

Text Preprocessing

Indonesia AI



OBJECTIVE & OUTLINE

Proprietary document of Indonesia AI 2023



Text Preprocessing

Objektif: Memahami konsep dan ragam teknik Text Preprocessing beserta contohnya.

Outline:

1. Konsep Dasar Text Preprocessing
2. Ragam Teknik pada Text Preprocessing
3. Text Cleaning
4. Text Normalization
5. Text Tokenization

Konsep Dasar Text Preprocessing

DEFINISI

Proprietary document of Indonesia AI 2023

Text Preprocessing adalah **rangkaian tahapan** yang **dilakukan pada data teks** sebelum melakukan modeling maupun analisis lebih lanjut.

Text preprocessing bertujuan untuk **membersihkan, mengubah format,** atau **menyajikan teks agar lebih mudah dipahami** oleh algoritma yang akan digunakan.



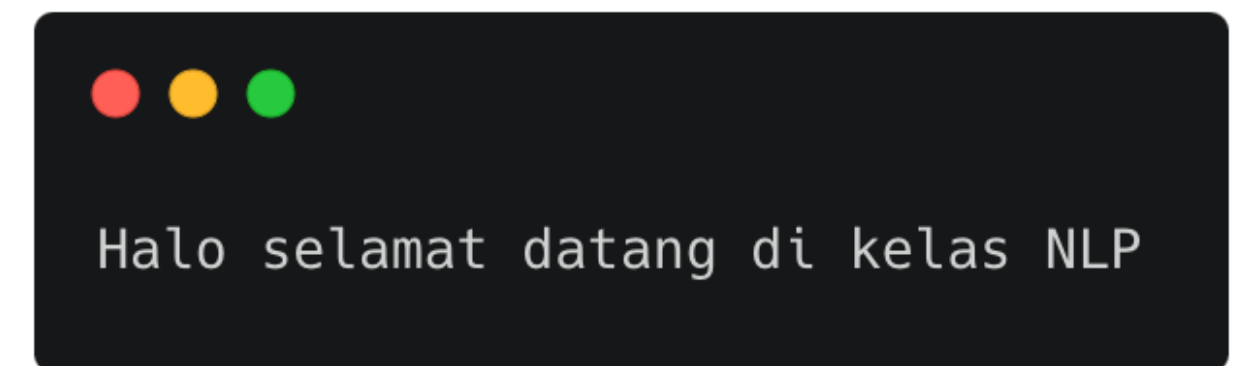
```
<p>  
<strong>Hello</strong> world!  
<a href="https://www.example.com">Click here</a>  
for more information.  
</p>
```

CONTOH LAIN RAW TEXT

Proprietary document of Indonesia AI 2023



**Text
Extraction**



Setelah berhasil mengekstraksi data teks dari tweet, data tersebut perlu diolah lagi.

TEXT EXTRACTION

Proprietary document of Indonesia AI 2023

```
<p>  
<strong>Hello</strong> world!  
<a href="https://www.example.com">Click here</a>  
for more information.  
</p>
```

Raw Text

Text
Extraction



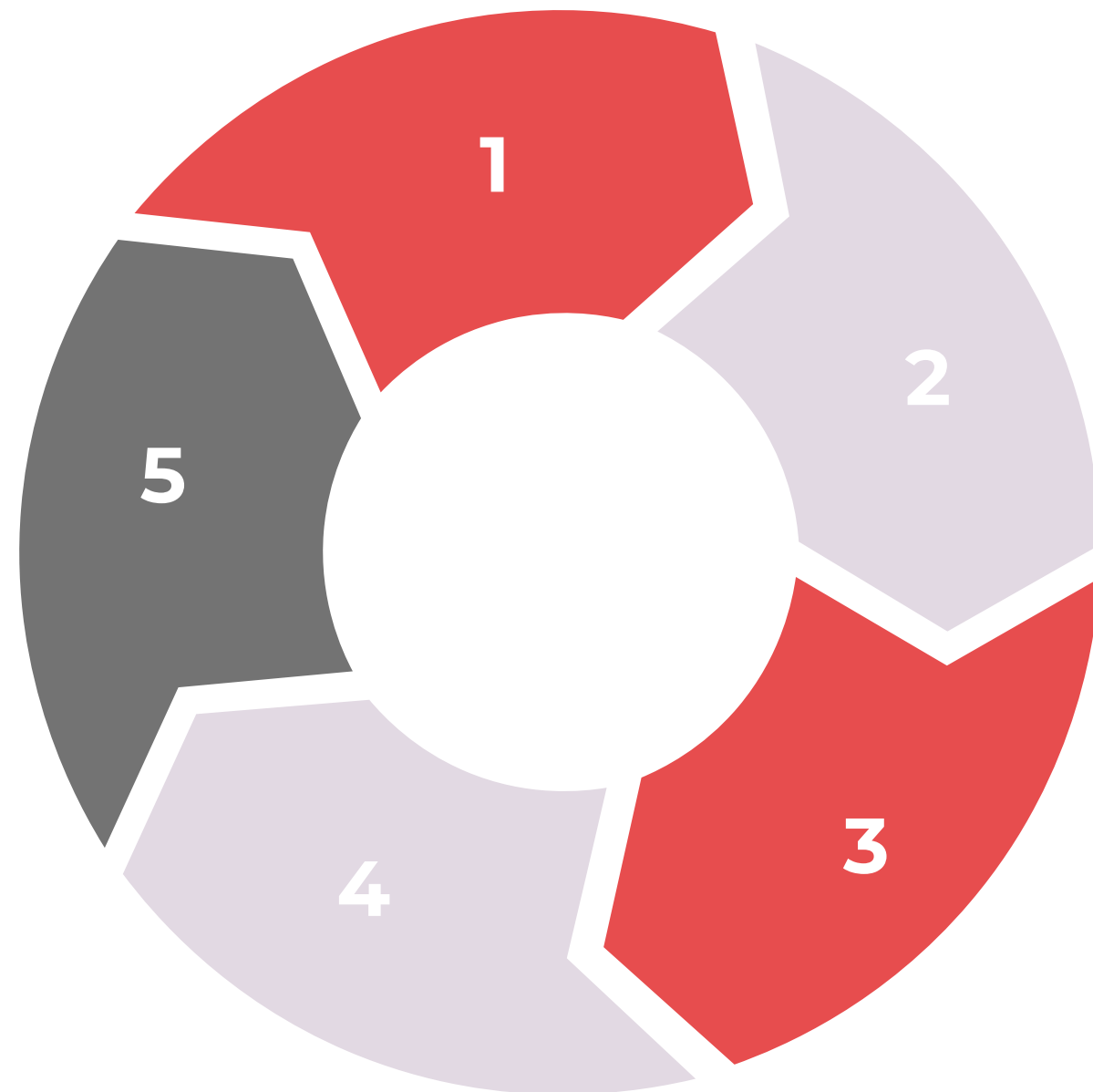
```
Hello world!  
Click here for more information.
```

Extracted Text

Setelah berhasil mengekstraksi data teks, kita perlu melakukan teks preprocessing agar data tersebut dapat dipahami oleh algoritma dengan lebih baik.

NLP LIFECYCLE

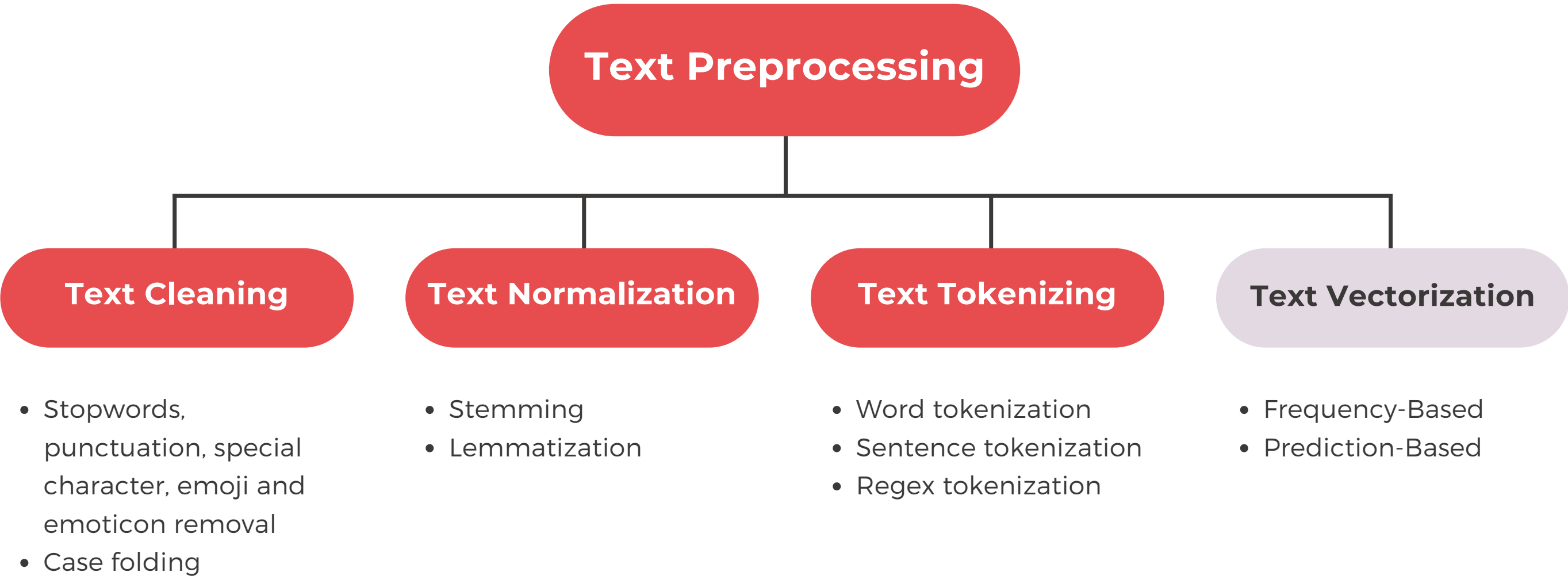
Proprietary document of Indonesia AI 2023



1. Data Collection & Text Extraction
2. **Text Preprocessing**
3. Data Modeling
4. Model Evaluation
5. Model Deployment & Maintenance

Kali ini kita akan mendalami tahapan text preprocessing, yuk kuy~

Ragam Teknik pada Text Preprocessing



Text cleaning, text normalization, dan text tokenizing memanfaatkan teknik regex secara langsung.

