Followers

Python, Stats, SQL, ML, MLProject

Computer-Vision

Natural Language Processing

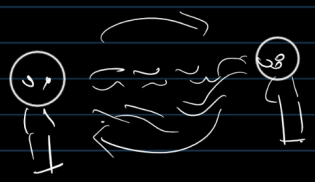
Agenda.

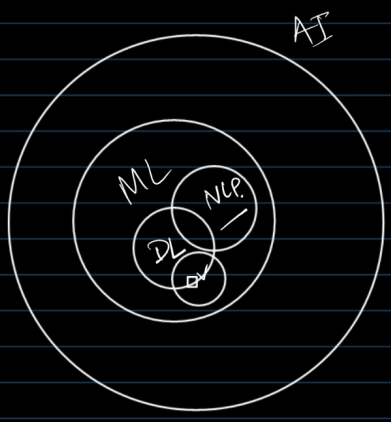
→ Overview and history of NLP.
→ Use of NLP
→ Text processing, Regex.
→ Text Normalization
→ lemmatization & stemming
→ Word embedding
→ Word 2 Vec and doc2 Vec.
→ NLtk & Spacy, TextBlob, Stanford NLP.

* NLP

↓

Natural language Processing





Part of AI that enables machines to Understand, interpret, generate and respond to human language.

✓ Ram & Shyam is going to temple. He will go movie after that.

Animal

cat dog.

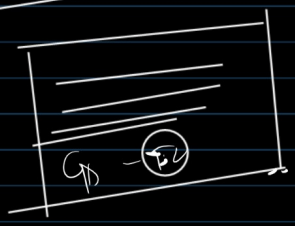
Motivation

- Chat Bots
- auto correction
- document Summarization
- translation
- Imaging → 
- Search engine
- Chat GPT
- Personal Assistants
- Medical health records
- Sentiment Analysis

Zoho

Recommendation

Lallantop → Fact check



History of NLP

1950's → during cold war → Machine translation

↳ Turing test → Alan Turing → Proposes the idea of Machine that can exhibit intelligent behaviours.

1954 → George town IBM Experiment → 60 russian sentence to English.

1980's → POS tagging.

↓

Statistical NLP

2010 → deep learning

Word embedding, RNN, LSTMs, Transformers (Bert ; GPT)

2020 → LLM → GPT3, GPT4, Grok 4

Few Terminologies / Concepts

① NLP → Natural Language Processing

② Phonetics and Phonology —
↳ Concerned with how sounds function

③ Morphology → It studies the structure of words

unhappiness

un (Prefix) + happy (root word) + ness (suffix)

④ Semantics
↳ literal meaning of words, phrases, sentences
↳ How meaning is constructed from words.

I went to the bank. (इसमें)

river bank

⑤ Pragmatics
↳ how context affects the meaning.

Can you explain me Normal distribution ??

↳ question, request

⑥ Corpus → large collection of text documents.

Corpus: ["Ajay Devgn is good friend", "this is not good", "---", "---"]

⑥ Document. → An individual piece of text within corpus

→ "Ajay Devgn and Ravi Kishan is friend"

⑦ Vocabulary

↳ Set of unique words without duplicates.

Ram is going to Delhi. He goes everyday to, is, he]

↳ [Ram, is, going, to, delhi, he, everyday]

⑧ Word (Tokens) → Individual component of sentence.

I am going

Task in NLP

① Spell check → detect and correct spelling error

② keyword based information retrieval →

Process of finding relevant information from a large collection of documents.

③ Topic Modelling

{ "The Government pass a law to regulate salary"
"Nvidia is releasing the most powered TPU"
"Tesla released GROK".

↳ Topic 1 → Politics → [Government, law, regulate]

" 2 → Technology → [Tesla, Nvidia, TPU, GROK]

④ Text classification

Sentiment Analysis → +ve, -ve, Neutral

Spam detection

News Categorization → Politics, Sports, Tech.

