NLP - Pipeline

1. Data Aquisition
2. Text preprocessing
3. Feature engineering
4. Model building
5. Deployment

Data Aquisition

Available
- Less data
  - Data augmentation method
    - n-grams
    - NER
    - POS Tagging
    - Tokenization
- Streaming
  - Shreatrive
  - Stakeholder
  - Repository
  - YYYYMMDD HH MM SS
- Database
  - Team
  - → Big data hadoop
  - → RDBMS
  - Plateform
  - Host team
  - Data engineering

Data Aquisition

(2) other source

- → web scrapping
- → API - Rapid
  - → state of the art
- → PDF
- → Audio/Video
- → public data
- → SQL / JSON / SAS / R

(3) Not available - Networking

Text Preprocessing

→ Cleaning – html tg, https, emoji, Grammer correction, Spelling mistake

→ Tokenization – Word, Sent, whitespace, Punctuation, Stopwords

→ Stemming / Lemmatization – , lower case, language detection

→ Symbols, Number digit

→ POS tagging

→ NER

run ran running run-run

go went gone going goes – go

cage = car

Feature engineering

Text → vector → Algorithm ← ML
DL (RNN) LSTM) GRU/Seq2Seq
Encoder - Decoder / Transformer
BERT / GPT - LLM

ML → Sparse ... very complex (computationally)
OHE X

→ Bag of words (BOW)

→ N-grams (uni-grams, Bi-grams, Tri-grams, Tetra gr- -etc.)

→ Term Frequency - Inverse Document Frequency (TF-IDF)

DL → Word 2 vec / Glove } Semantic meaning - Dense

happy → enjoy

Model building

ML Model

- → RF, XGB
- → Naive Bayes theorem
- → Logistic / DT / SVM / KNN etc

DL

- RNN / LSTM / GRU
- Transformer (Seq2Seq, Encoder Deco-
  - BERT
  - GPT
  - BARD

Deployment → chatBot, API, Flask, Heroku, AWS, Azure, GCP etc