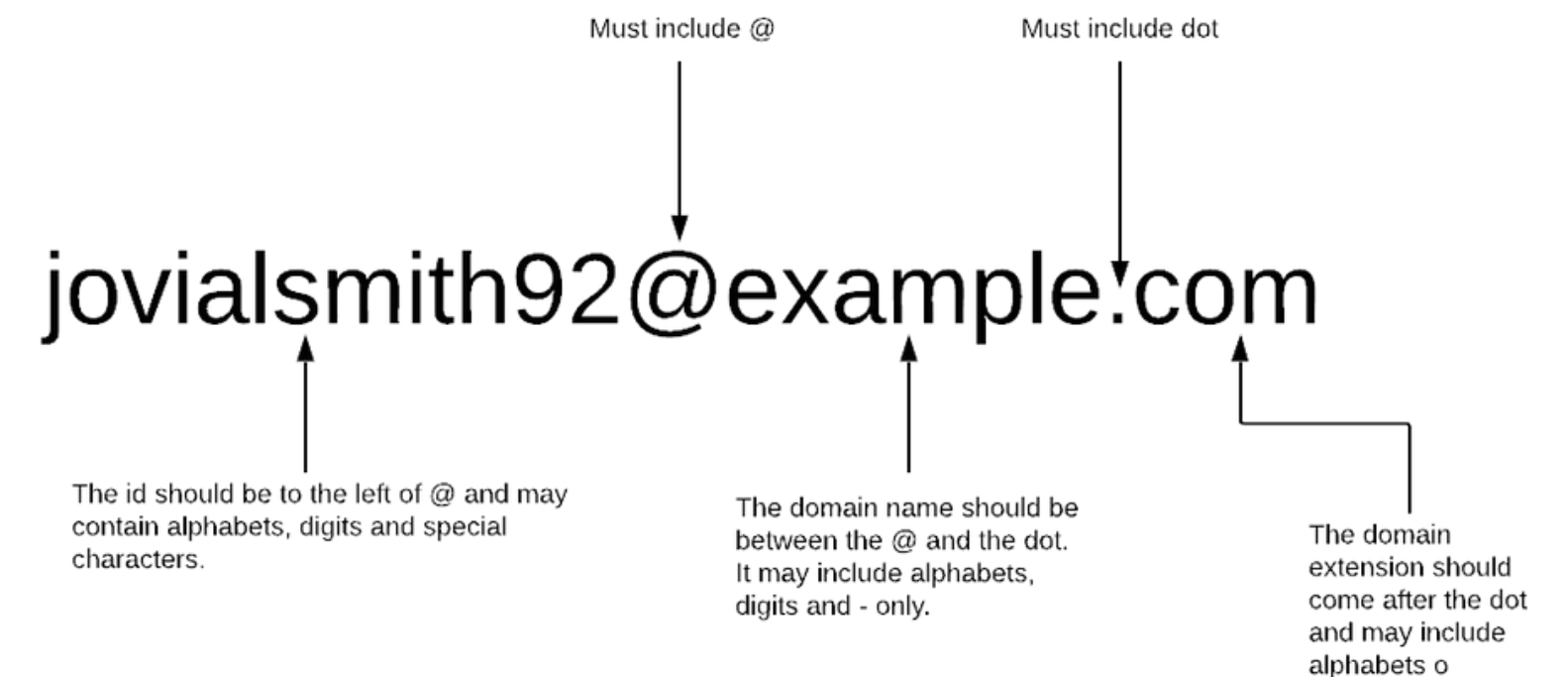
TEKNIK REGEX

Pada tahap text preprocessing, ada satu teknik yang sering diadopsi pada teknik text preprocessing yang lain, yaitu teknik regex.

Regex atau **Regular Expression** adalah pola yang menunjukkan urutan karakter. Teknik regex memanfaatkan pola yang sudah ditentukan untuk melakukan pencocokan teks.

Indonesia AI

Proprietary document of Indonesia AI 2023



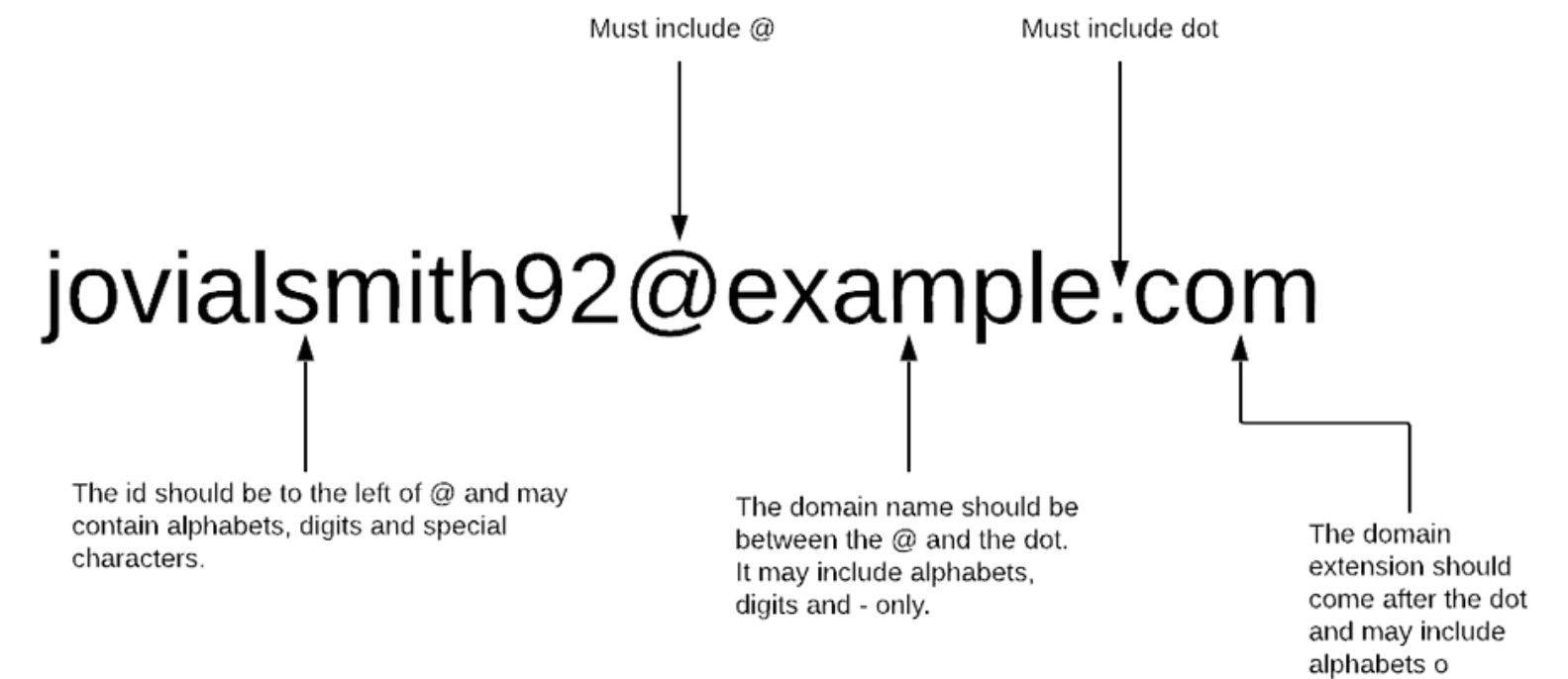
Gambar: r-bloggers

TEKNIK REGEX

Proprietary document of Indonesia AI 2023

Dalam regex, kita dapat menentukan **pola** yang harus dicocokkan, seperti urutan karakter tertentu, angka, atau karakter khusus.

Regex memungkinkan kita untuk melakukan preprocessing teks dengan lebih **fleksibel** dan efisien berdasarkan pola yang diinginkan.



