Reza Shokrzad

March 2025

NLP Workshop

Session 1 - NLP Basics & Text mining

Content

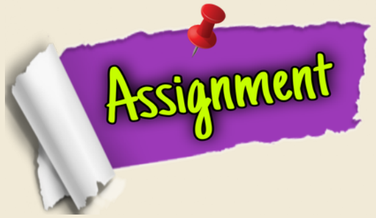
- Course Structure
- Where are we?
- What is NLP?
- NLP Challenges
- Text Data
- Preprocessing
- Regex
- Python Packages

Course Structure

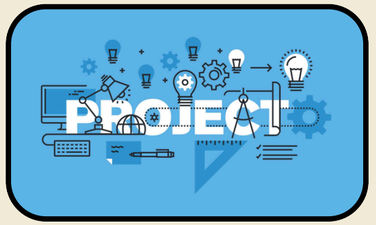
YouTube Workshops Plan (NLP) | Reza Shokrzad

1 

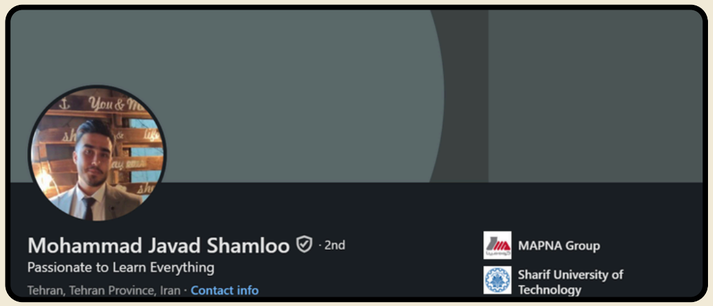
2 

3 

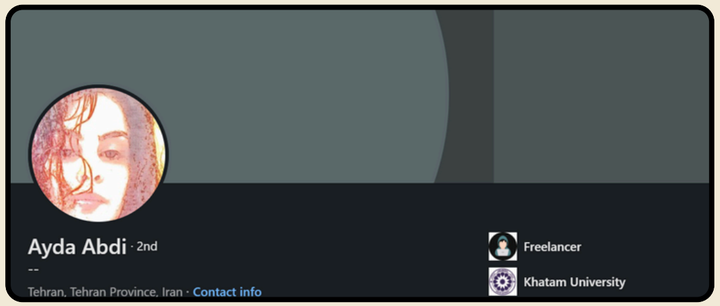
4 

5 

TAs


Mohammad Javad Shamloo · 2nd
Passionate to Learn Everything
Tehran, Tehran Province, Iran · Contact info
MAPNA Group
Sharif University of Technology


Roqayeh Mohajeri · 1st
Aspiring Data Science | Machine Learning Trainee | Biotech MSc
BNUT - Babol Noshirvani University of Technology

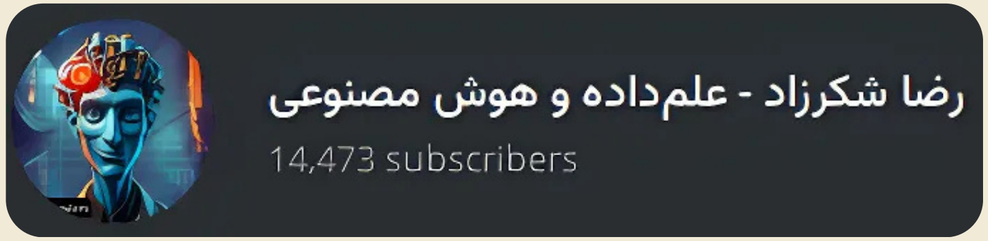

Ayda Abdi · 2nd
--
Tehran, Tehran Province, Iran · Contact info
Freelancer
Khatam University


Hamidreza KAZEMI · 2nd
Enthusiastic about Data Science and Business Intelligence| IUST student
Tehran, Tehran Province, Iran · Contact info
Iran University of Science and Technology