Language detection → French, English

⑤ Information Extraction :-

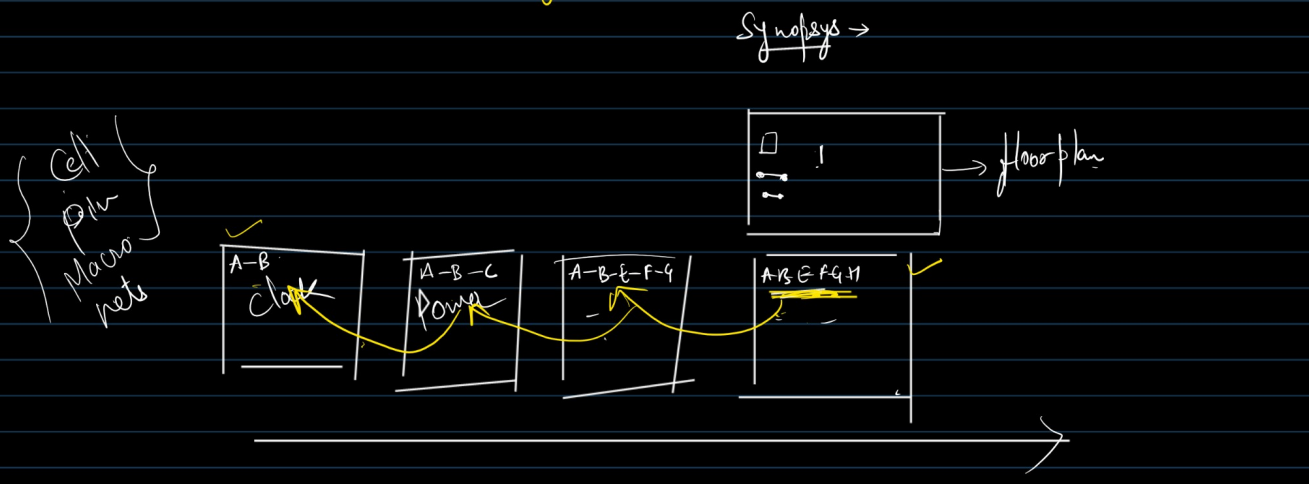
↳ Process of automatically extracting structured information like names, dates, relationship, events from Unstructured data.

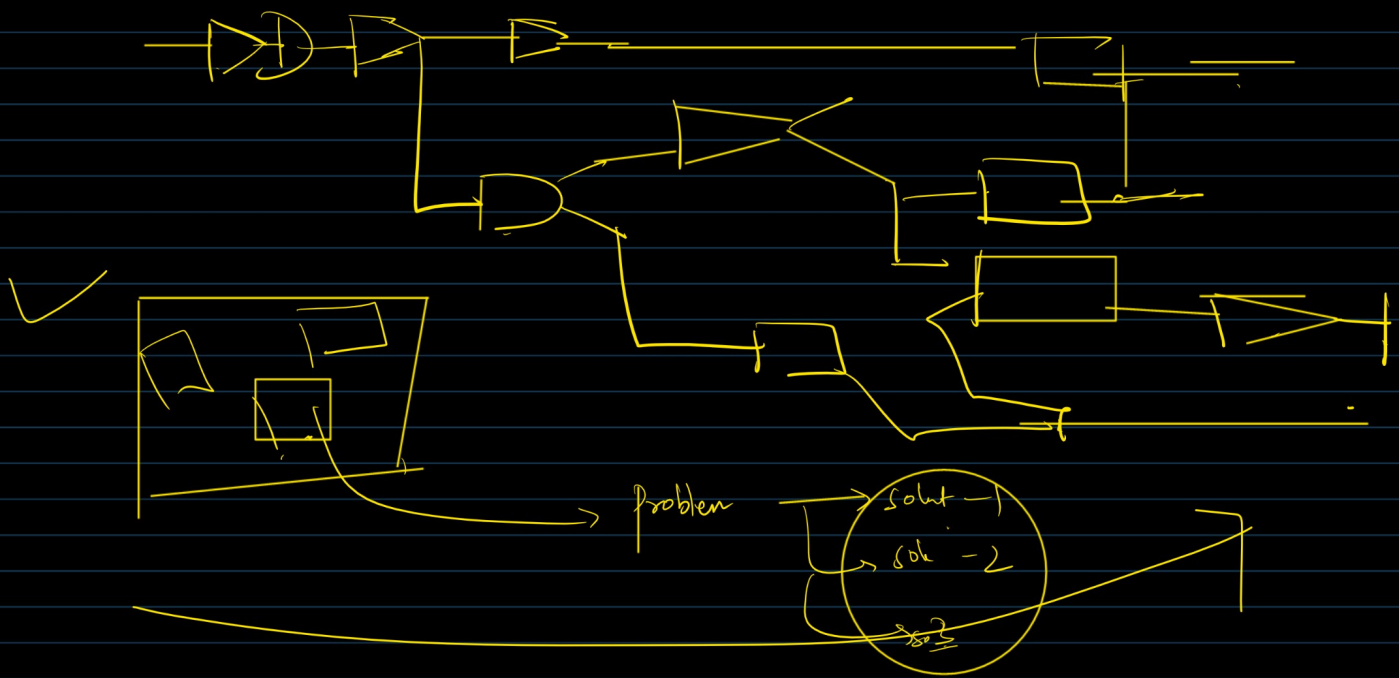
* → Named Entity recognition :-

Ajay was born in Bhagalpur

(Person: Ajay
"location": Bhagalpur)

