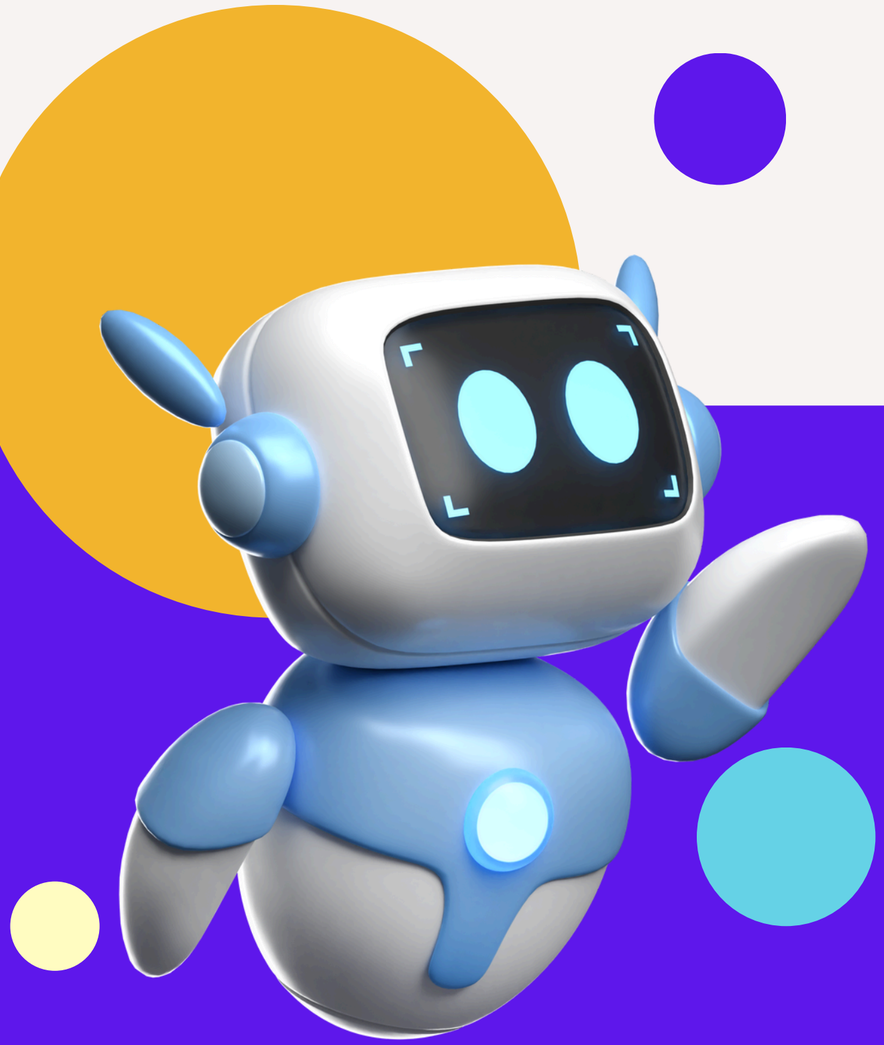


IFB 304 - Sistem Pakar & Bahasa Alamiah

Text Preprocessing

Chalifa Chazar

Course Outline

1. Dasar-dasar sistem pakar dan karakteristik utama sistem pakar.
2. Knowledge Engineering
3. Model Inferensi
4. Faktor Ketidakpastian (Uncertainty factor) dan kepastian (Certainty facto
5. Perangkat Lunak (Tools) untuk sistem pakar dan proyek akhir sistem pakar
6. Pemrosessan sintaks dan Interpretasi atau penterjemahan semantik
7. Pembuatan perangkat lunak Bahasa Alamiah untuk proyek akhir bahasa alamiah

