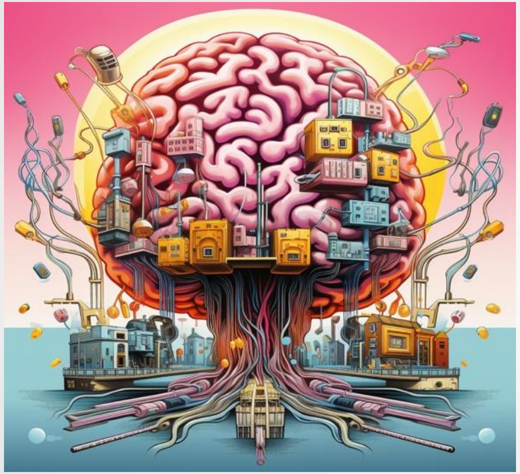
esprit
Se former autrement
HONORIS UNITED UNIVERSITIES



Chapter 2: Data pre-processing

Unit: Advanced Deep Learning

Introduction

- Text preprocessing is an essential step in **natural language processing** (NLP).
- Involves cleaning and transforming unstructured text data to prepare it for analysis.
- Includes:
  - tokenization,
  - Data cleaning
  - stemming,
  - lemmatization,
  - part-of-speech tagging.
  - Name Entity Recognition

Framework

We assume:

- A token is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1, ..., V.
- A document is a sequence of N words denoted by d = (w1,w2, ...,wN), where wn is the N-th word in the sequence.
- A corpus is a collection of M documents denoted by D = (d1, d2, ..., dM)

Example: Wikipedia, All the articles of the NYT in 2021…

Document

A Document can be:

- A Sentence
- A Paragraph
- A sequence of characters

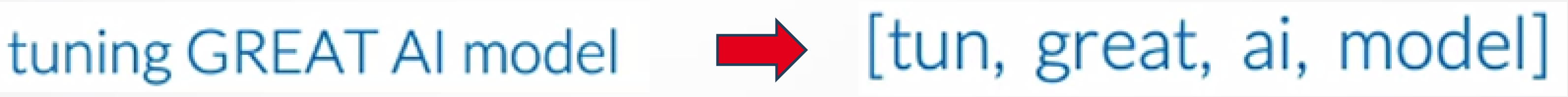
Token

With regard to our end task, a token can be:

- A word
- A sub-word: e.g. a sequence of 3 characters
- A character
- An sequence of characters (sometimes a word, sometimes several words, sometimes a sub-word…)