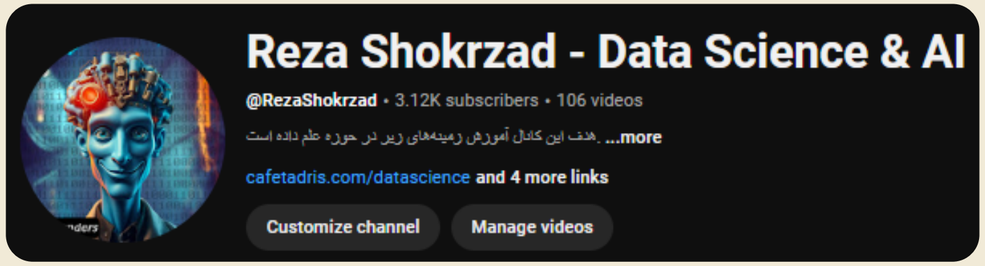

#OPENTOWORK
Afsaneh Shamsaddini · 1st
Artificial Intelligence Engineer /Machine Learning/Python Developer
Kerman, Kerman Province, Iran · Contact info
Shahid Bahonar University


M
Mohammadmahdi Rouholamini · 2nd
--
Shahid Beheshti University

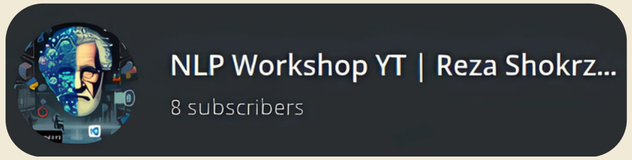
Communities



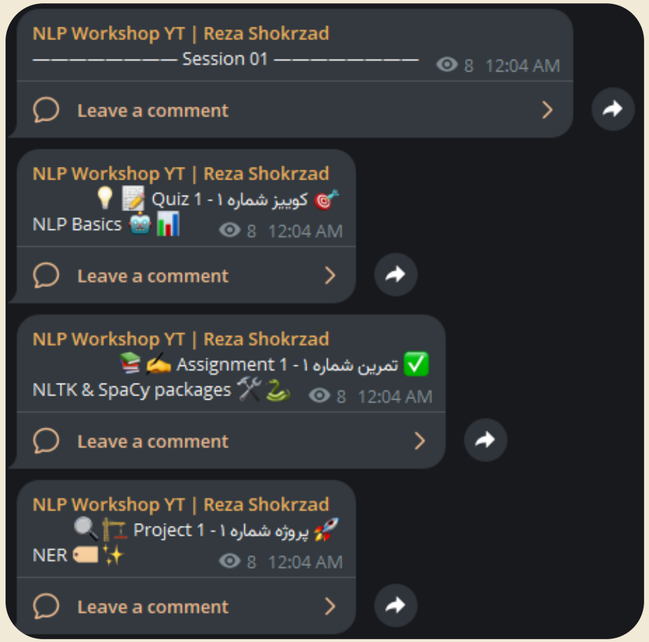
@DSLanders



@RezaShokrzad



Workshop Channel



Where were we?

Can LLMs Generate Novel Research Ideas?

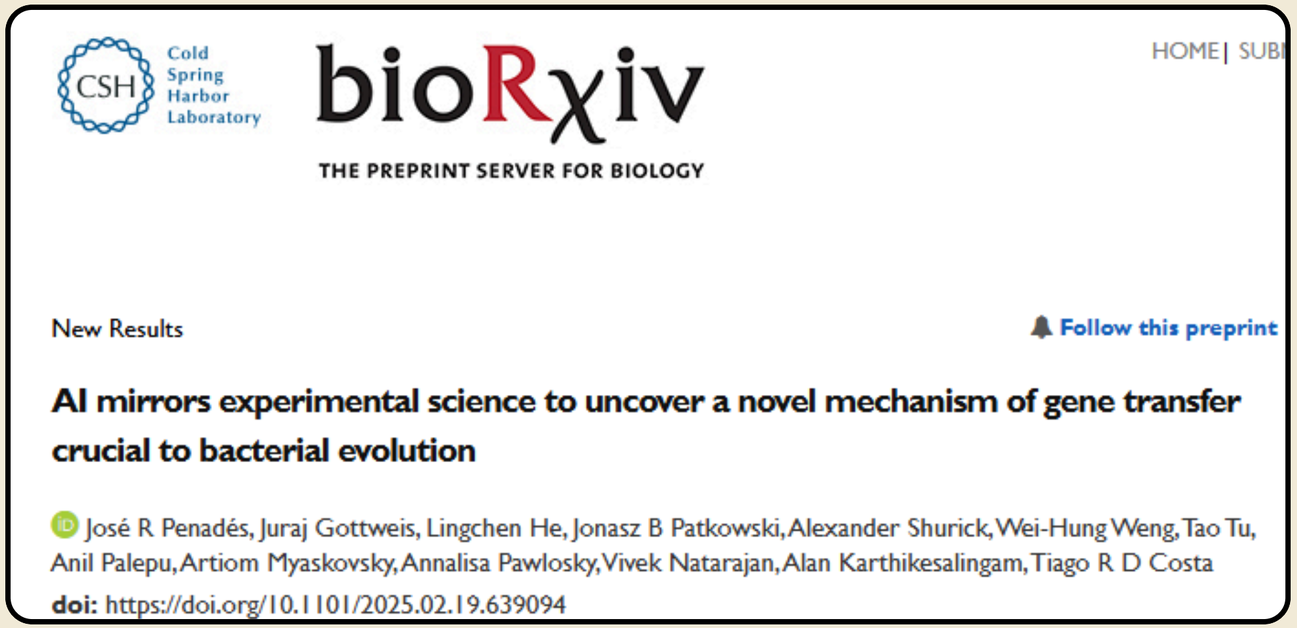
A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Sep 2024