Ruang Lingkup NLP

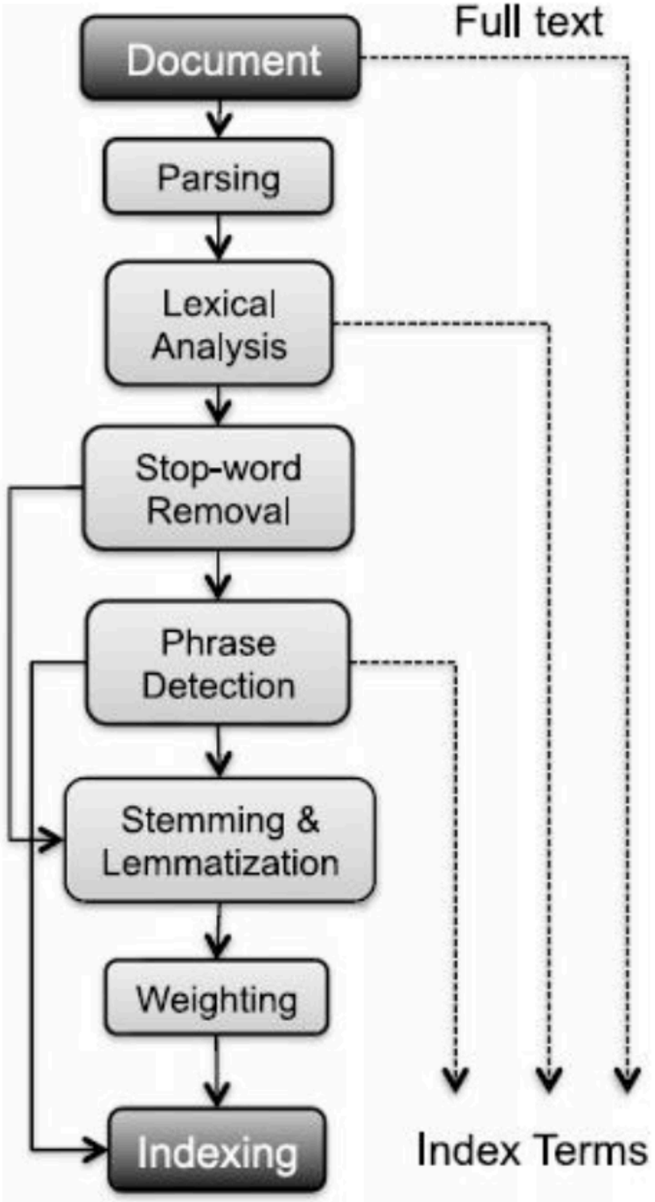
- Natural Language Processing (NLP) → pemrosesan secara simbolik terhadap bahasa tulisan
- Text to Speech (TTS) → pemrosesan text (bahasa tulisan) menjad ucapan (bahasa lisan)
- Speech Recogniion (SR) → pemrosesan ucapan (bahasa lisan) menjadi text (bahasa tulisan)

Text Preprocessing

- Merupakan langkah penting dalam NLP
- Proses ini mencakup serangkaian teknik yang digunakan untuk membersihkan dan mempersiapkan data text sebelum digunakan dalam model machine learning atau analisis teks
- Text preprocessing yang buruk menyebabkan kualitas data yang buruk dan hasil analisis yang tidak akurat

Tujuan Text Preprocessing

- **Meningkatkan Kualitas Data:** Memperbaiki kualitas teks dengan menghapus informasi yang tidak relevan
- **Mengurangi Dimensi Data:** Mengurangi kompleksitas data teks dengan menghapus informasi yang tidak penting
- **Meningkatkan Akurasi Model:** Memberikan data teks yang lebih bersih dan terstruktur agar model machine learning dapat bekerja dengan lebih efisien dan efektif.



Langkah-langkah dalam text preprocessing

Parsing

- Tulisan dalam sebuah dokumen bisa jadi terdiri dari berbagai macam bahasa, character sets, dan format
- Parsing Dokumen berurusan dengan pengenalan dan “pemecahan” struktur dokumen menjadi komponen-komponen terpisah.
- Contoh:
  - Satu tweet bisa dijadikan sebagai 1 dokumen
  - Email dengan 4 lampiran bisa dipisah menjadi 5 dokumen

Tokenization atau Lexical Analysis

- Tokenisasi adalah proses memecah teks menjadi unit-unit kecil yang disebut "tokens" (kata, karakter, atau kalimat)

- 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 dilakukan juga case folding dan juga cleaning

  case folding → merubah semua huruf menjadi huruf kecil, seperti **"Apple"** dan **"apple"** dianggap sama.

  cleaning → membersihkan dokumen dari komponen-komponen yang tidak memiliki hubungan dengan informasi yang ada (exp. link, tag html, script, dll)

Hasil Tokenization

Text	Hasil Token
“apakah culo dan boyo bermain bola di depan rumah boyo?”	“culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”

Hasil Tokenization

- Type adalah token yang memperhatikan adanya duplikasi kata. Ketika ada duplikasi hanya dituliskan sekali saja

Text	Hasil Type
“apakah culo dan boyo bermain bola di depan rumah boyo?”	“culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”