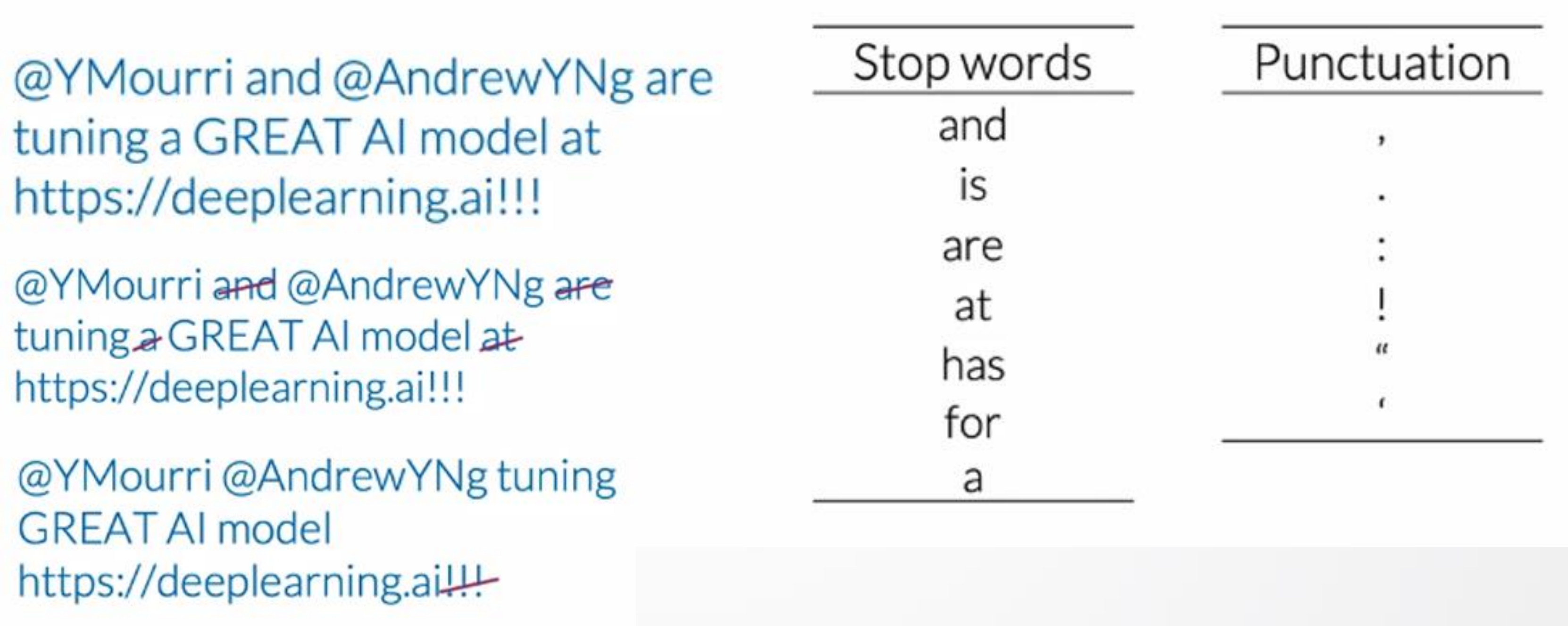
Tokenization

- Tokenization is the process of breaking down large blocks of text such as paragraphs and sentences into smaller, more manageable units.
- Objectif: obtain a more accurate representation of the underlying patterns and trends present in the text data.



tuning GREAT AI model → [tun, great, ai, model]

Data cleaning : Stop words and punctuation



@YMourri and @AndrewYNg are tuning a GREAT AI model at https://deeplearning.ai!!!

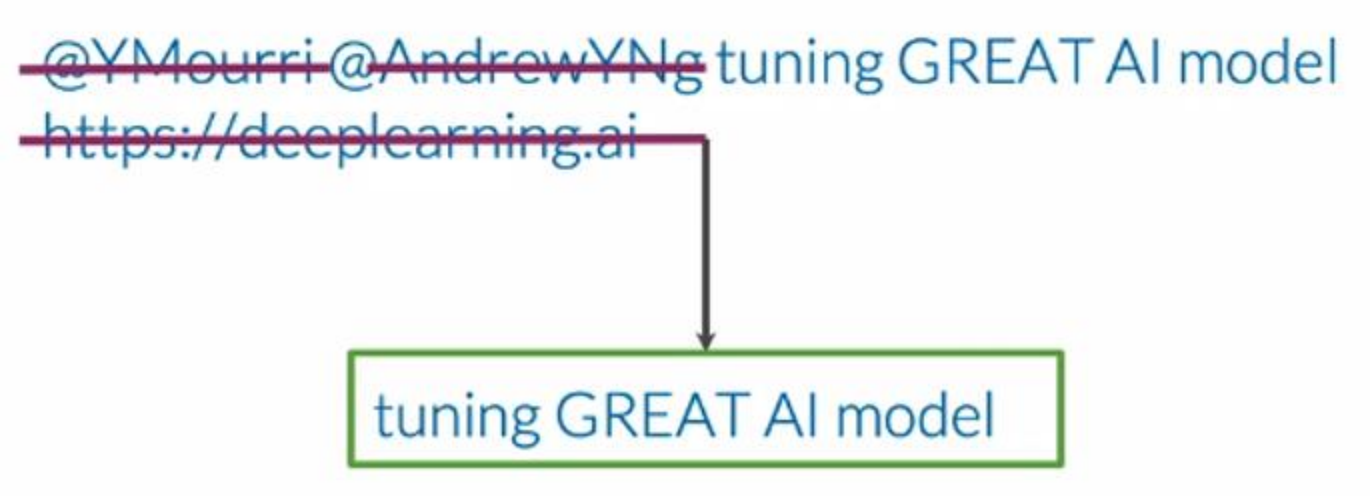
@YMourri ~~and~~ @AndrewYNg ~~are~~ tuning ~~a~~ GREAT AI model ~~at~~ https://deeplearning.ai!!!

@YMourri @AndrewYNg tuning GREAT AI model https://deeplearning.ai~~!!!~~

Stop words
and
is
are
at
has
for
a

Punctuation
,
.
:
!
“
‘

Data cleaning : Handles and URLs

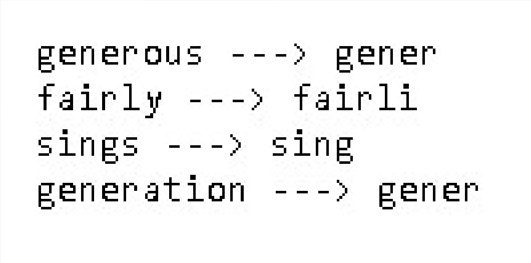


~~@YMourri @AndrewYNg~~ tuning GREAT AI model
~~https://deeplearning.ai~~

tuning GREAT AI model

Stemming

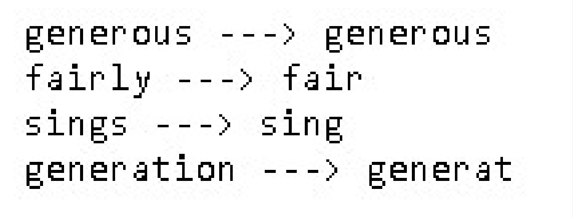
- This step, known as text standardization, stems or reduces words to their root or base form.
- Stemming can cause the root form to lose its meaning or not reduce to a proper English word.
- Stemming is beneficial in scenarios where speed is crucial, such as search engines and text mining.
- It helps in reducing the dimensionality of text data, allowing for faster processing and retrieval of relevant information



```
generous ---> gener
fairly ---> fairli
sings ---> sing
generation ---> gener
```