https://arxiv.org/abs/2409.04109

Where are we?



February 2025

Paper: https://www.biorxiv.org/content/10.1101/2025.02.19.639094v1.article-info

News: https://longportapp.com/en/news/229180857

Natural Language Processing (NLP)

NLP enables computers to interpret, generate, and interact with human language through computational techniques.

NLP Challenges

- Ambiguity & Context Sensitivity
- Cultural & Linguistic Nuances
- Handling Massive Datasets
- Continuous Innovation for Greater Accuracy & Relevance

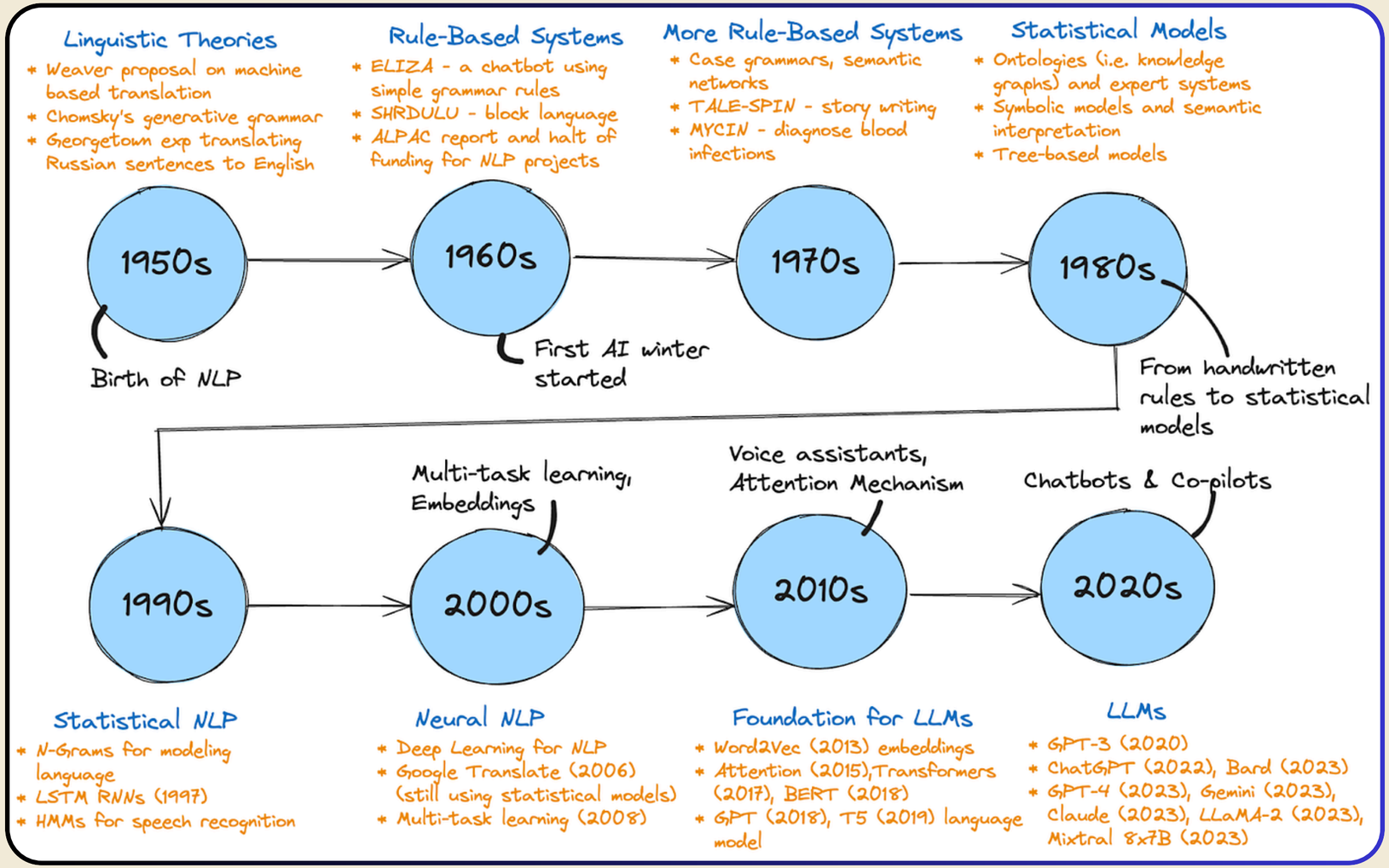
Words like "bank" (river vs. financial institution)