⑥ Closed Domain Conversational Agent





Banking → What is my account balance? Transfer 500 to Utkarsh

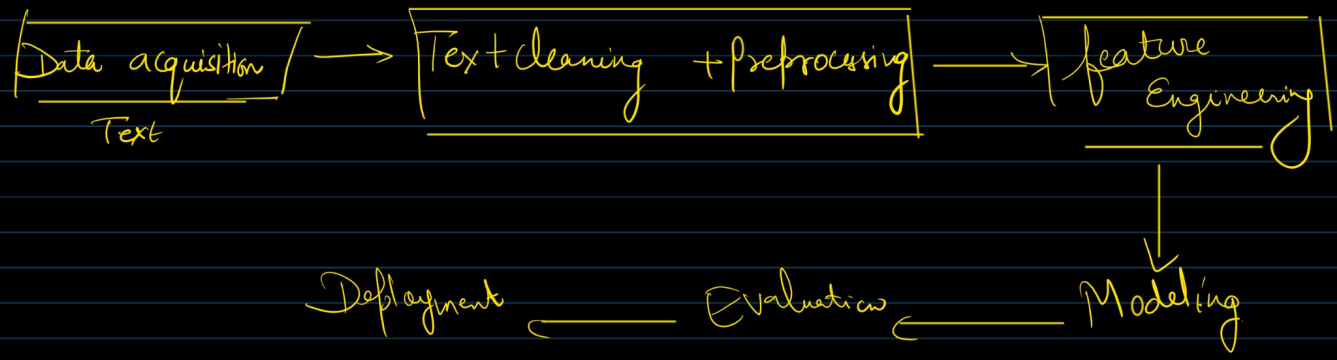
Healthcare → image → diagnosis → disease → solution

⑦ Text Summarization

⑧ Question Answering

⑨ Translation

ML → CRISP framework



① data acquisition

↳ collection of textual data
- → web scraping
- → User input
- → data repositories
- → API's
- → business team

② text cleaning

"I'm going ☺ to DELHI TO Morrow I afte r oon."

I am. ✗ Caps. Space? in Extra space (n)

→ once data is acquired, it has noise → (punctuation, extra space, Caps, incomplete words) missing value

③ Text processing

After cleaning, the text/data undergoes for pre-processing for modelling.

① Tokenization → (A small amount is token)

"I am Ajay"

Tokenization of Sentence: ("I", "am", "Ajay")

Tokenization of word: "Ajay"

["A", "j", "a", "y"]

Tokenization of corpus → "I am Ajay. I am learning ds"

"I am Ajay", "I am learning ds"

Tokenization → Tokenization is the process of breaking a text string into smaller meaningful units called tokens.

The tokens can be

→ Sentences
→ Words
→ Subwords
→ Characters.

② Removing punctuation & Special Character

I don't play $, @, ?

③ Stemming & lemmatization

I love reading → read
I love eating → eat
He goes to market daily

goes → go

good better best → good

④ Stop word removal

→ eliminating very common words → to, the, a, an, and, is

I love going to delhi

⑤ N-gram

→ A contiguous sequence of n items from given text.

items: words, characters

I love NLP → Unigram → "I", "love", "NLP"

" → Bi-gram → ("I love", "Love NLP")

" → Tri-gram → ["I love NLP"]

* In text preprocessing → you would like to understands :→

ⓐ word context (which words appear together)
ⓑ Phrase pattern (frequent phrases or expressions)
ⓒ Probabilities of sequences (LLMs)

A language model estimates the prob of a sequence of words:

$$P(w_1, w_2, w_3 \dots w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|(w_1, w_2)) \cdots$$

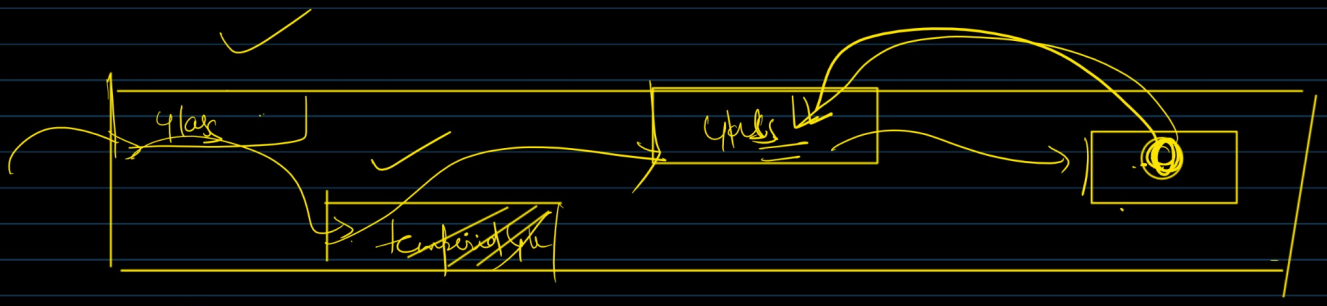
↓

Markov Assumption

↓

n−1 words matter

$$P(w_3|(w_1, w_2)) \approx P(w_3|w_2)$$



⑥ POS → noun, verb → structure of sentence.

Feature Engineering

↳ preprocessed text to number

① TF-IDF
② Word Embedding
③ N-Grams → talked above

modelling

Demonstration

Mai Jaata Hun

$x_1\ x_2\ x_3\ x_4\ x_5 \dots x_{100}$ | y

corr (x_1, y) — +0.8 — x_1
corr (x_2, y) — −0.9 → x_2
corr (x_3, y) — 0.6 — x_3
→ x_4
— 0.01