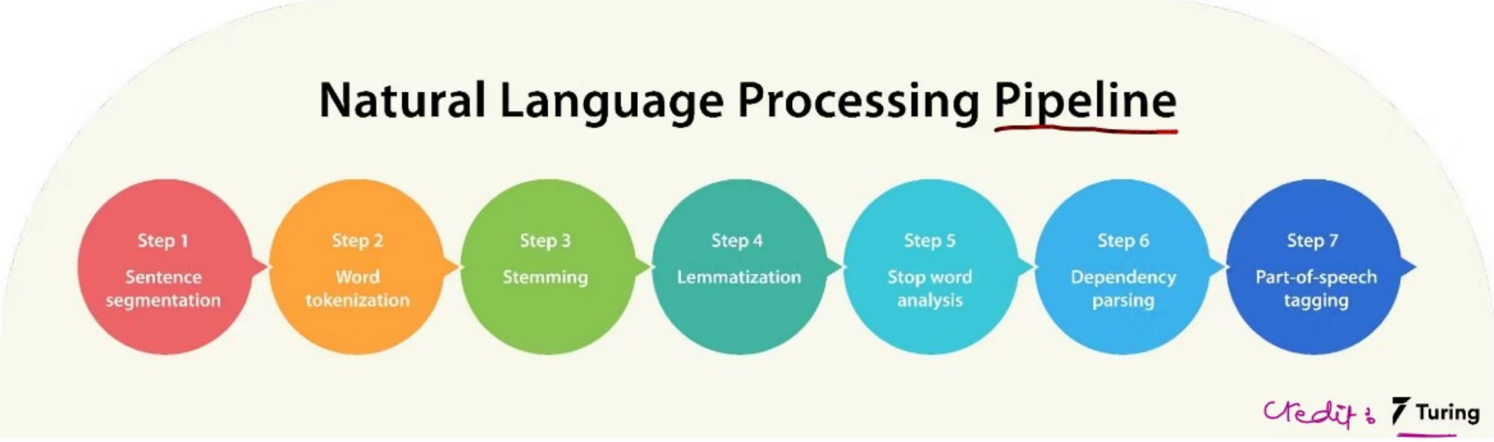
NLP Pipeline

02 November 2024 22:15

Focus on steps → (not on the sequence)



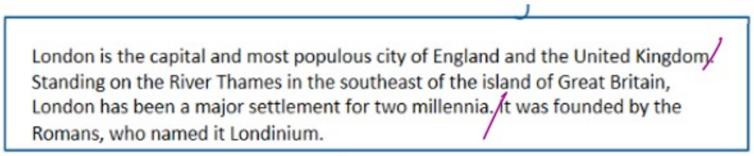
Natural Language Processing Pipeline

Step 1 Sentence segmentation
Step 2 Word tokenization
Step 3 Stemming
Step 4 Lemmatization
Step 5 Stop word analysis
Step 6 Dependency parsing
Step 7 Part-of-speech tagging

Credit: Turing

step#① Sentence Segmentation

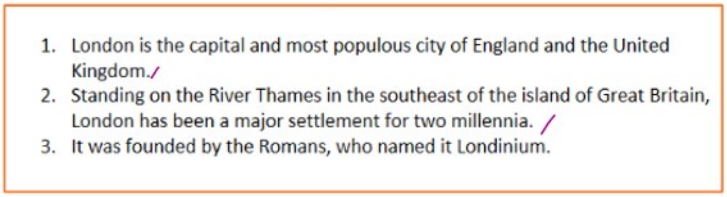
- it is the very first step in NLP pipeline.
- divides the entire paragraph into different sentences for better understanding.



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

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 $\boxed{\text{un}}$happy with you.

~~+ve → happy sentiment~~

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

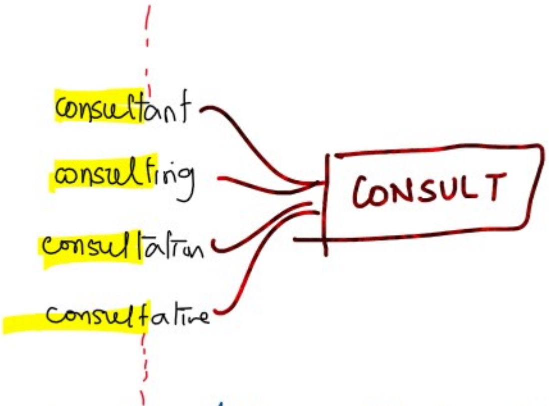
character token: ["c", "a", "t"]

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.