Flying planes can be dangerous

Idioms such as "break a leg"

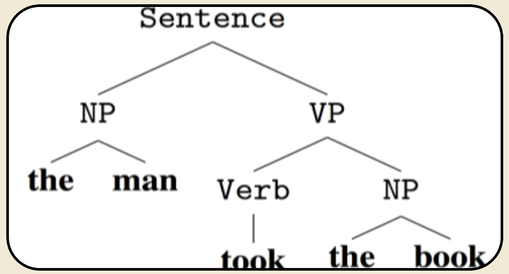
She's on fire

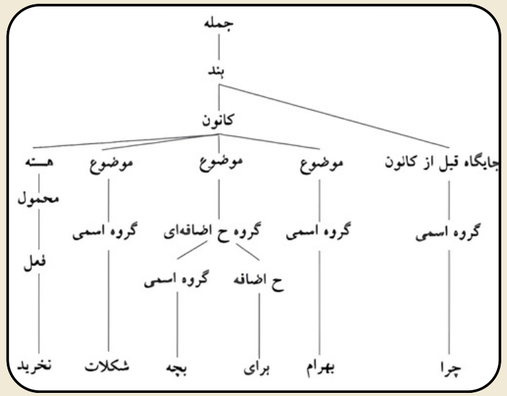
NLP Timeline

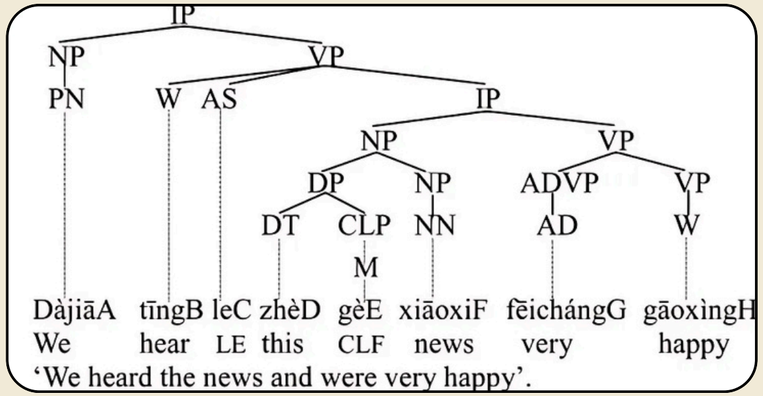


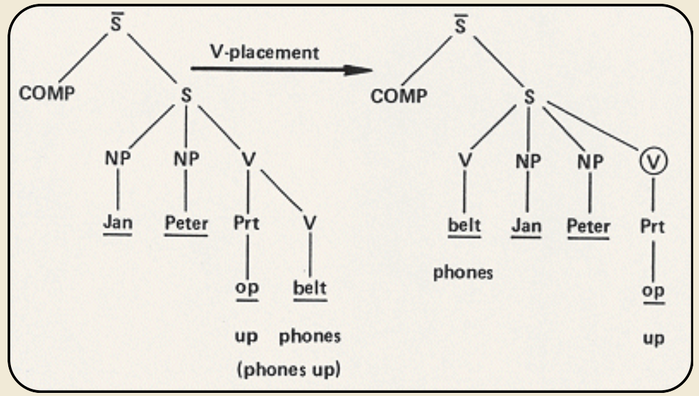
Noam Chomsky Contribution

- Pioneered generative grammar, introducing tree-based syntactic representation.
- Developed phrase-structure rules and transformational approaches.
- Revolutionized cross-linguistic parsing and modern syntactic theory.









