umum Web search engines seperti ***google*** sebenarnya **tidak menghilangkan** stop word, karena algoritma yang mereka gunakan berhasil memanfaatkan stopword dengan baik.

Langkah 4 : *Phrase Detection*

- Langkah ini bisa menangkap informasi dalam teks **melebihi** kemampuan dari metode tokenisasi / bag-of-word murni.
- Pada langkah ini tidak hanya dilakukan tokenisasi per kata, namun juga mendeteksi adanya 2 kata atau lebih yang menjadi **frase**.
- Contoh, dari dokumen ini : *"search engines are the most visible information retrieval applications"*
- Terdapat dua buah **frase**, yaitu *"search engines"* dan *"information retrieval"*.

Langkah 4 : *Phrase Detection*

- *Phrase detection* bisa dilakukan dengan beberapa cara : menggunakan **rule/aturan** (misal dengan menganggap dua kata yang sering muncul berurutan sebagai frase), bisa dengan ***syntactic analysis***, and **kombinasi** keduanya.
- Metode umum yang diguakan adalah **penggunaan thesauri** untuk mendeteksi adanya frase.
- Contoh : Pada thesauri tersebut terdapat **daftar frase-fase** dalam bahasa tertentu, kemudia kita bandingkan kata-kata dalam teks apakah mengandung frase-frase dalam thesauri tersebut atau tidak.

Langkah 5 : Stemming

- **Stemming** adalah proses pengubahan **bentuk kata** menjadi **kata dasar** atau tahap mencari root kata dari tiap kata hasil filtering.
- Dengan dilakukanya proses stemming setiap kata berimbuhan akan berubah menjadi kata dasar, dengan demikian dapat lebih **mengoptimalkan** proses **teks mining**.

Langkah 5 : Stemming

Hasil Token	Hasil Filtering	Hasil Stemming
namanya	namanya	nama
adalah	-	-
santiago	santiago	santiago
santiago	santiago	santiago
sudah	-	-
memutuskan	memutuskan	putus
untuk	-	-
mencari	mencari	cari
sang	-	-
alkemis	alkemis	alkemis

Langkah 5 : Stemming

- Implementasi proses **stemming** sangat beragam , tergantung dengan **bahasa** dari dokumen.
- Beberapa metode untuk Stemming :
  - Porter Stemmer (English & Indonesia)
  - Stemming Arifin-Setiono (Indonesia)
  - Stemming Nazief-Adriani (Indonesia)
  - Khoja (Arabic)

Stemming : Metode

- Algorithmic: Membuat sebuah **algoritma yang mendeteksi imbuhan**.
- Jika ada awalan atau akhiran yang seperti imbuhan, maka akan dibuang.

Porter's algorithm

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Stemming : Metode

- Kelebihan : relatif **cepat**
- Kekurangan : beberapa algoritma **terkadang salah mendeteksi imbuhan**, sehingga ada beberapa kata yang bukan imbuhan tapi dihilangkan
- Contoh : makan -> mak; **an** dideteksi sebagai akhiran sehingga dibuang.

Stemming : Metode

- Metode Lemmatization
- Lemmatization : Stemming berdasarkan **kamus**
- Menggunakan *vocabulary* dan *morphological analysis* dari kata untuk menghilangkan imbuhan dan dikembalikan ke bentuk dasar dari kata.
- Stemming ini bagus untuk kata-kata yang mengalami **perubahan tidak beraturan** (terutama dalam english)
- Contoh : “see” -> “see”, “saw”, atau “seen”
- Jika ada kata “see”, “saw”, atau “seen”, bisa dikembalikan ke bentuk aslinya yaitu “see”