Text Cleaning

TEXT CLEANING

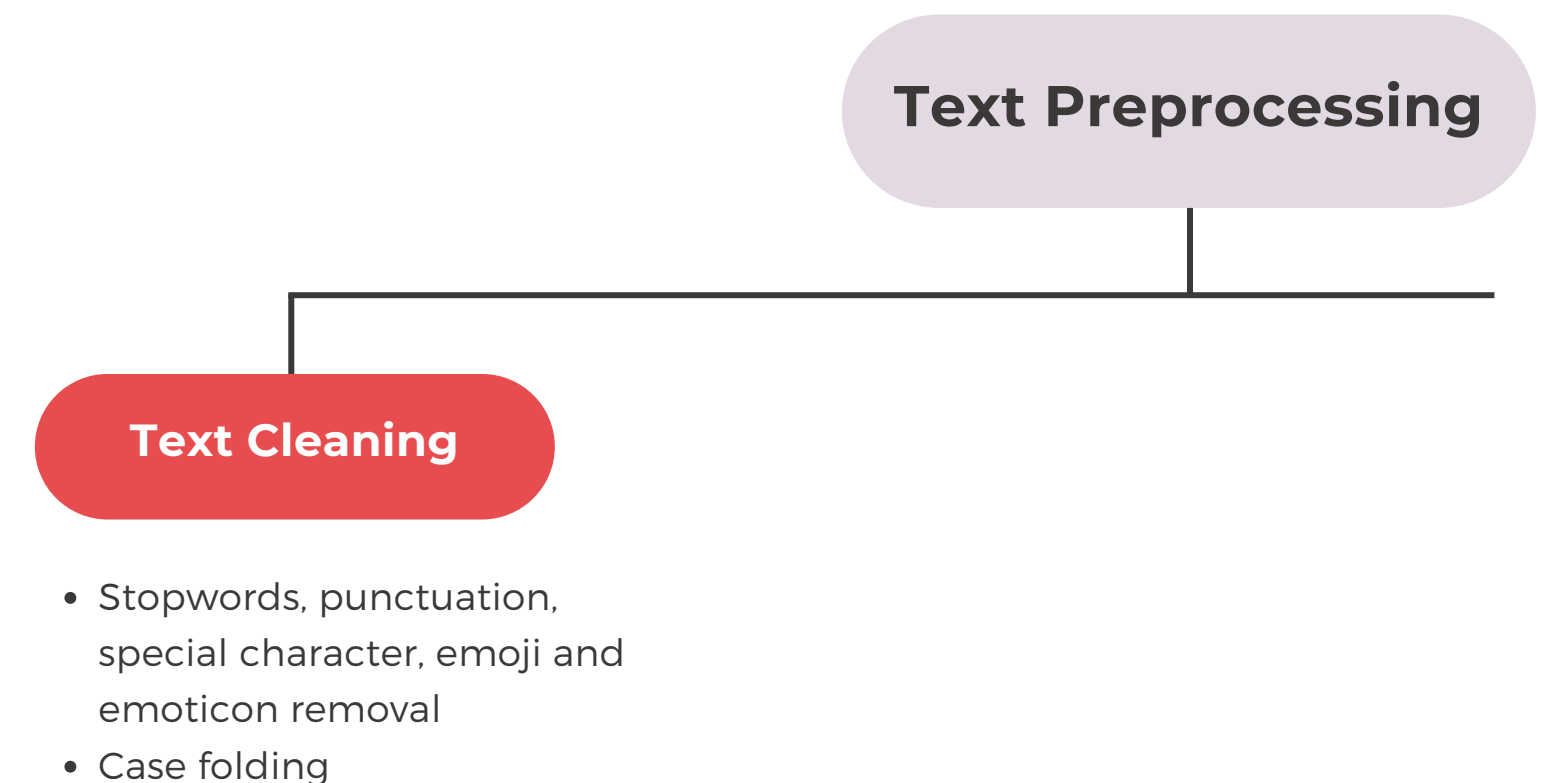
Proprietary document of Indonesia AI 2023

Text Cleaning adalah proses membersihkan teks dari karakter khusus, tanda baca, dan elemen yang tidak relevan atau mengganggu.

Umumnya text cleaning terdiri dari beberapa proses, yaitu:

- Menghilangkan stopwords,
- Menghilangkan punctuation,
- Menghilangkan special character,
- Case folding: lower case, upper case, title.

Indonesia AI



STOPWORDS (1)

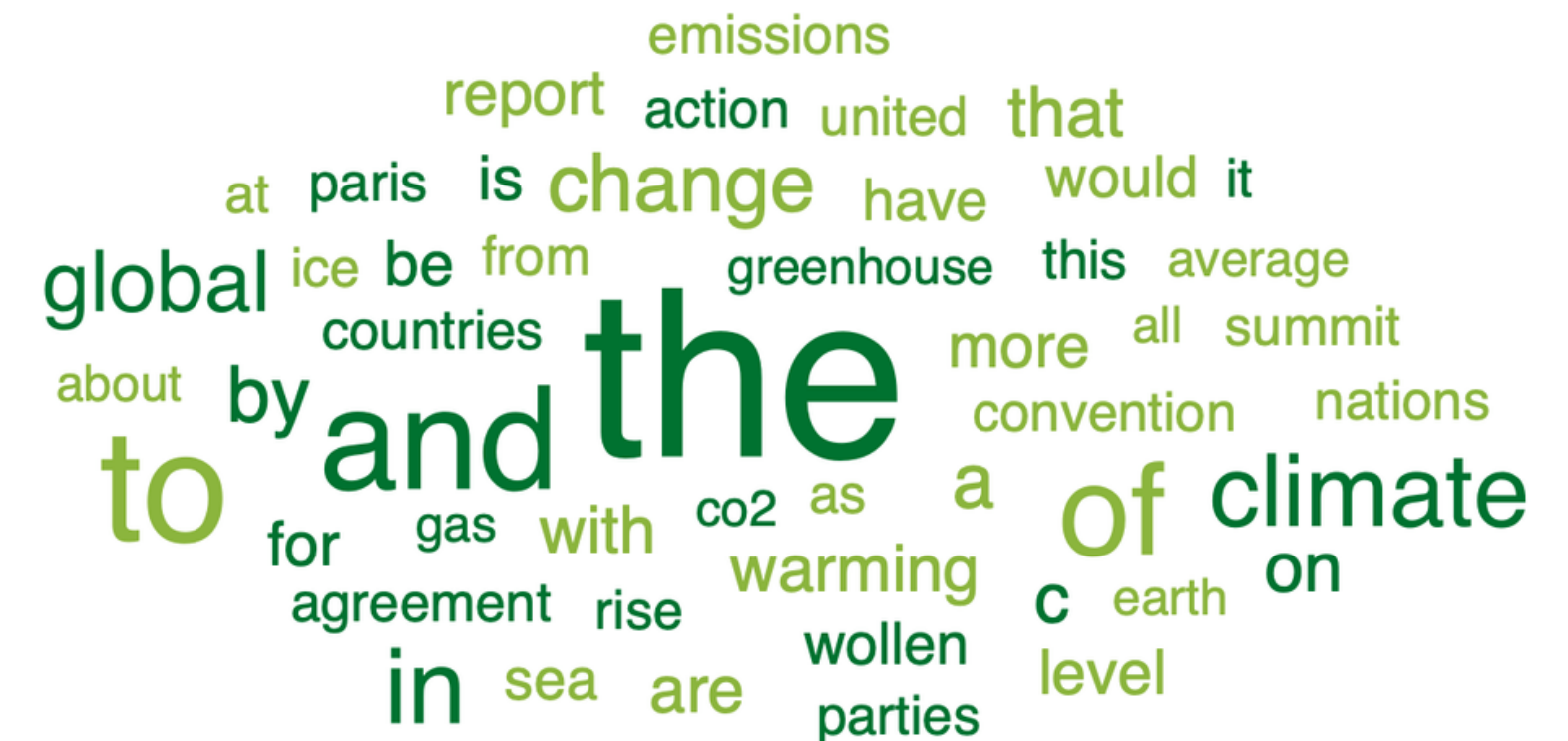
Proprietary document of Indonesia AI 2023

Stopwords adalah kata-kata yang umum dan tidak memberikan banyak informasi penting dalam analisis atau pemrosesan teks.

Kata-kata ini sering muncul secara berulang dalam teks dan cenderung tidak memiliki makna khusus yang relevan dalam konteks tertentu.

Beberapa contoh stopwords dalam bahasa Inggris meliputi "the", "is", "and", dan sebagainya.

Indonesia AI



Gambar: maxqda

STOPWORDS (2)

Proprietary document of Indonesia AI 2023

Menghilangkan stopwords penting untuk dilakukan karena dapat mempercepat pemrosesan teks, mengurangi dimensi data, dan memfokuskan analisis pada kata-kata yang lebih informatif atau signifikan dalam teks.

Sebelum menghilangkan stopwords, kita perlu meninjau konteks analisis yang akan dilakukan.

the which
and
is
do are

PUNCTUATION

Tanda baca digunakan dalam teks untuk memberikan struktur, memisahkan kalimat, menunjukkan kutipan, dan memberikan penekanan pada kata-kata tertentu.

Keberadaan tanda baca seperti titik, koma, tanda tanya, dan tanda seru sering tidak memberikan kontribusi signifikan terhadap pemahaman teks dalam beberapa kasus analisis teks.

What Are the 16 Punctuation Marks?