Text Dataset

Corpora

Corpus

Document

Token

- **Brown Corpus:** Foundational American English dataset.
- **Penn Treebank:** Syntactically annotated text collection.
- **British National Corpus:** Comprehensive modern British English.
- **Wikipedia Corpus:** Vast, diverse contemporary text.
- **Google Books Ngram Corpus:** Historical language trends via n-grams.

Text

An unstructured data ...

```
dataset['Reviews'].iloc[2]
```

```
"Hello e-v-e-r-y-o-n-e!!!@@@@!!!!!! 😊 @DONT BUY THIS PHONE at all-first of al
l that say the phone in new , i took it to the lab after 6 month the phone is d
ead dead ,you can save itthay open the phone in the lab and say!!!!the phone is
renew ,and its cheepe commponents.I payed 400$ for only 6 month ,now i need to
buy new one this LG G4 is dead .not nice 'people say to me dont buy from http
s://www.amazon.com/gp/aw/d/B01GYUDMFY/ref=ya_aw_od_pi?ie=UTF8&psc=1 at al!!!"
```

... needs preprocessing

Preprocessing

1. Lowercasing
2. Punctuation & Special Character Removal
3. Tokenization
4. Stop-Word Removal
5. Stemming and Lemmatization

https://en.wikipedia.org/wiki/Text_processing

1. Lowercasing

- Convert text to lowercase.
- Unify case for consistent tokens.
- Reduce vocabulary redundancy.

Pizza
pizza
PIZZa
PIZZA
piZza

pizza

2. Punctuation & Special Character Removal

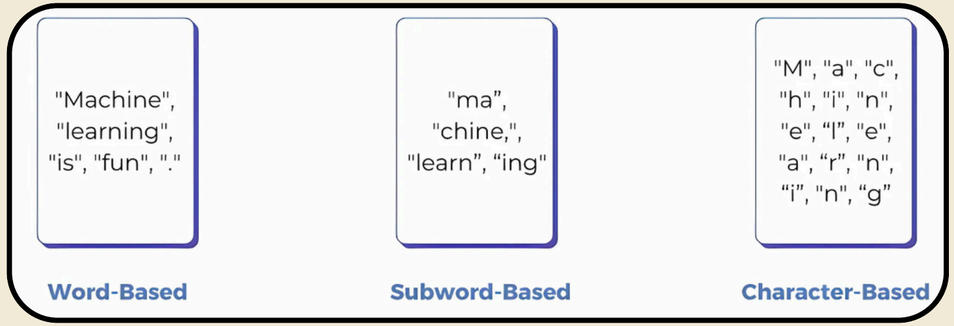
- Strip punctuation from text.
- Remove special characters systematically.
- Eliminate non-alphanumeric symbols.
- Reduce noise by cleaning marks.

**Wow!!! Visit https://example.com now – save 30%
 #Deal @limitedTime**

wow visit examplecom now save 30 deal limitedtime

3. Tokenization

- Tokenization converts text into manageable units (tokens) for easier processing by NLP models.
- **Types:** Includes word, subword (e.g., Byte-Pair Encoding), and character-level tokenization, each offering different granularity.
- **Impact on Model Performance:** The choice of tokenization affects model size, training time, and accuracy, influencing how models understand text.
- **Handling Special Characters:** Proper tokenization accounts for punctuation, contractions, and special characters to maintain context.
- **Language-Specific Considerations:** Tokenization varies across languages (e.g., word boundaries in English vs. character-based in Chinese).



https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization

4. Stop-Word Removal

- Eliminate common stop words.
- Remove non-informative terms.
- Reduce dataset vocabulary.
- Enhance semantic clarity.