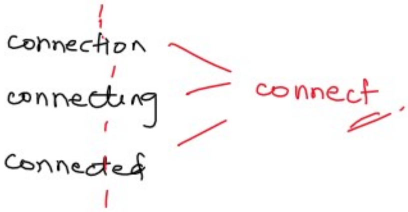
step#3 stemming

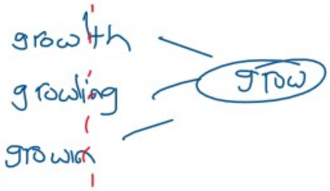
step#3 stemming



stemming is a text normalization technique to reduce the words to their base or root form* → (lemmatization is the right approach) known as stem.

* If only we follow the contextual approach.

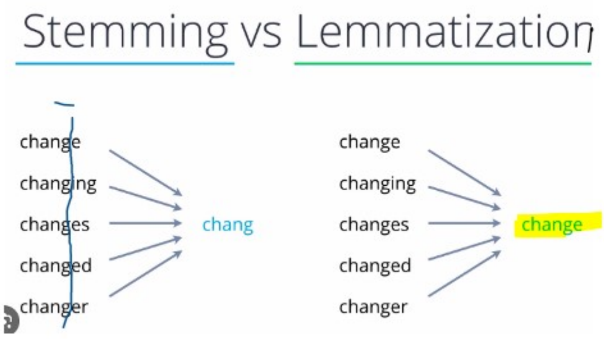




step# 4 Lemmatization

It is a text normalization technique that reduces words to their base or dictionary form → known as "lemma"

Unlike stemming which often uses heuristics (rules) to cut-off affixes while lemmatization involves a more sophisticated process of deriving the base form of word based on its meaning and context.


Stemming vs Lemmatization

ran → run

wrote → write

written ↗
(lemmatization)
writing ↗

Stemming:	Lemmatization:
"running" → "run"	"running" → "run"
"better" → "bett"	"better" → "good"
"flies" → "fli"	"flies" → "fly"
"drove" → "drove"	"drove" → "drive"
"cats" → "cat"	"cats" → "cat"
"worse" → "wor"	"worse" → "bad"

selecting | cutting
(3 characters)

Q. Well in case, it seems that why someone should do `stemming` rather than doing 'lemmatization' ?

R. **Speed and Simplicity**

stemming algorithms are generally simpler and faster than lemmatization algorithms.

stemming can be less resource intensive making it suitable for 'large-scale applications'

Choice between stemming vs lemmatization

Classification problems

In general, stemming is often used in classification problems which can help to reduce the dimensionality of text data by grouping different forms of a word together.

Treating related words as the same feature

Text Generation Problems:

In general, lemmatization is typically preferred in text generation tasks as it preserves the meaningful base form of the words.

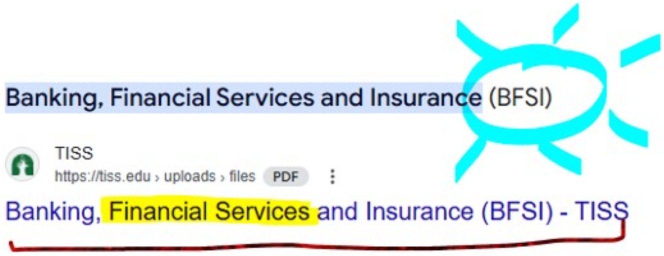
it maintains the meaning and context of words → leads to more coherent and accurate generate text.

Step#5 stop words analysis

- Words that are commonly used in a language but are often deemed irrelevant for specific NLP tasks

  Examples in English: a, an, the, and, or, is, in, of, to, etc.

- Primary goal of removing stop words is to focus on more meaningful words that carry substantial information for text analysis such as sentiment analysis, topic modeling



Banking, Financial Services and Insurance (BFSI)

TISS
https://tiss.edu › uploads › files PDF ⋮

Banking, Financial Services and Insurance (BFSI) - TISS

Topic Modeling

↳ it is an NLP technique used to discover the underlying Themes or topics within a collection of text documents.