<p>Period</p> <p>.</p> <p>Indicates that a thought or sentence is complete</p>	<p>Question Mark</p> <p>?</p> <p>Makes a sentence into a question</p>	<p>Exclamation Point</p> <p>!</p> <p>Shows emphasis or emotion in a sentence</p>	<p>Comma</p> <p>,</p> <p>Provides pauses by separating parts of a sentence</p>
<p>Semicolon</p> <p>;</p> <p>Connects two separate but related independent clauses</p>	<p>Colon</p> <p>:</p> <p>Introduces or points to related text</p>	<p>En Dash</p> <p>–</p> <p>Shows number ranges and connections between similar words</p>	<p>Em Dash</p> <p>—</p> <p>Interrupts a sentence to add information or show emphasis</p>
<p>Hyphen</p> <p>–</p> <p>Joins related words together to create compound words</p>	<p>Parentheses</p> <p>()</p> <p>Enclose extra information in sentences</p>	<p>Brackets</p> <p>[]</p> <p>Add context to a quote or offset text within parentheses</p>	<p>Braces</p> <p>{ }</p> <p>Set off incidental or optional information or thoughts from the rest of the sentence</p>
<p>Apostrophe</p> <p>'</p> <p>Indicates that letters have been omitted or shows a noun's possession</p>	<p>Quotation Marks</p> <p>“ ”</p> <p>Mark quotes or citations in a sentence</p>	<p>Single Quotation Marks</p> <p>‘ ’</p> <p>Set off quotes inside larger quotes</p>	<p>Ellipsis</p> <p>...</p> <p>Omits parts of your writing</p>

YOURDICTIONARY

Gambar: yourdictionary

SPECIAL CHARACTER

Special characters (karakter khusus) merujuk pada karakter non-alfanumerik atau karakter yang tidak termasuk dalam kategori huruf atau angka.

Karakter khusus seperti simbol mata uang, tanda hubung, tanda kurung, atau karakter lainnya yang tidak relevan dengan analisis teks dapat mengganggu pemrosesan dan perlu dihapus.

Queries

fx = Table.AddColumn("#Added Custom", "Special Characters", each Te

	ABC 123 Text String	ABC 123 Clean Data	ABC 123 Special Characters
1	ABC--123	ABC123	--
2	ABC***	ABC	***
3	<>ABC<>	ABC	<>
4	!@#\$123	123	!@#\$
5	(ABC)	ABC	()
6	###xln cad	xln cad	###
7	XL n CAD 789!!!	XLnCAD789	!!!
8	sachin TENDULKAR	sachinTENDULKAR	
9	SACHIN tendulkar	SACHINTendulkar	
10	Sachin Tendulkar	SachinTendulkar	
11	Excel 2019	Excel2019	
12	AutoCAD 2020	AutoCAD2020	

Gambar: xln cad

EMOJI DAN EMOTICON

Proprietary document of Indonesia AI 2023

Emoji dan emoticon adalah simbol grafis yang digunakan untuk menyampaikan emosi atau ekspresi dalam teks, tetapi dalam beberapa kasus analisis teks, mereka mungkin tidak relevan atau mengganggu.

Sebagai alternatif, preprocessing terhadap emoji dan emoticon dapat berupa konversi emoji atau emoticon pada teks sesuai dengan dictionary tertentu.

```
1 emoji_dict = {  
2     '😊': 'Happy',  
3     '😄': 'Smile',  
4     '😴': 'Yawn',  
5     '👌': 'Okay',  
6     '👍': 'Thumbs up',  
7     '👎': 'Thumbs down',  
8     '🙏': 'Praying hands',  
9 }  
10
```

CASE FOLDING

Proprietary document of Indonesia AI 2023

Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil atau huruf besar, sehingga tidak ada perbedaan antara huruf besar (uppercase) dan huruf kecil (lowercase) dalam teks.

Tujuan dari case folding adalah untuk menyederhanakan teks dan menghilangkan perbedaan kapitalisasi yang mungkin tidak relevan dalam analisis teks.

Indonesia AI

Input:

```
[  
    'Hello World',  
    'HEllo WoRld'  
]
```

Output:

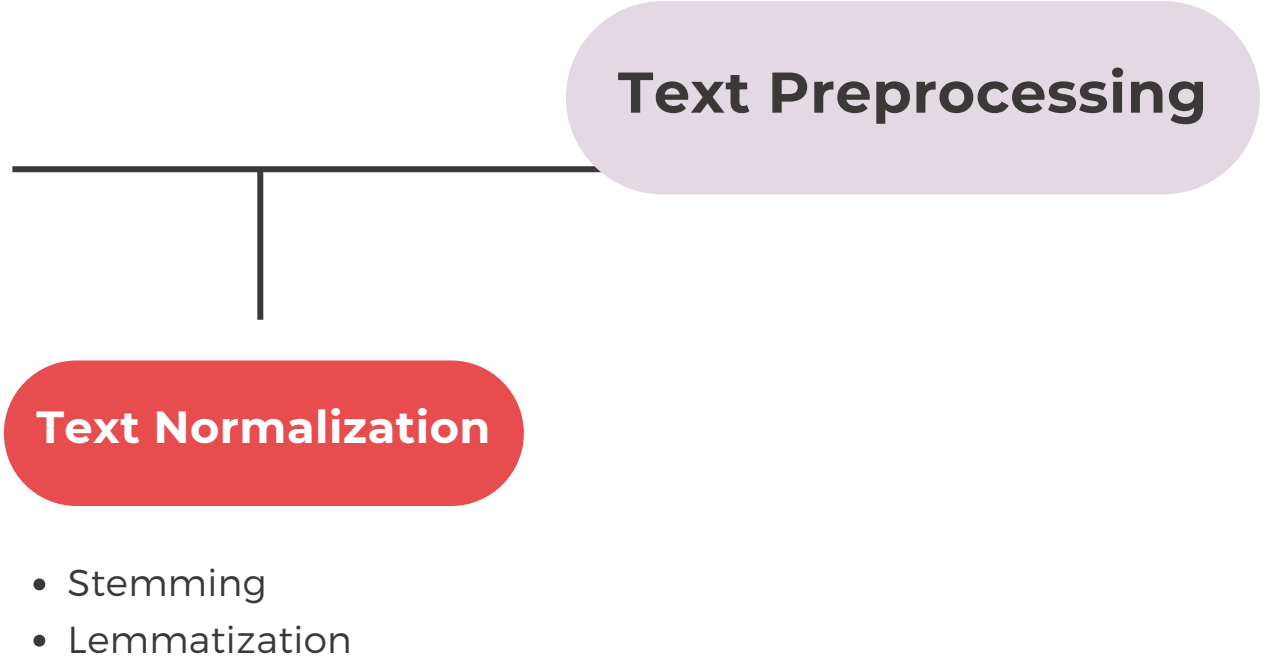
```
[  
    'hello world',  
    'hello world'  
]
```

Text Normalization

TEXT NORMALIZATION

Text normalization adalah proses mengubah teks ke dalam bentuk standar atau normal agar lebih konsisten dan dapat diolah secara lebih efektif.

Tujuannya adalah untuk menyederhanakan teks, menghilangkan variasi yang tidak relevan, dan memastikan konsistensi dalam representasi kata-kata.



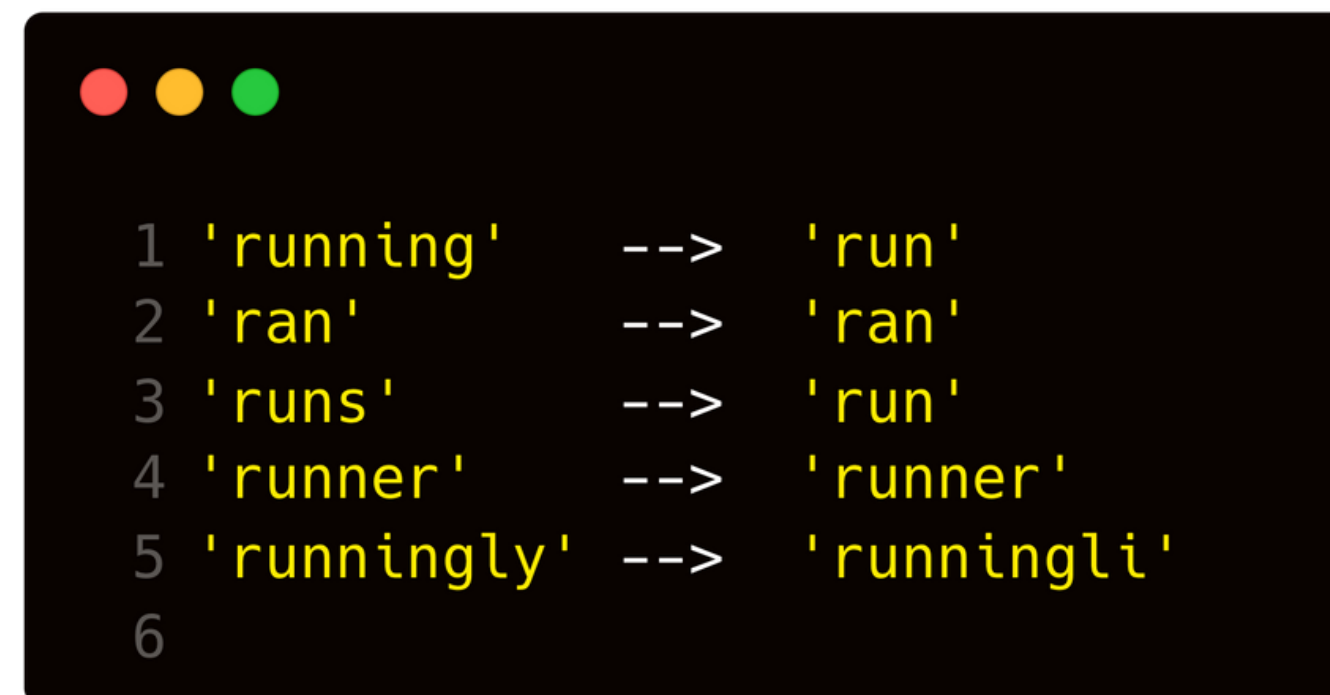
STEMMING

Proprietary document of Indonesia AI 2023

Stemming adalah proses menghilangkan imbuhan atau akhiran kata untuk menghasilkan bentuk dasar kata, yang disebut "kata dasar" atau "stem". Tujuannya adalah untuk mengurangi variasi bentuk kata yang serupa agar dapat diperlakukan sebagai entitas yang sama dalam analisis teks.