Stopwords		
a	it	these
about	its	they
again	itself	this
all	just	those
almost	kg	through
also	km	thus
although	made	to
always	mainly	upon
among	make	use
an	may	used
and	mg	using
another	might	various
any	ml	very
are	mm	was
as	most	we
at	mostly	were

1	
2	و
3	در
4	به
5	از
6	که
7	مي
8	اين
9	است
10	را
11	با
12	هاي
13	براي
14	آن
15	يك

https://en.wikipedia.org/wiki/Stop_word

https://github.com/kharazi/persian-stopwords/tree/master

5. Stemming and Lemmatization

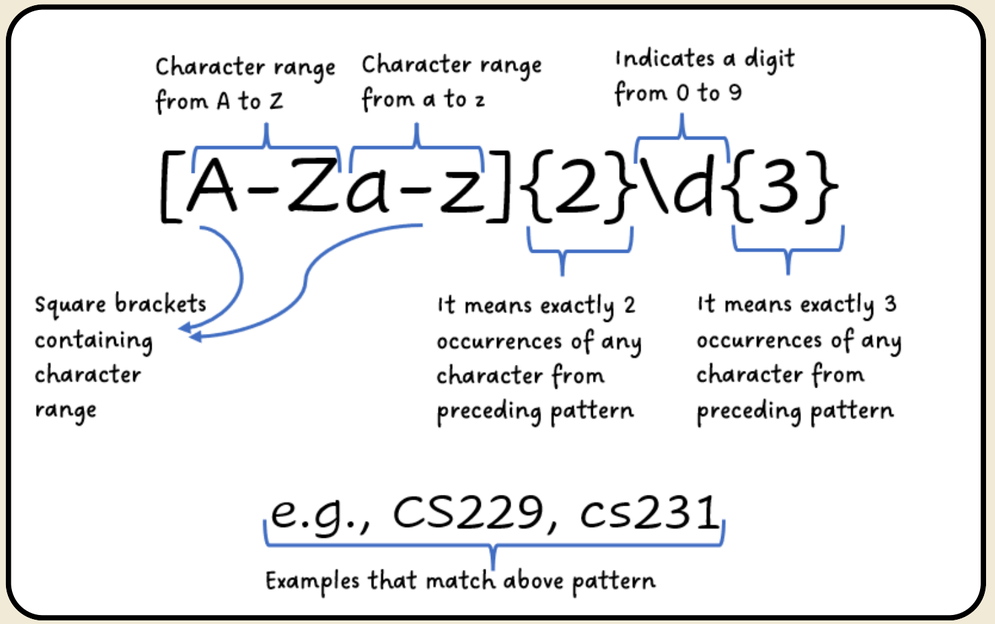
- Remove affixes (stemming).
- Map words to base forms (lemmatization).
- Reduce vocabulary diversity.
- Enhance text consistency.

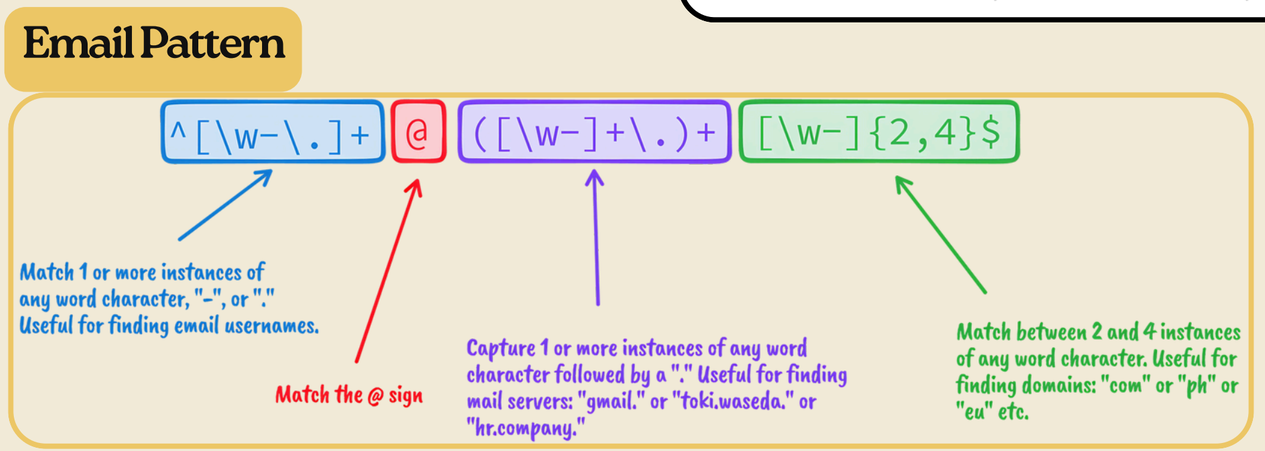
https://en.wikipedia.org/wiki/Stemming

https://en.wikipedia.org/wiki/Lemmatization

Regular Expression

- Match specific text patterns.
- Extract relevant substrings.
- Replace or remove patterns.
- Filter unwanted text noise.