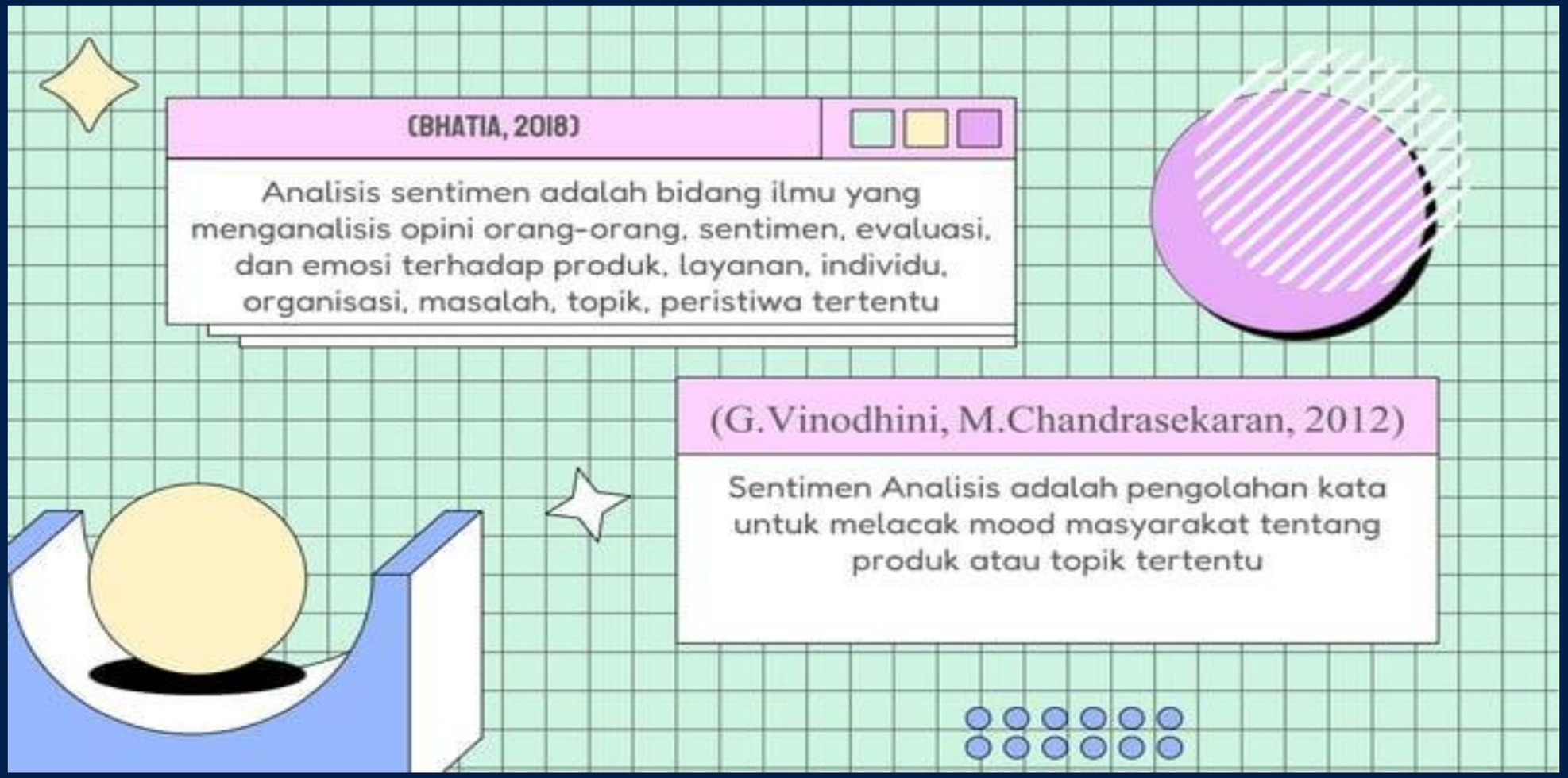
Stemming : Metode

- Algoritma Porter Stemming merupakan algoritma yang paling populer. Ditemukan oleh Martin Porter pada tahun 1980.
- Mekanisme algoritma tersebut dalam mencari kata dasar suatu kata berimbuhan, yaitu dengan membuang imbuhan–imbuhan (atau lebih tepatnya akhiran pada kata–kata bahasa Inggris karena dalam bahasa Inggris tidak mengenal awalan).

Stemming : Metode

Hasil Token	Hasil Filtering	Hasil Stemming	Type	Term
namanya	namanya	nama	nama	nama
adalah	-	-	-	-
santiago	santiago	santiago	santiago	santiago
santiago	santiago	santiago	-	-
sudah	-	-	-	-
memutuskan	memutuskan	putus	putus	putus
untuk	-	-	-	-
mencari	mencari	cari	cari	cari
sang	-	-	-	-
alkemis	alkemis	alkemis	alkemis	alkemis

Analisis Sentimen



Analisis Sentimen

KEBUTUHAN ANALISIS SENTIMEN

Meningkatnya penggunaan social media di masyarakat, berdampak pada **bertambahnya peran berbagi infromasi di ruang public**, yang selanjutnya menyebabkan **berkembangnya opini publik**.
Kemudian hal tersebut dimanfaatkan menggunakan metode tertentu untuk tujuan **pengawasan terhadap suatu objek**.