Stemming lebih sederhana dan cenderung menghasilkan hasil yang lebih kasar atau tidak sempurna.

Indonesia AI



```
1 'running' --> 'run'
2 'ran' --> 'ran'
3 'runs' --> 'run'
4 'runner' --> 'runner'
5 'runningly' --> 'runningli'
6
```

Stemming pada Bahasa Inggris

KEUNTUNGAN STEMMING

Proprietary document of Indonesia AI 2023

Text normalization menggunakan stemming memberikan 2 keuntungan utama, yakni:

1. Pengurangan dimensi
2. Peningkatan akurasi

Namun pada praktiknya, keputusan untuk melakukan stemming dibutuhkan peninjauan sesuai objektif dari proyek yang dikerjakan.

A terminal window with a dark blue background and three colored window control buttons (red, yellow, green) at the top left. It displays a list of Indonesian words and their stemmed forms, numbered 1 through 8. The words are in green, and the stemmed forms are in white. The mapping is as follows: 1. 'berjalan' to 'jalan', 2. 'jalan-jalan' to 'jalan', 3. 'lari' to 'lari', 4. 'lari-lari' to 'lari', 5. 'lari-larian' to 'lari', 6. 'membaca' to 'baca', 7. 'mengaji' to 'aji', and 8. (empty line).

```
1 'berjalan'      --> 'jalan'
2 'jalan-jalan'   --> 'jalan'
3 'lari'          --> 'lari'
4 'lari-lari'     --> 'lari'
5 'lari-larian'   --> 'lari'
6 'membaca'       --> 'baca'
7 'mengaji'       --> 'aji'
8
```

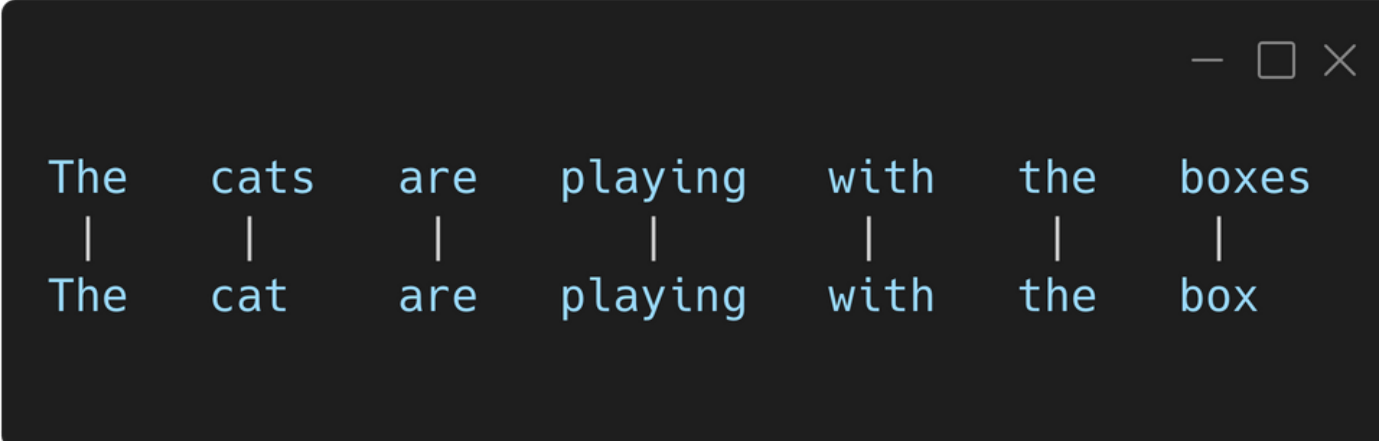
Stemming pada Bahasa Indonesia

LEMMATIZATION

Proprietary document of Indonesia AI 2023

Lemmatization adalah proses mengubah kata-kata dalam teks menjadi bentuk dasarnya yang disebut "lemma". Lemma adalah kata dasar yang memiliki makna yang sama dengan kata tersebut.

Lemmatization menggunakan informasi morfologi bahasa, kamus, atau aturan linguistik untuk mengubah kata ke bentuk dasarnya.



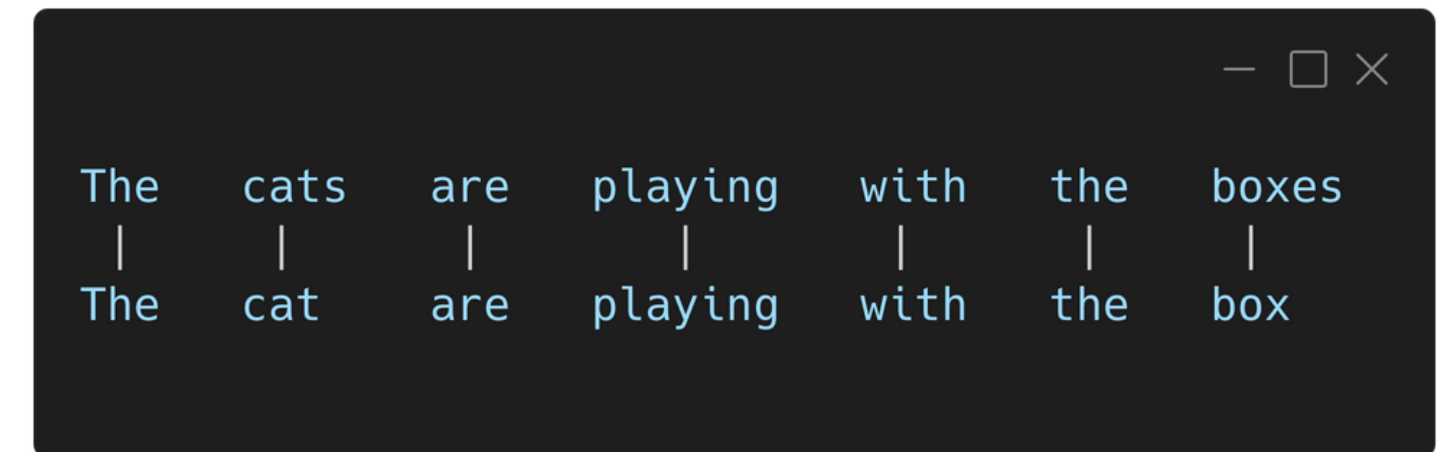
The	cats	are	playing	with	the	boxes
The	cat	are	playing	with	the	box

KEUNTUNGAN LEMMATIZATION

Proprietary document of Indonesia AI 2023

Text normalization menggunakan lemmatization memberikan 2 keuntungan utama, yakni:

1. Pengembalian kata yang lebih bermakna
2. Mempertahankan struktur grammatical



STEMMED VS LEMMATIZED

Proprietary document of Indonesia AI 2023

Perbedaan utama antara stemming dan lemmatization terletak pada tingkat keakuratan dan validitas hasil. Stemming cenderung menghasilkan bentuk dasar yang sederhana, tetapi tidak selalu valid atau umum.

Sementara itu, lemmatization menghasilkan bentuk kata dasar yang valid dan relevan dengan konteks, dengan mempertimbangkan aturan bahasa secara lebih komprehensif.

