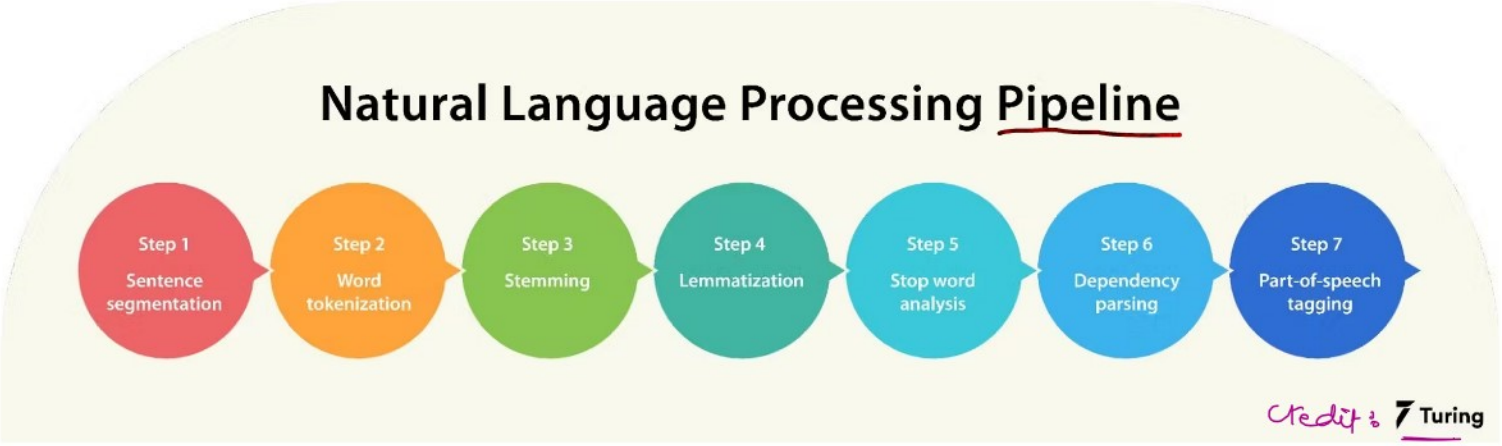
NLP Pipeline

02 November 2024 22:15

Focus on steps → (not on the sequence)



step #① Sentence Segmentation

- it is the very first step in NLP pipeline.
- divides the entire paragraph into different sentences for better understanding.

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

source: wikipedia

sentence segmentation

1. London is the capital and most populous city of England and the United Kingdom.
2. Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia.
3. It was founded by the Romans, who named it Londinium.

step #② Word Tokenization

- it is the process of splitting a text into individual words or tokens.

Sentence: I love NLP

words or tokens:

Sentence: I love NLP

word tokens : ["I", "love", "NLP"]

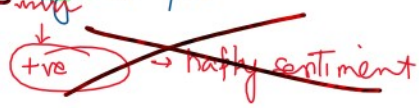
Note: It is a crucial step as it transforms raw text data into a format that can be processed by ML/AI algorithms

subword tokens:

- a technique to handle out-of-vocabulary words by breaking them into smaller but meaningful units.

For sentiment analysis:

2017 – my first "baby project" → I am unhappy with you.



words: unhappiness, unwell

subword tokens: ["un", "happiness"]

character tokens:

- In some cases, tokenization is done at the characters level. where each token is a single character

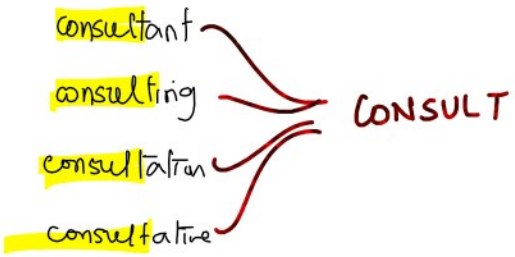
word: cat

character token: ["c", "a", "t"]

tL, dr: too long, didn't read

- it is a short hand used on the internet to summarize long content, articles etc. into a brief, digestible form.

step#3 stemming



stemming is a text normalization technique to reduce the words to their base or root form known as stem.