NLP

1. Data ingestion
2. EDA
3. Pre-Processing
4. Models
5. Evaluation

ML/DL → CV → image | video → collection of frames (FPS)

ML/DL → NLP → Natural Language Processing ⇒ text (sequence) → (audio)

1. What is NLP ?
2. Natural Language
3. NLP usecase | NLP task
4. NLP Application | Domains.
5. Approaches to NLP → ~~RB~~, ML, DL
6. Challanges in NLP
7. NLP Pipeline
8. GenAI → LLM

Primary, LLM

openai, mistral

usecase → Application → till deployment

Practical

RNN, LSTM, GRU, Encoder, decoder

Attention, Transformer

Bert, GPT, ...

text Processing, word embedding

What is a NLP ?

Natural language processing (NLP) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

Natural + Language + Processing

↓

medium of communication

NLP → Computer Science

→ (text) (Copora, Corpus)

Primary goal

So Machine can understand Human Language

10-15

ATM

H ⟷ H

H ⟷ Comm M. ⟷ H

Language

NLP → Processing (Natural Lan.) (text | audio)

Datascience

ML | DL

Data

- Structure data (tabular data)
  - Num
    - Dis
    - conti
  - Cat
    - ord
    - nom.
- unstructure data (images | videos | text | audio)

Regression, classification

Regression / classification

ML ⇒ Lin reg, ridge, Lasso, EN
SVM
DT
Log reg
Ensemble methods

N | C | N | C

← Domain / industry

→ tabular data.

Prediction

ANN

{Image / video} ⇒ CNN ⇒ (ANN) ⇒ image classification

DL — CNN ⇒ {Convolution, Neural Net}

- Object detection
- Object seg.
- Object tracking
- OCR
- Usecase

(ANN + CNN) ⇒ Application ⇒ Industry / Dom.

ML, ANN, CNN

NLP ⇒ field ⇒ (text + audio) ⇒ Property ⇒ Sequence

Seq mapping.

"my name is Sunny and I am a DS one Monday I am working on Genji Project and after that I will explore the RL."

Machine

Program Arm ⇒

RNN ⇒ Stacked RNN, Deep RNN, Bi-directional RNN

LSTM, GRU

NLP

Task / Usecases → Application → Industry / Domain

1. text Summarization
2. text classification / Document classification (Language Detection) (Sentiment classification)
3. text generation (Machine translation) (Sentence pred, word pred)
4. Information / keyword extractor (POS (Part of speech tagging))
5. Knowledge graph
6. topic modelling
7. Question Answering
   → {Question → Answer}
8. Speech recognization
9. Chatbot { Conversional ai }

10. NER

11. Sentence similarity

12. fill in the blanks / Spell corrector

Application

1. Spam filtering ⇒ mail → spam, ham
2. next word pred. ⇒ gmail (auto completion)
3. Inshorts (text summarization) ⇒ 30 word
4. Grammarly (spell corrector)
5. Paraphrasing (quillbot)
6. Google translator (text generation)
7. google assistant (speech to text)
8. tweet classification / sentiment analysis
9. Contextual Adv. ⇒ Insta., FB, twitter
   (text, audio)
   Process
   model
   Context
10. Chatgpt (Chatbot)
    (Conversational ai)