Indonesia AI

```
# stemming #
```

running	mice	better
run	mice	better

```
# lemmatization #
```

running	mice	better
run	mouse	good

Text Tokenization

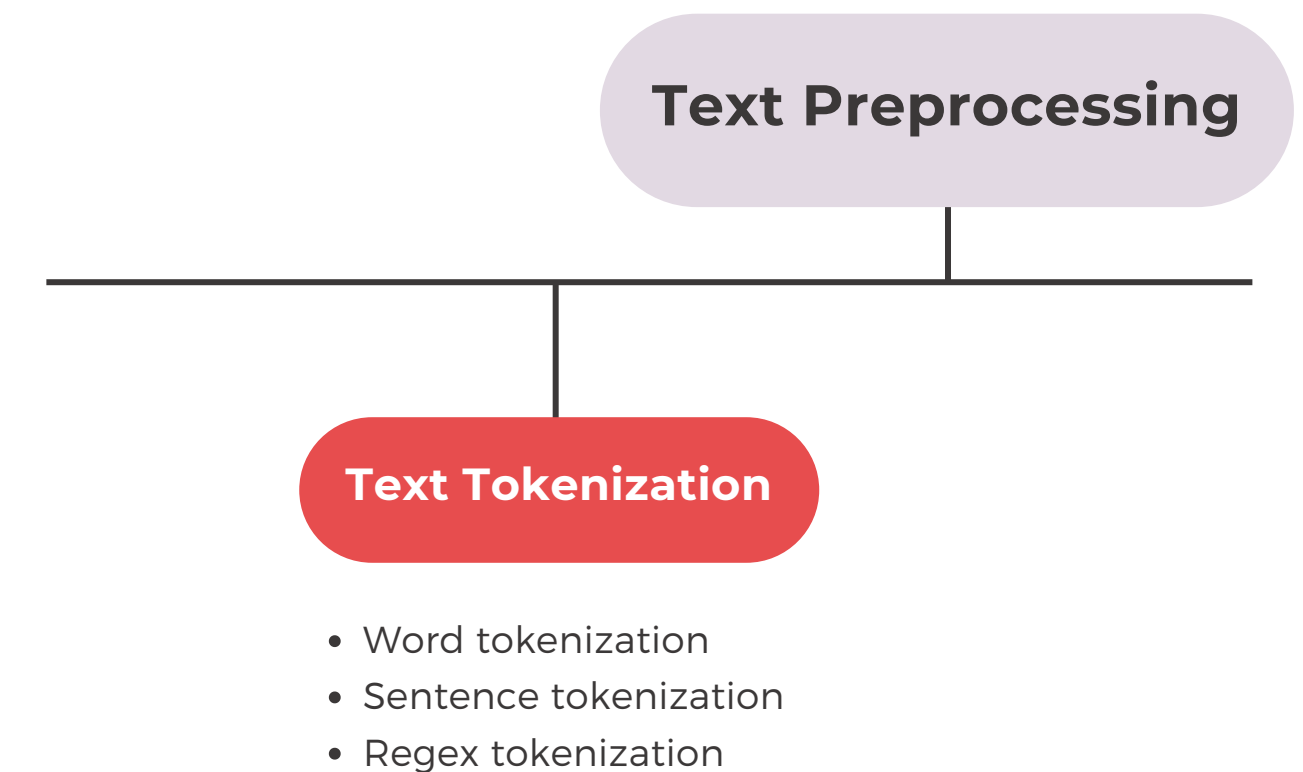
TEXT TOKENIZATION

Proprietary document of Indonesia AI 2023

Text tokenization adalah proses membagi teks menjadi unit-unit yang lebih kecil, yang disebut "token". Token bisa berupa kata, frasa, atau simbol lainnya, tergantung pada konteks dan tujuan analisis.

Tujuan utama dari text tokenization adalah untuk memisahkan teks menjadi unit-unit terpisah agar dapat diolah lebih lanjut dalam pemrosesan bahasa alami dan analisis teks.

Indonesia AI



RAĢAM TEXT TOKENIZATION

Proprietary document of Indonesia AI 2023

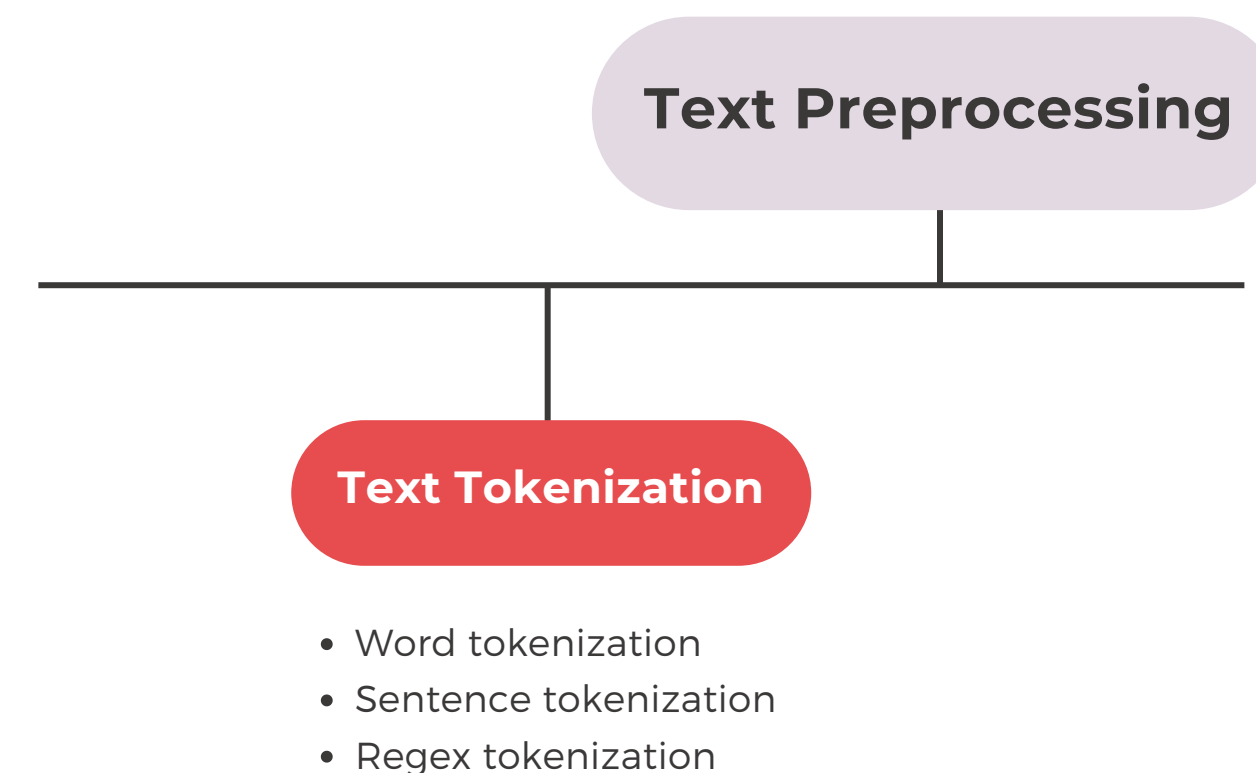
Pada proyek NLP, text tokenization yang umum digunakan adalah sebagai berikut:

1. Word Tokenization
2. Sentence Tokenization

Namun ada juga variasi lain, beberapa diantaranya:

1. Regex Tokenization
2. Tweet Tokenization
3. N-gram Tokenization

Indonesia AI



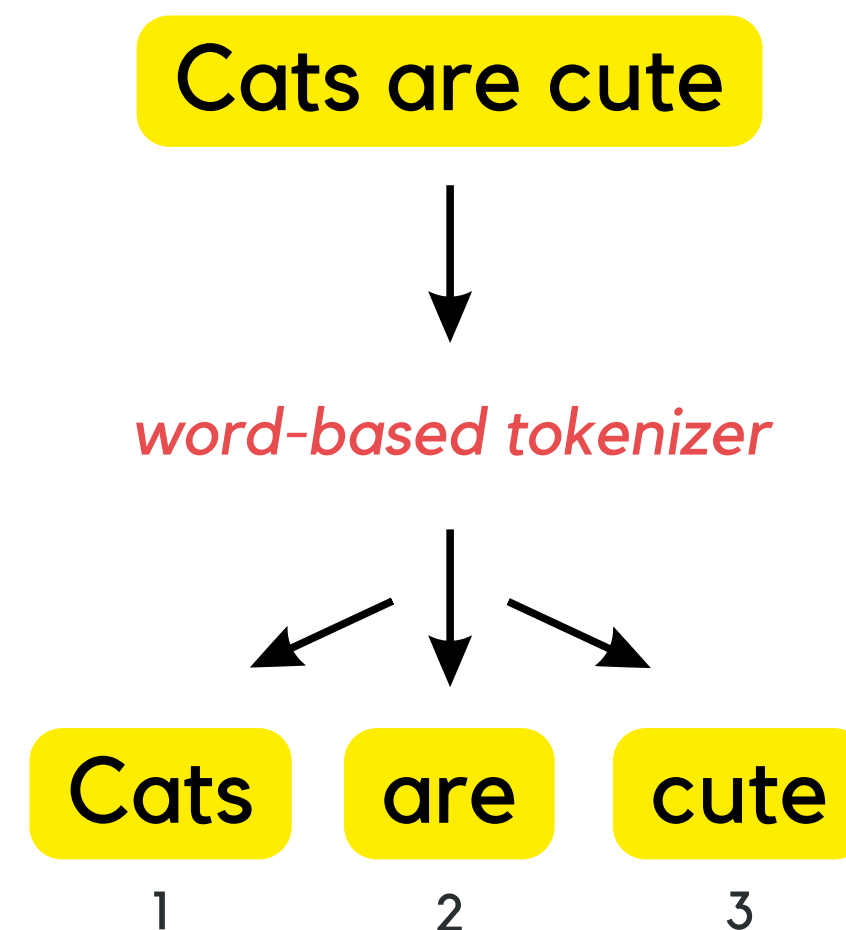
WORD TOKENIZATION

Proprietary document of Indonesia AI 2023

Word Tokenization adalah proses membagi teks menjadi unit-unit yang lebih kecil berdasarkan kata. Teks dipecah menjadi kata-kata terpisah.