Social Media

Berbagi Informasi

Opini Publik

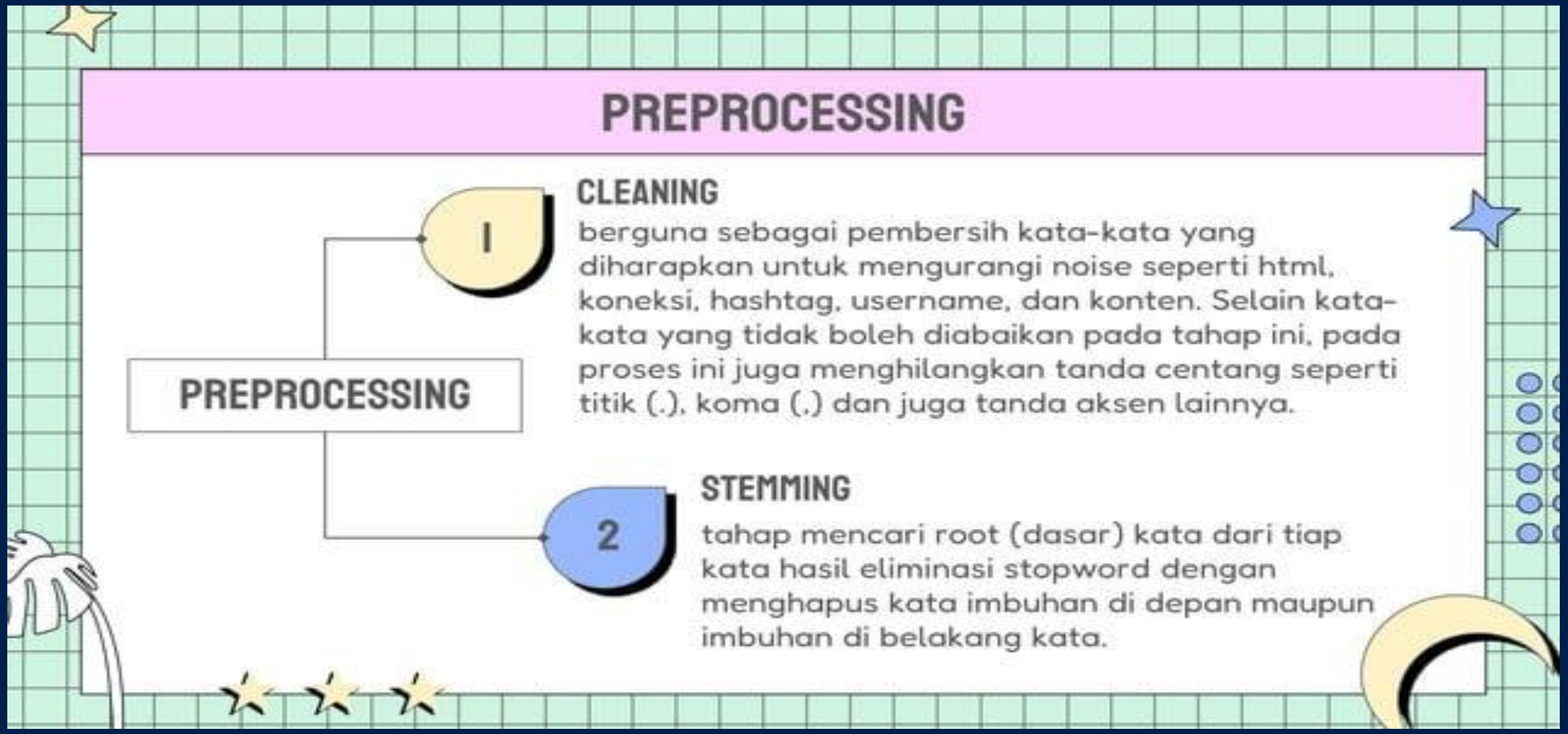
Peran Pengawasan

Analisis Sentimen

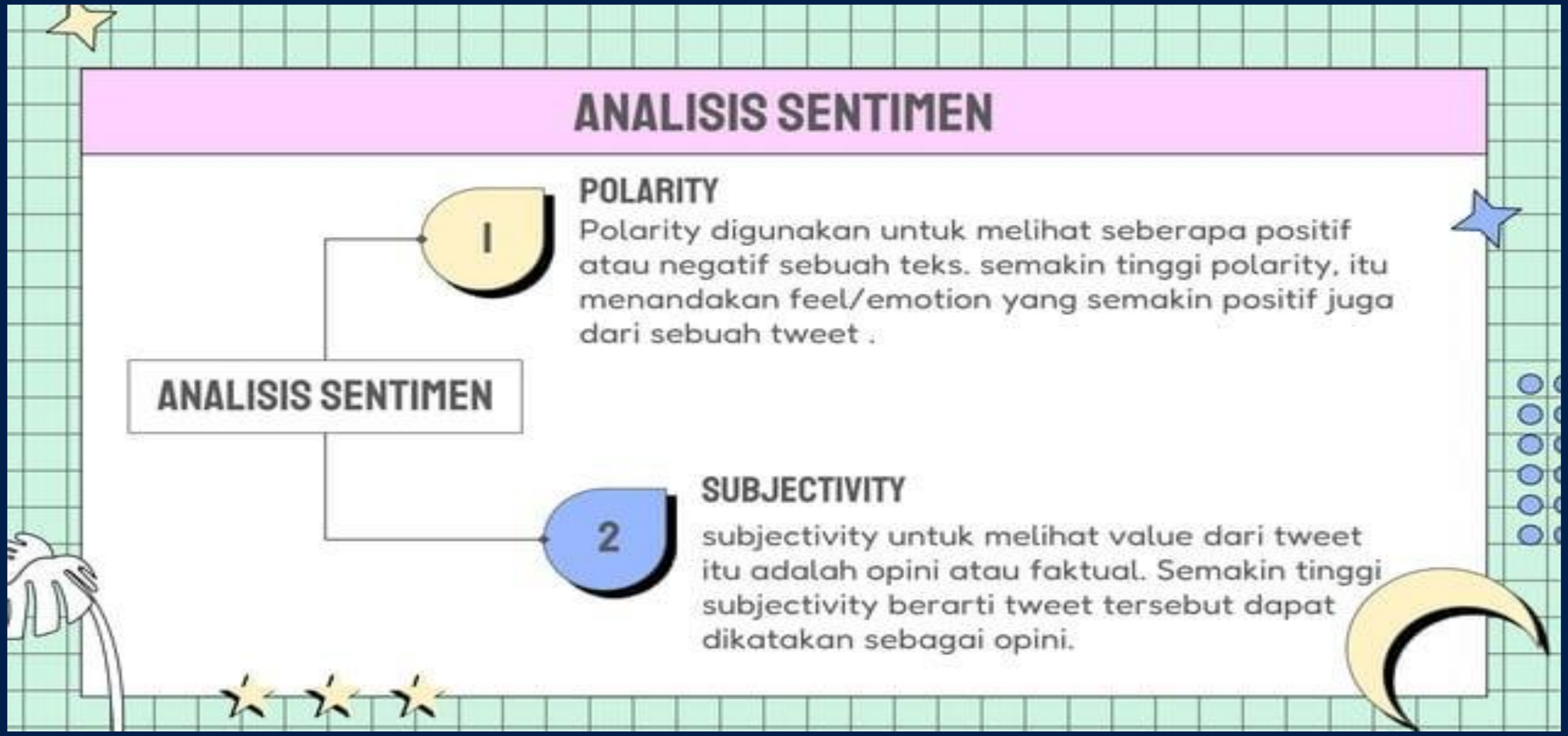
Analisis Sentimen

Analisis Sentimen

Analisis Sentimen



Analisis Sentimen



Penerjemahan Mesin

Penerjemahan Mesin

Penerjemahan Mesin

Latihan : Tentukan hasil Tokenisasi, Filtering dan Stemming setiap dokumen tersebut

Dokumen (Doc)	Isi (Content)
Doc 1	elearning di PTIIK diatas jam 6 malam kok selalu gak bisa dibuka ya?
Doc 2	ub tidak punya lahan parkir yang layak. Dan jalanan terlalu ramai karena di buka untuk umum. Seperti jalan tol saja. Brawijaya oh brawijaya
Doc 3	Kelas Arsitektur dan Organisasi Komputer penuh, apakah tidak dibuka kelas lagi. Rugi kalo saya bisa ngambil 24 SKS tapi baru 18 SKS yg terpenuhi
Doc 4	Informasi tata cara daftar ulang bagi mahasiswa baru PTIIK kurang jelas. Sehingga ketika tanggal terakhir syarat penyerahan berkas daftar ulang, banyak mahasiswa baru yang tidak membawa salah satu syarat daftar ulangnya.

CONCLUSION

Fill in .....

REFERENCES

Fill in IEEE Style

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”