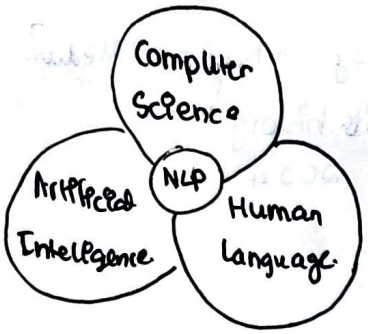
NATURAL LANGUAGE PROCESSING



NLP is a subfield of linguistics, computer science & AI concerned with interactions b/w computers & human language. In particular how to program computers to process & analyse large amounts of natural language data.

Need for NLP

- ★ Natural language ★ evolved naturally in humans through use & repetition
- ★ Ordinary langue.
- ★ without conscious planning or premeditation.

REAL WORDD APPLICATIONS

- ★ Contextual Advertisements
- ★ Email clients – spam filtering, smart reply.
- ★ Social Media - removing adult content, opinion mining
- ★ Searches Engines.
- ★ Chatbots

Common NLP Tasks

- ★ Text / document classification (News)
- ★ Sentiment Analysis (Reviews) (customer reviews)
- ★ Information retrieval.
- ★ Parts of Speech Tagging (noun, verb, adjective).
- ★ Language Detection & Machine translations
- ★ Conversational Agents.
- ★ Knowledge graphs & QA systems (who is prime min of India?)
- ★ Text summarization (In Shorts)
- ★ Topic modelling
- ★ Text Generation (Keyboard)
- ★ Spell Checking & Grammar correction
- ★ Text parsing
- ★ Speech-to-Text

APPROACHES TO NLP

- ★ Heuristic Methods
- ★ Machine learning Based Methods
- ★ Deep learning Based Methods

Heuristic Method : Hard rules

Not efficient

Example

- ★ Regular Expressions
  - → removing HTML tags
  - → searching for salutation
- ★ Wordnet
  - Lexical dictionary
- ★ Open Mind common sense.

Advantages

- ★ Quick approaches.
- ★ Current scenario

Machine learning approach :

The Big advantage.

ML workflow

Algorithms used

- → Naive Bayes
- → logistic Regression
- → SVM
- → LDA
- → Hidden Markov Models

Deep Learning Approach

* Architectures used
  - RNN
  - LSTM
  - GRU / CNN.
  - Transformers (BERT
  - Auto encoders.

Challenges in NLP

* Ambiguity.

  I saw a the boy on the beach with my binoculars

  I have never tasted a cake quite like that one before.

* Contextual Words

  I ran to the store because we ran out of milk

* Colloquialisms & Slang

  Piece of cake, pulling your legs.

* Synonyms

* Irony, Sarcasm and tonal difference

  That's just what I needed today!

* Spelling Errors

* Creativity

  Poems, dialogue, scripts.

* Diversity

Assignment

1) make a blog on NLP. at Medium
   → NLP & its history.
   → 1960s to 2024

End to End NLP Pipeline

→ Set of steps followed to build end-to-end NLP software

→ NLP software consists of following steps:

★ Data Acquisition

★ Text Preparation
- Text cleanup
- Basic preprocessing
- Advance ~~processing~~ preprocessing

★ Feature Engineering
- Bag of words
- TFIDF

★ Modelling
- →Applying algorithm - Model Building
- →Evaluation

★ Deployement
- Deployement
- Monitoring
- Model update.

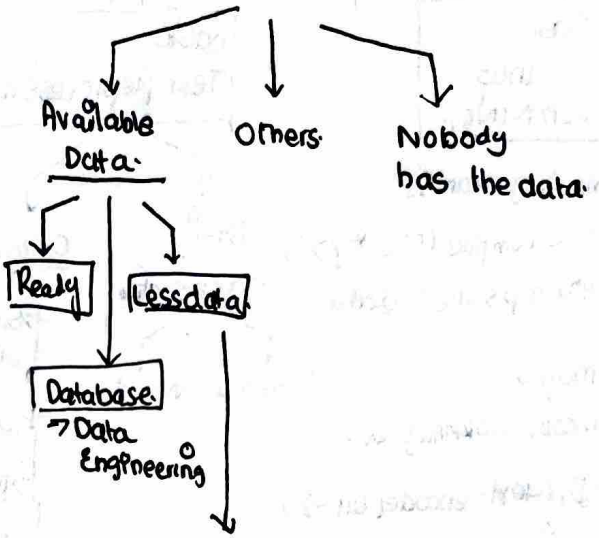
Points to remember

★ Its not universal.

★ Deep Learning Pipelines are slightly different

★ Pipeline is Non-Linear.