Puma Shoes
Nike shoes

Insta
FB

11. Search engine
    (NLP)

12. Alexa (speech to speech)
    text
    NLP

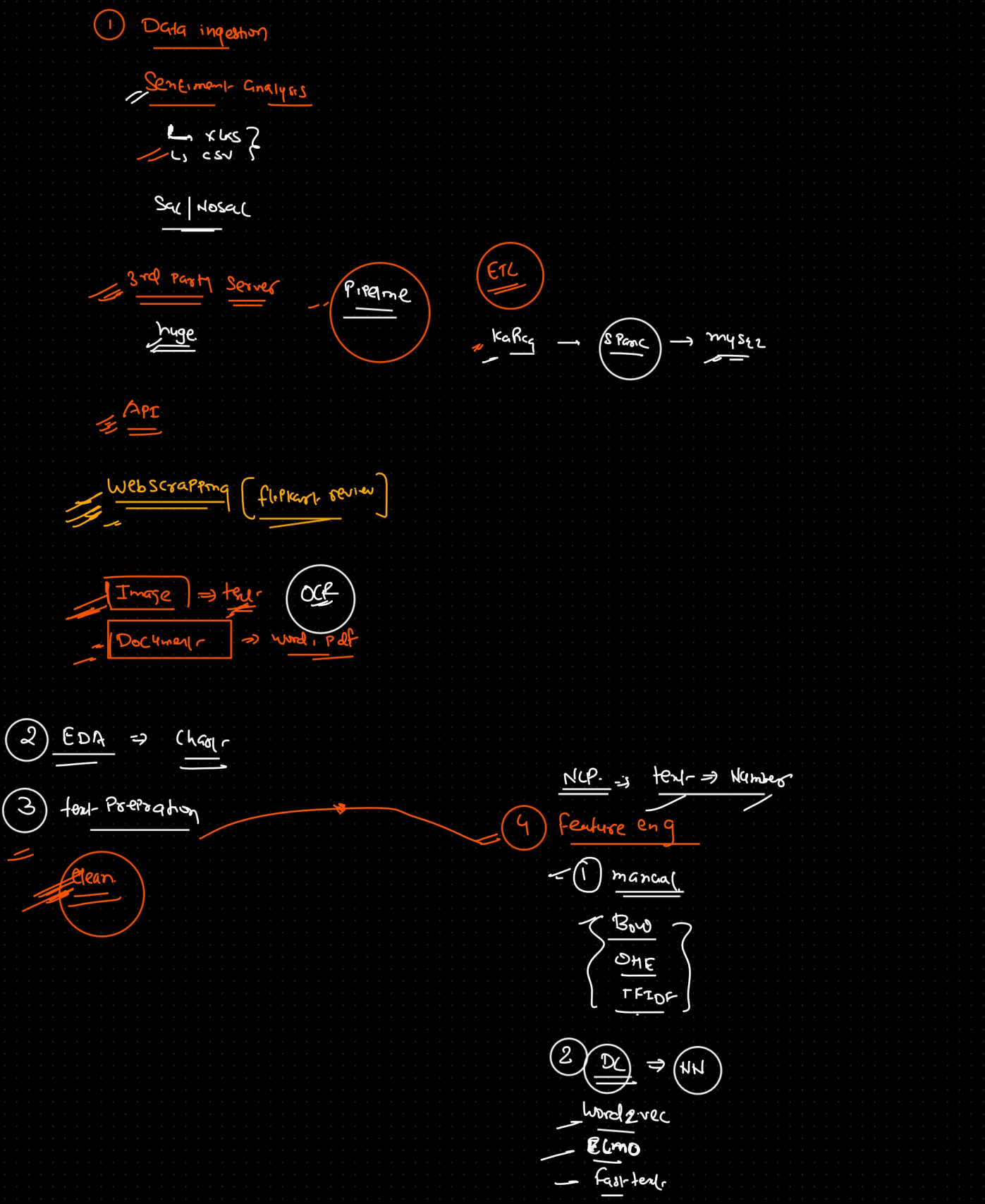
NLP Pipeline

1. Data ingestion
2. text Preparation
   - Basic cleaning
   - Basic to advance Processing

(EDA ⇒ text)

3. Feature Eng
   - Manual approach
   - DL approach
4. Model building
5. Model evaluation

6. Deployment


1 Data ingestion
Sentiment Analysis
xlxs
csv
Sql | Nosql
3rd party Server
huge
Pipeline
ETL
kafka
Spark
mysql
API
Webscrapping [flipkart review]
Image ⇒ text
Document ⇒ word, pdf
OCR
2 EDA ⇒ chart
3 text Preprocessing
clean
NLP ⇒ text ⇒ Numbers
4 Feature eng
1 manual
BoW
OHE
TFIDF
2 DL ⇒ NN
word2vec
Elmo
fasttext
5 Model building
1 Rule based approach (1950-1960)
(Heuristic approach)
2 ML
Naive bayes (Cond. Prob)
NLP ⇒ Seq mapping
Bigger Sentance ⇒ Fail

3 DL ⇒ RNN : LSTM, GRU

LLM ⇒ Base Archi. → {transformer}

Encoder, Decoder
Attention
apir
Transformer
chatgpt
↳ 2017

5 Model evaluation

1. Confusion mat (Acc, F-1score, Pre, Recall)
2. BLEU
3. Perplexity (gpt)
   ↳ Prob

Challanges

1. Ambiguity
2. Contextual words

I ran to the store because we ran out of milk

3. Slang
4. Synonyms
5. Sarcasm.

Prompt → LLM → output