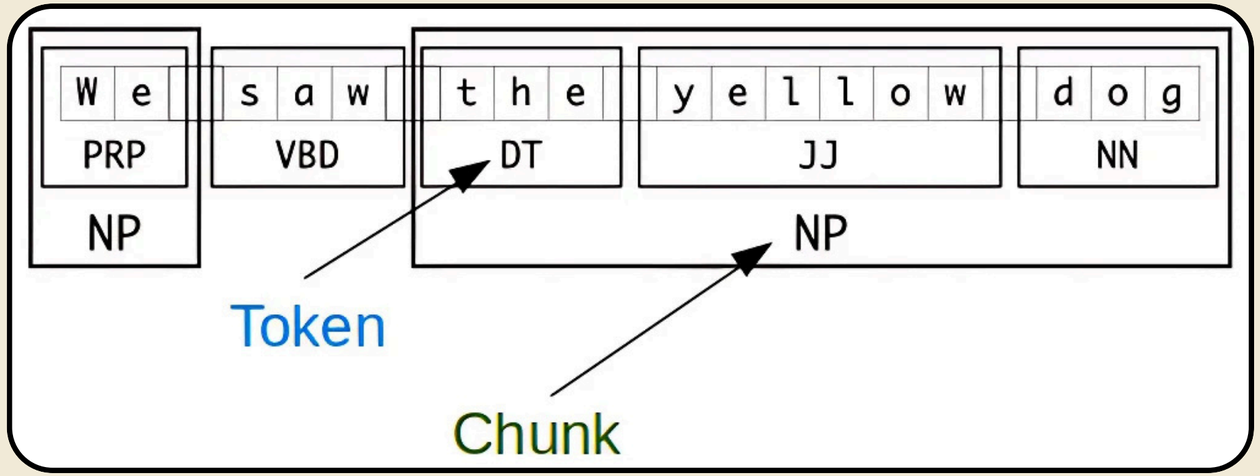




https://en.wikipedia.org/wiki/Regular_expression

Chunking and Parsing

- Group words into meaningful chunks.
- Identify phrase boundaries.
- Analyze grammatical structure.
- Reveal sentence hierarchy.



https://en.wikipedia.org/wiki/Shallow_parsing

POS tagging

- **Noun (N)**- Daniel, London, table, dog, teacher, pen, city, happiness, hope
- **Verb (V)**- go, speak, run, eat, play, live, walk, have, like, are, is
- **Adjective(ADJ)**- big, happy, green, young, fun, crazy, three
- **Adverb(ADV)**- slowly, quietly, very, always, never, too, well, tomorrow
- **Preposition (P)**- at, on, in, from, with, near, between, about, under
- **Conjunction (CON)**- and, or, but, because, so, yet, unless, since, if
- **Pronoun(PRO)**- I, you, we, they, he, she, it, me, us, them, him, her, this
- **Interjection (INT)**- Ouch! Wow! Great! Help! Oh! Hey! Hi!

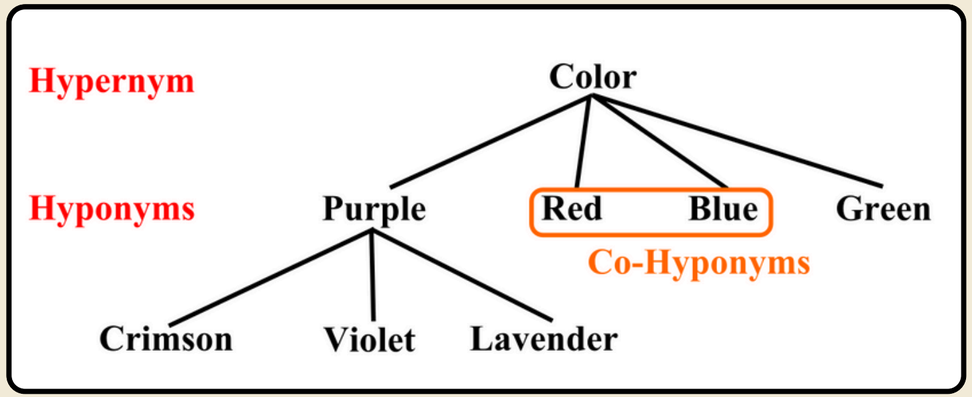
Why POS tagging?