Word Tokenization dapat dilakukan dengan menggunakan metode sederhana seperti membagi teks berdasarkan spasi atau menggunakan algoritma yang lebih kompleks seperti Tokenizer dari library NLP seperti NLTK.

Indonesia AI

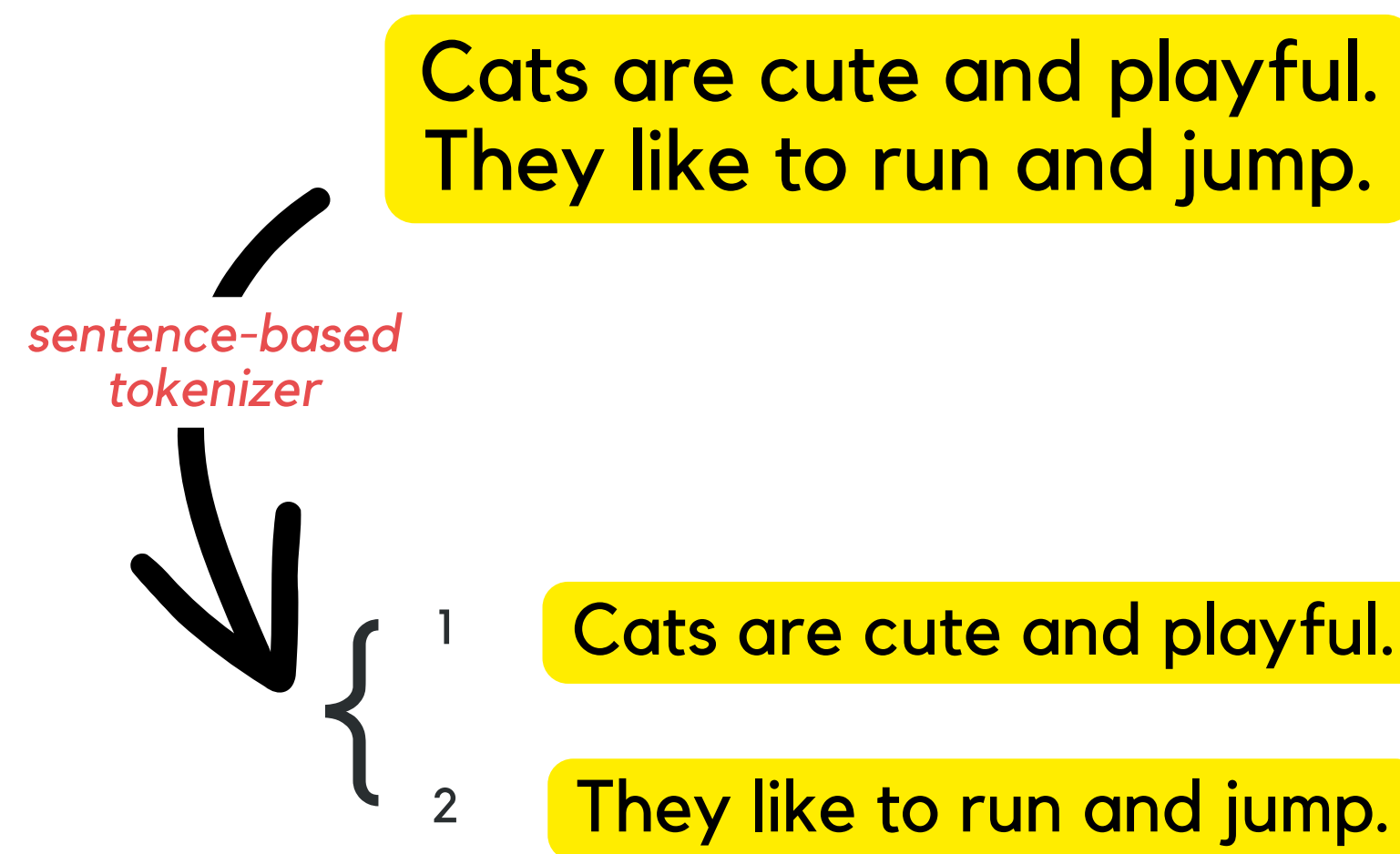


SENTENCE TOKENIZATION

Proprietary document of Indonesia AI 2023

Sentence Tokenization adalah proses membagi teks menjadi unit-unit yang lebih kecil berdasarkan kalimat. Teks dipecah menjadi kalimat-kalimat terpisah.

Sentence Tokenization dapat dilakukan dengan menggunakan aturan heuristik seperti pemisahan berdasarkan tanda baca, penggunaan model statistik, atau menggunakan library NLP yang menyediakan fitur Sentence Tokenizer.



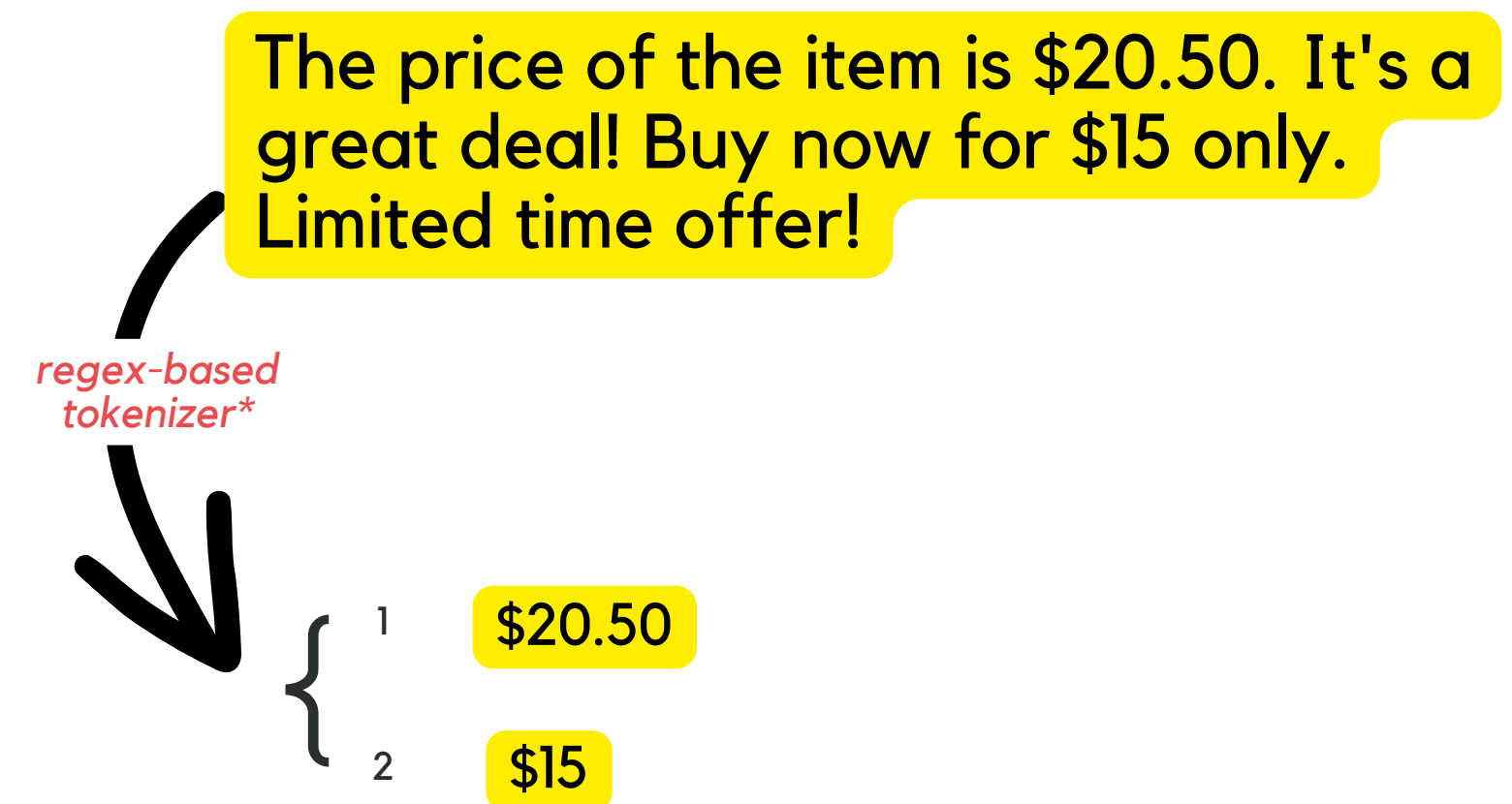
REGEX TOKENIZATION

Proprietary document of Indonesia AI 2023

Regex Tokenization adalah proses membagi teks menjadi token berdasarkan pola tertentu yang ditentukan menggunakan regex. Pola regex yang digunakan dalam tokenization dapat bervariasi tergantung pada kebutuhan dan jenis teks yang akan diproses.

Regex tokenization berguna dalam kasus-kasus di mana pemisahan token tidak dapat diatasi dengan metode tokenization konvensional.