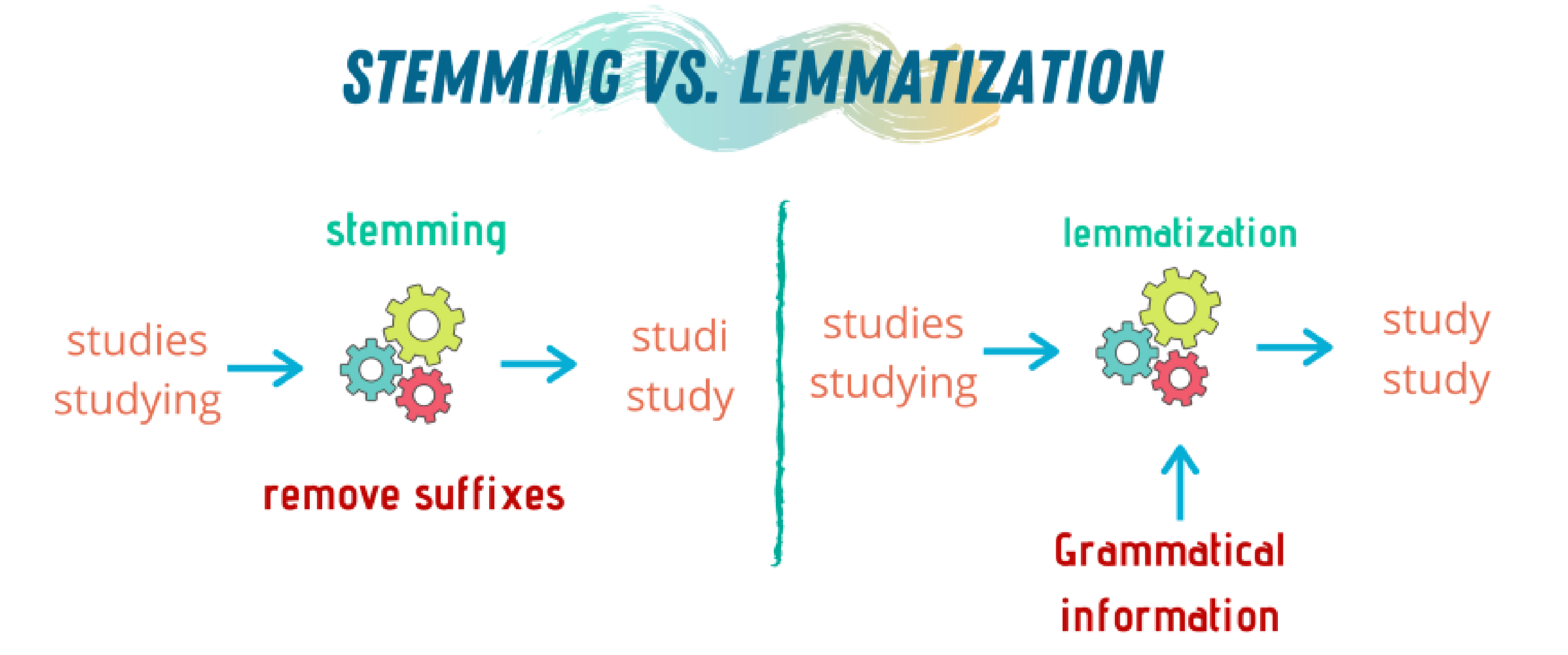
Lemmatization

- It stems from the word but ensures it does not lose meaning.
- Lemmatization has a pre-defined dictionary that stores the context of words and checks the word in the dictionary while diminishing.
- Lemmatization is preferred when accuracy is essential, especially in applications requiring semantic understanding.
- It ensures that words with similar meanings are treated as the same entity, improving the quality of analysis in sentiment analysis and chatbots.



```
generous ---> generous
fairly ---> fair
sings ---> sing
generation ---> generat
```

Stemming vs lemmatization



Part-of-speech tagging.

- Part-of-Speech (POS) tagging involves assigning a grammatical category (such as noun, verb, adjective, etc.) to each word in a sentence.
- POS tags include:
  - Noun (NN)
  - Verb (VB)
  - Adjective (JJ)
  - Adverb (RB)
  - Pronoun (PRP)
  - Preposition (IN)

Part-of-speech tagging.

POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

[PRON , NOUN, VERB, NOUN, CC, PRON, VERB, PREP, NOUN, PUNCT]

Name Entity Recognition

NER: Find the **Name-Entities** in a sentence

[My, name, is, Bob, and, I, live, in, New, York, !]

*[O , O, O, **B-PER**, O, O, O, O, **B-LOC**, **I-LOC,** O]*

PER: PERSON

LOC: LOCATION