- Term adalah type yang sudah dinormalisasi (dilakukan stemming, filtering, dsb)

Text	Hasil Term
“apakah culo dan boyo bermain bola di depan rumah boyo?”	“culo”, “boyo”, “main”, “bola”, “depan”, “rumah”

Stopword Removal

- 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
- Algoritma yang digunakan pada Filtering yaitu stoplist dan wordlist

  Stoplist atau stopword adalah kata-kata yang tidak deskriptif (tidak penting) yang dapat dibuang dengan pendekatan bag-of-words (BoW)

  Wordlist adalah kata-kata yang deskriptif (penting) yang harus disimpan dan tidak dibuang dengan pendekatan bag-of-words (BoW)

Stoplist atau Stopword

Hasil Token	Hasil Filtering
they	-
are	-
applied	applied
to	-
the	-
words	words
in	-
the	-
texts	text

Wordlist

Hasil Token	Hasil Filtering
they	-
are	-
applied	applied
to	-
the	-
words	words
in	-
the	-
texts	text

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:

  “search engines are the most visible information retrieval applications” .

  Terdapat dua buah frase, yaitu “search engines” dan “information retrieval”.

Phrase Detection

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

Stemming

- Stemming adalah proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari root kata dari tiap kata hasil filtering

- Proses stemming setiap kata berimbuhan akan berubah menjadi kata dasar, (dapat lebih mengoptimalkan proses teks mining)

Stemming

Hasil Token	Hasil Filtering	Hasil Stemming
they	-	-
are	-	-
applied	applied	apply
to	-	-
the	-	-
words	words	word
in	-	-
the	-	-
texts	text	text

Stemming

Hasil Token	Hasil Filtering	Hasil Stemming	Type	Term
they	-	-	-	-
are	-	-	-	-
applied	applied	apply	apply	apply
to	-	-	-	-
the	-	-	-	-
words	words	word	word	word
in	-	-	-	-
the	-	-	-	-
texts	text	text	text	text

Lemmatization

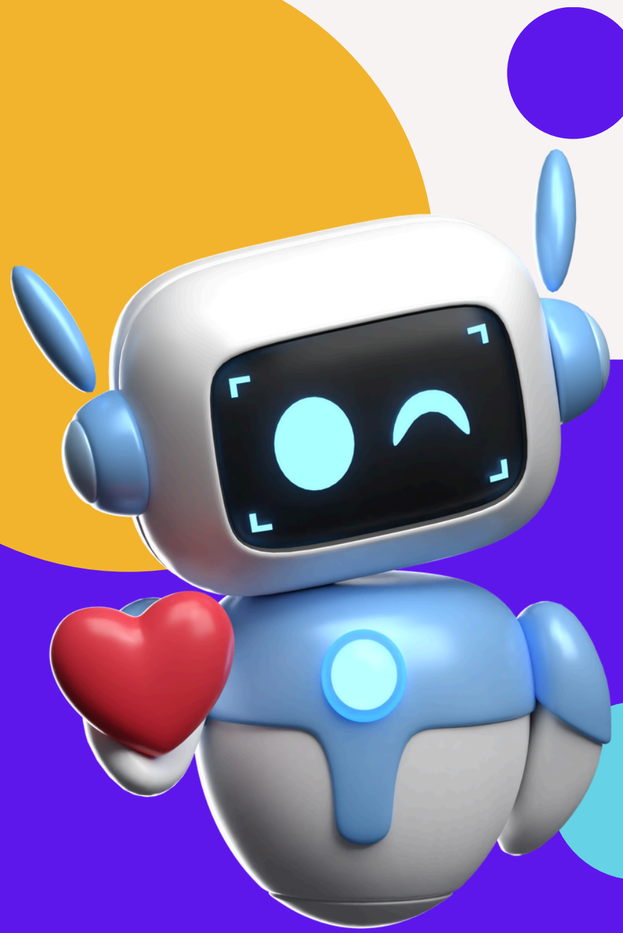
- Lemmatization mirip dengan stemming, tetapi lebih canggih karena mempertimbangkan konteks
- Lemmatization mengubah kata ke bentuk dasar yang valid secara linguistik (lemma)
- Contoh:

Input	Output
running	run
better	good
studies	study

Tugas 2 (Kelompok)

Buat sebuah tema aplikasi NLP (tiap kelompok berbeda-beda)

1. Latar belakang aplikasi
2. Jelaskan tahapan aplikasi
3. Implementasikan dalam bentuk aplikasi



See u next..

Thanxs!

chalifa.id