Indonesia AI



TEKNIK TOKENIZATION LAINNYA

Proprietary document of Indonesia AI 2023

```
Input:
[
  'wah belajar NLP #menyenangkan #serubanget ketagihan~'
]

Output:
[
  'wah',
  'belajar',
  'NLP',
  '#menyenangkan', '#serubanget', 'ketagihan',
  '~'
]
```

Tweet Tokenization

```
Input:
[
  'I love to eat pizza.'
]

# ngram with n=3

Output:
('I', 'love', 'to')
('love', 'to', 'eat')
('to', 'eat', 'pizza')
('eat', 'pizza', '.')
```

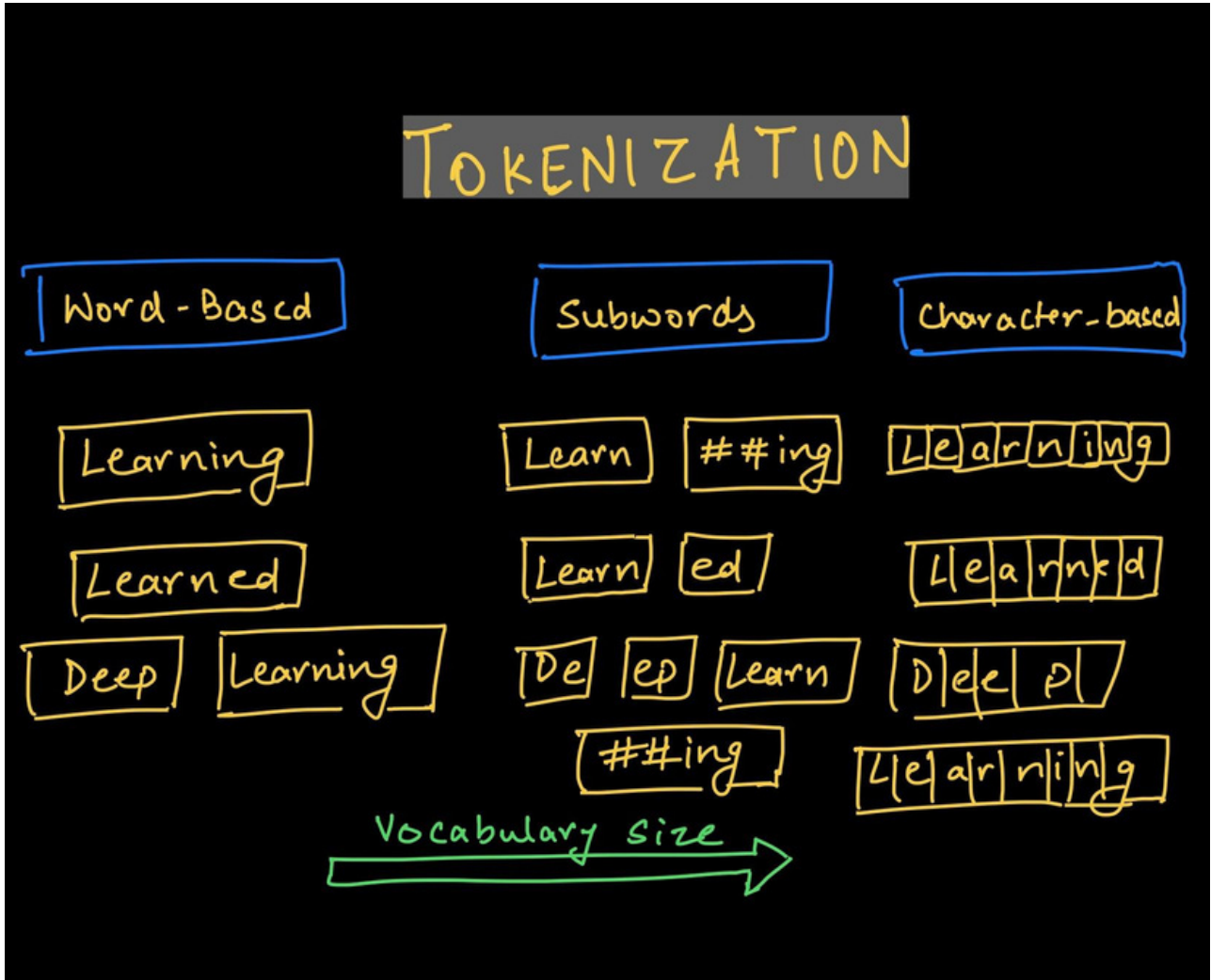
N-gram Tokenization

PEMILIHAN TEKNIK TOKENIZATION

Proprietary document of Indonesia AI 2023

Dalam memilih teknik tokenization, perlu mempertimbangkan ukuran unit yang digunakan untuk membagi teks menjadi token. Semakin kecil unit yang digunakan, semakin banyak fitur atau informasi yang dapat diperoleh.

Penting untuk mempertimbangkan trade-off antara jumlah fitur dan tingkat detail informasi dalam memilih teknik tokenization yang paling sesuai untuk kebutuhan analisis.



Gambar: freecodecamp

Summary

HOW MUCH IS ENOUGH FOR TEXT PREPROCESSING?

Proprietary document of Indonesia AI 2023

Sebelum melakukan Text Preprocessing penting untuk meninjau beberapa hal sebelum memilih teknik preprocessing teks yang sesuai.

Hal yang perlu ditinjau sebelum melakukan teks preprocessing:

1. Tujuan Analisis
2. Format jenis teks
3. Kondisi data mentah
4. Konteks dan domain
5. Ukuran teks
6. Special case-handling



HOW MUCH IS ENOUGH FOR TEXT PREPROCESSING?

Level of text preprocessing needed		
	Domain Specific / Noisy Texts	General / Well Written Texts
Lots of data	<ul style="list-style-type: none">- <u>Moderate</u> pre-processing- Text enrichment <u>could be helpful</u>	<ul style="list-style-type: none">- <u>Light</u> pre-processing- Text enrichment could be helpful, but <u>not critical</u>
Sparse data	<ul style="list-style-type: none">- <u>Heavy</u> pre-processing- Text enrichment is <u>important</u>	<ul style="list-style-type: none">- <u>Moderate</u> pre-processing- Text enrichment <u>could be helpful</u>

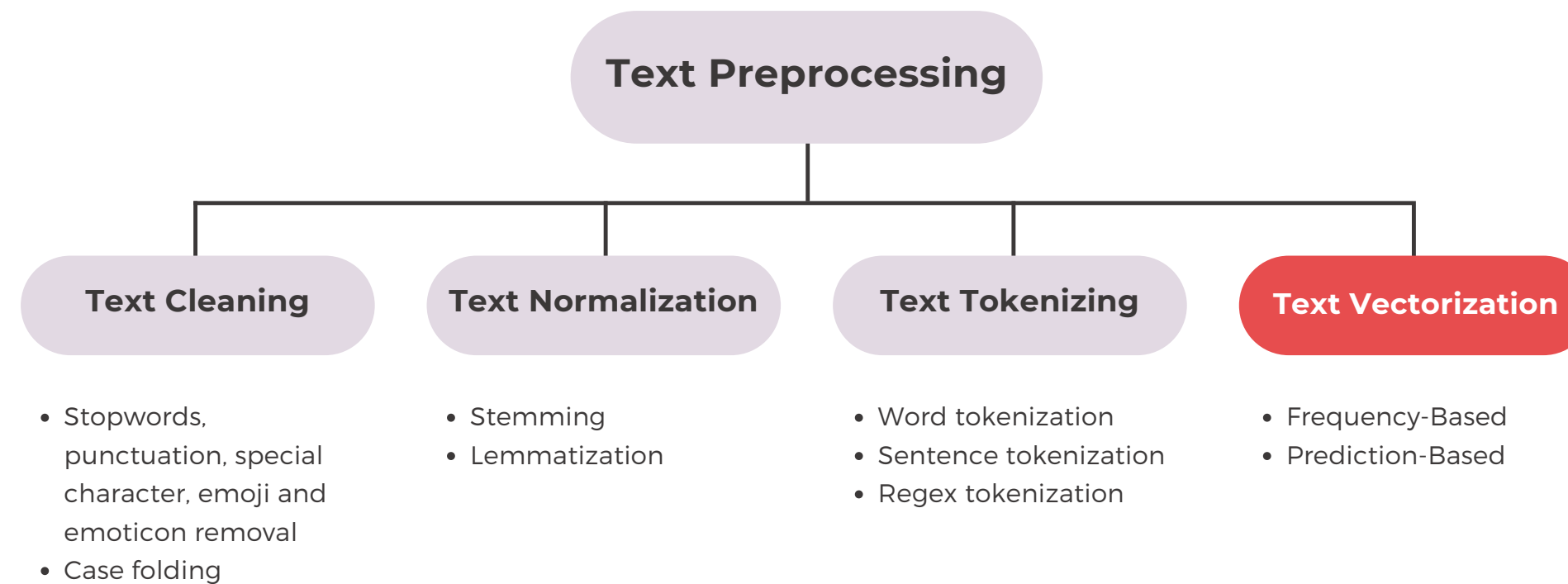
By: Kavita Ganesan

Gambar: kdnuggets - Kavita Ganesan

Any question guys ~

MASIH INGAT DIAGRAM INI?

Proprietary document of Indonesia AI 2023



Bagaimana dengan Text Vectorization?

Selain Text Vectorization, tahap ini juga dikenal dengan nama Text Representation.

Nah text representation ini terdiri dari beberapa jenis, masing-masing jenis ada variasi algoritmanya.

Sehingga butuh sesi terpisah agar bahasannya dapat maksimal~



Hands-on~

Terimakasih!