Data Acquisition



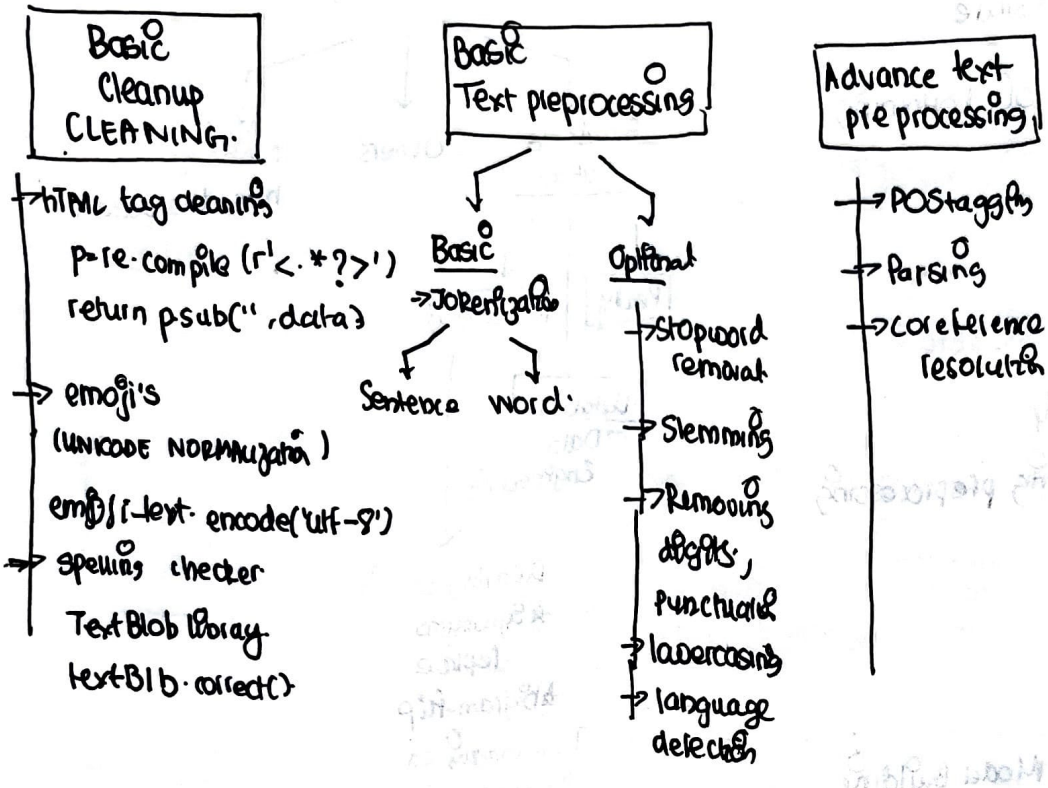
Others

→ Public dataset
→ Webscrapping (beautiful soup)
→ API
→ pdf → RapidAPI → requests → JSON.
→ Image → OCR
→ audio → speech-to-text libraries

Nobody

Text preparation

HTML tag
Hello



Tokenization

Converting text to tokens

```
from nltk.tokenize import sent_tokenize,
                          word_tokenize.
```

Optional

* Stop word removal
  and, for, the, an.

* Stemming & lemmatization

POStagging

→ Parts of speech tagging

Feature Engineering

→ Converting text to numbers.

Bag of words
↳ Tfidf
↳ OHE
↳ word2vec.

Sentiment analysis of movie reviews

	review_text	sentiment
1		0 Negative Sentiment
		1 Positive Sentiment

(50,000, 4) - Text vectorization -

#of +ve words	# of -ve words	# of Neutral words	sentiment
5	2	6	0
3	1	6	1

Feature Engineering in

ML Algo

Pre → Feature eng... → Apply Algo

Advantage.
Can justify the interpretability

Disadvantage
* Domain knowledge required to find features

DL Algo

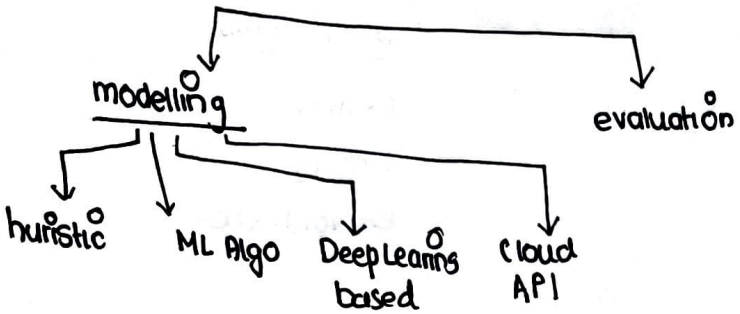
Data → Pre → Algo.

Deep learning model generates feature by itself.

Advantage
* Features are generated automatically

Disadvantage
* Cannot justify why something is working.

Modelling



Which approach to apply?
→ Amount of data
→ Nature of problem.

Email Spam classifier

Less data → heuristic approach.

Increased data → NL algo.

Huge data → DL approach

# ve	# -ve	from spammer
3	6	0
4	6	1

Cloud API : Comes at a cost

DL → Transfer learning

Eg: BERT → Trained on 40GB data on the internet

Evaluation

eval

- Intrinsic evaluation
- Extrinsic evaluation

Eg: Text generation

Next word while typing in keypad.

Metric : perplexity.

↓

Tells how confused is NLP model

Deployment

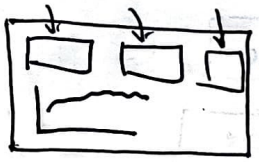
Deployment Monitoring Update

★ Deployment

→ Deploy model on any cloud service as API (microservice)

→ Chatbot

★ Monitoring



→ Dashboarding

★ Update

NLP Pipeline Questions

1. Data Acquisition
   a. From where would you acquire the data?
2. Text preparation
   a. What kind of cleaning steps would you perform?
   b What kind of text processing would you apply.?
   c. Is aavanced text processing required?
3 Feature Engineering
   a. What kind of features would you create?
4. Modelling
   a. What algorithm would you use to solve the problem at hand?
   b What intrinic evaluation metrics would you use?
   c) what extrincs evaluation metrics would you use?
5 Deployement
   a) ~~what~~ How would you deploy your solution into the entire product?
   b) How & what things will you monitor?
   c) what would be your model update strategy?