Products

ICICI
↓
[OD against Salary]
↓
(overdraft) ✓

[Bajaj Finance] →

SEO

Application

Content recommendation:

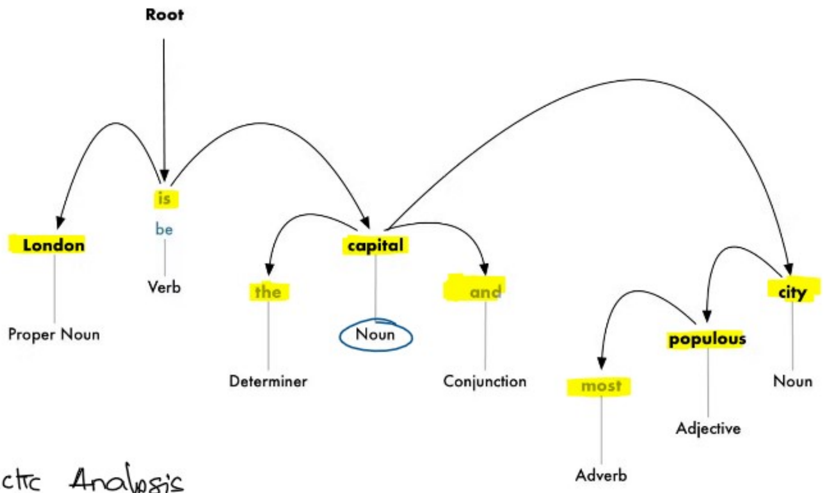
- suggesting articles or content based on the identified topics

Disclaimer # Discussing these things for learning not undermining any brand/institution

[generate articles/blogs]

Step #6 Dependency Parsing

Grammer



Syntactic Analysis

- is crucial for understanding the grammatical structure of sentences and how words relate to each other.

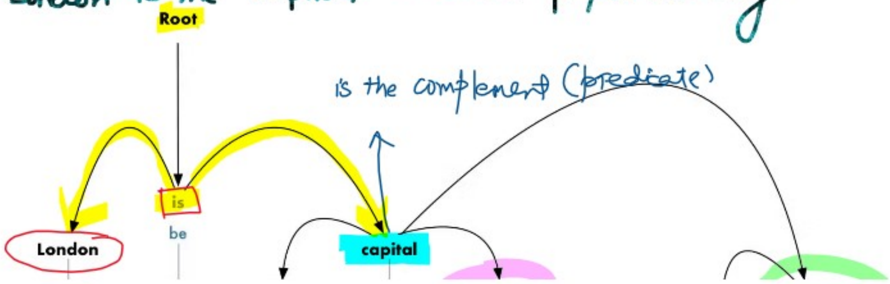
- it is the process of analyzing a sentence to determine its grammatical structure

Dependency Parsing

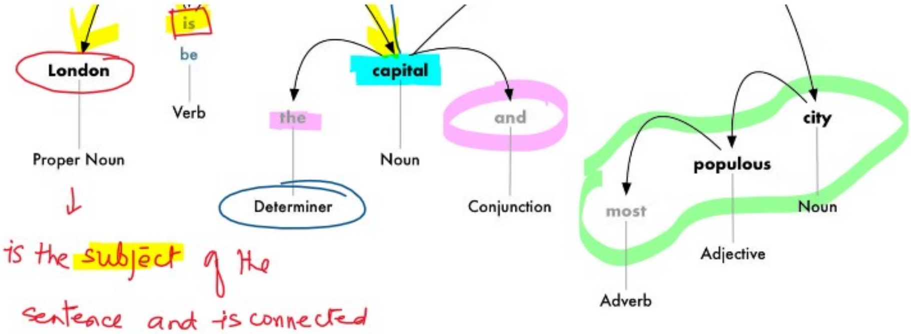
- Once we have identified the grammatical relationships between words, dependency parsing constructs a dependency tree or graph that shows how words in a sentence depend on each other.

Dependency Parse Tree

London is the capital and most populous city.

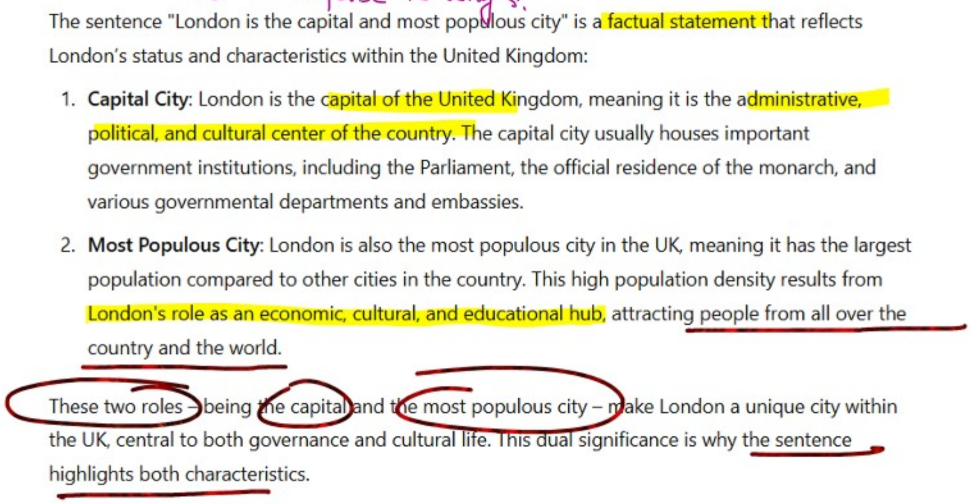


is the complement (predicate)



↓ is the subject of the sentence and is connected to the root verb "is"

ChatGPT 4o → response to why??