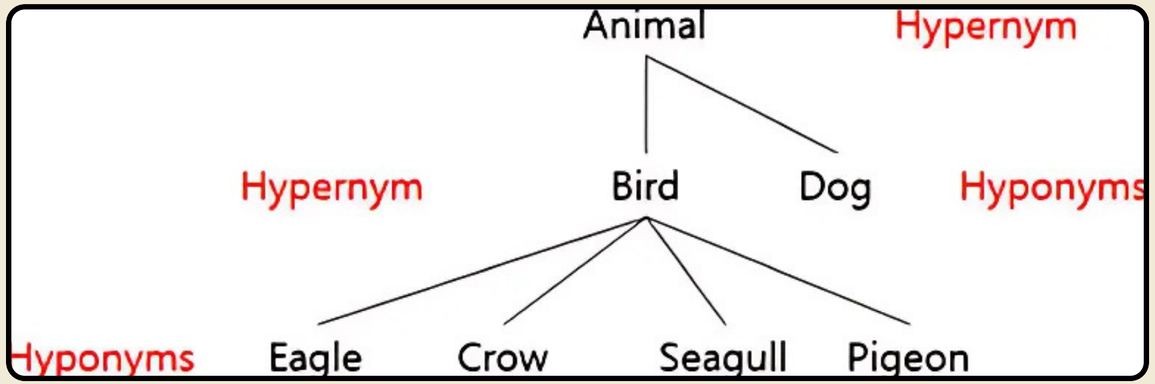
- Clarifies grammatical roles
- Aids in disambiguation
- Enhances downstream NLP tasks

https://en.wikipedia.org/wiki/Part-of-speech_tagging

Hypernyms and Hyponyms

- Identify broad category (hypernym).
- Recognize specific instance (hyponym).
- Enhance semantic hierarchy.
- Improve lexical organization.





https://en.wikipedia.org/wiki/Hypernymy_and_hyponymy

NLTK

- Comprehensive NLP toolkit.
- Supports tokenization, tagging, parsing.
- Extensive corpora and lexicons.
- Widely used in education and research.

https://en.wikipedia.org/wiki/Hypernymy_and_hyponymy

spaCy

- Fast, efficient NLP library.
- Robust tokenization, parsing, NER.
- Pretrained language models included.
- Industrial-Strength library.

https://spacy.io

https://en.wikipedia.org/wiki/Hypernymy_and_hyponymy

WordNet

- Lexical database for English words.
- Organizes words into synonym sets.
- Maps semantic relationships (e.g., hypernyms, antonyms).
- Supports NLP tasks like similarity analysis.

https://wordnet.princeton.edu

https://en.wikipedia.org/wiki/WordNet

جلسه ۱ - NLP (YTW) ← نصر از کتاب Jurafsky

Semantic معنایی

Syntactic ساختاری (structrol)

Sentiment احساس

Task) NER : Name Entity Recognition یعنی اسم موجودیت‌ها

↓

Apple is an American Company.

lower → apple is an american Company.

Token: a, p, l, e, i, S, ..., 1, 2, ...,

Char_level → _, -, !, ...

Subword_level → apple, is, an, america, #n, ...

word level → apple, is, an, american, ...

Learning learnable learned learnt

establish

establishment

in establish

in establish ment

⋮

assessment

assortment

⋮

establish + ment + in

Computation power ≡ hardware { Processor, RAM, Disk }

26 (a-z)

a9 (10) → 200

!?# --- (...)

Stemming

ریشه‌یابی کلمات در زبان

morphology

active
action
activity
act
$\xrightarrow{\text{stem}}$ act

Lemmatization

{ good, better, best, well, ⋮ } $\longrightarrow$ good

{ go, went, gone } $\longrightarrow$ go

Regular Expression (Regex)

Python $\xrightarrow{}$ import re

عبارت منظم

+ : one or more — a+ : "a", "aa", "aaa", ...

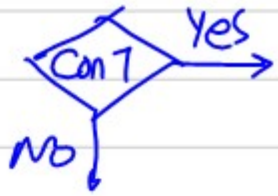
→ astrisk

* : zero or more — a* : "", "a", "aa", ...

^ (caret) → شروع رشته و نوشتن char

? → انتهای عبارت

Intelligence
1. Rule_based (NLTK_spaCy)



2. Machine learning

- Sup (label)
- UnSup
- Self_Sup
- Reinforcement Learning