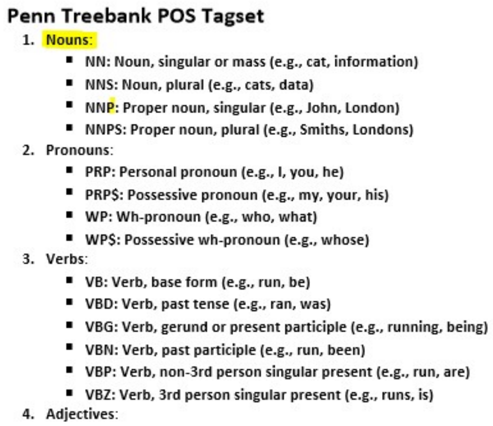
The sentence "London is the capital and most populous city" is a factual statement that reflects London's status and characteristics within the United Kingdom:

1. **Capital City**: London is the capital of the United Kingdom, meaning it is the administrative, political, and cultural center of the country. The capital city usually houses important government institutions, including the Parliament, the official residence of the monarch, and various governmental departments and embassies.
2. **Most Populous City**: London is also the most populous city in the UK, meaning it has the largest population compared to other cities in the country. This high population density results from London's role as an economic, cultural, and educational hub, attracting people from all over the country and the world.

These two roles – being the capital and the most populous city – make London a unique city within the UK, central to both governance and cultural life. This dual significance is why the sentence highlights both characteristics.

Step #7 POS Tagging
↓
(Parts of speech)

POS tags contains verbs, adverbs, nouns, pronouns, adjectives etc. that help to understand the meaning of words in a grammatically correct way in a sentence.



Penn Treebank POS Tagset

1. Nouns:
   - NN: Noun, singular or mass (e.g., cat, information)
   - NNS: Noun, plural (e.g., cats, data)
   - NNP: Proper noun, singular (e.g., John, London)
   - NNPS: Proper noun, plural (e.g., Smiths, Londons)
2. Pronouns:
   - PRP: Personal pronoun (e.g., I, you, he)
   - PRP$: Possessive pronoun (e.g., my, your, his)
   - WP: Wh-pronoun (e.g., who, what)
   - WP$: Possessive wh-pronoun (e.g., whose)
3. Verbs:
   - VB: Verb, base form (e.g., run, be)
   - VBD: Verb, past tense (e.g., ran, was)
   - VBG: Verb, gerund or present participle (e.g., running, being)
   - VBN: Verb, past participle (e.g., run, been)
   - VBP: Verb, non-3rd person singular present (e.g., run, are)
   - VBZ: Verb, 3rd person singular present (e.g., runs, is)
4. Adjectives:

   - **JJ: Adjective (e.g., big, blue)**
   - **JJR: Adjective, comparative (e.g., bigger, bluer)**
   - **JJS: Adjective, superlative (e.g., biggest, bluest)**

5. **Adverbs**:
   - **RB: Adverb (e.g., quickly, well)**
   - **RBR: Adverb, comparative (e.g., more quickly, better)**
   - **RBS: Adverb, superlative (e.g., most quickly, best)**
6. **Determiners**:
   - **DT: Determiner (e.g., the, a)**
   - **PDT: Predeterminer (e.g., all, both)**
   - **WDT: Wh-determiner (e.g., which, what)**
7. **Prepositions and Conjunctions**:
   - **IN: Preposition or subordinating conjunction (e.g., in, of, because)**
   - **CC: Coordinating conjunction (e.g., and, but, or)**
8. **Auxiliaries and Modals**:
   - **MD: Modal (e.g., can, should, will)**
9. **Interjections**:
   - **UH: Interjection (e.g., oh, wow)**
10. **Particles**:
    - **RP: Particle (e.g., up, off)**
11. **Other**:
    - **EX: Existential there (e.g., there is a problem)**
    - **FW: Foreign word (e.g., bonjour)**
    - **LS: List item marker (e.g., 1., 2., 3.)**
    - **NN: Noun, singular or mass (e.g., car)**
    - **SYM: Symbol (e.g., $, %, &)**
    - **TO: To (e.g., to go)**
    - **VBG: Verb, gerund or present participle (e.g., running)**