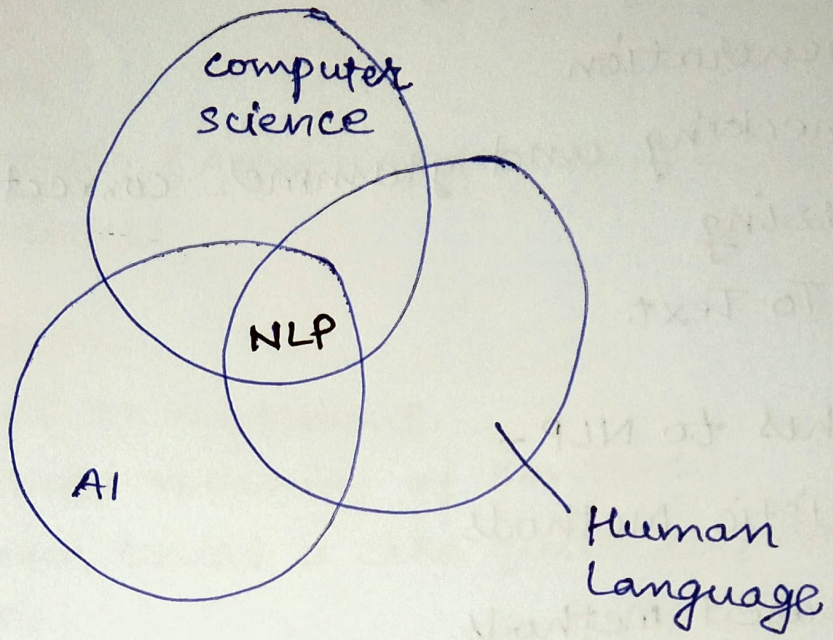
NLP

- Introduction - subfield of linguistics computer science



- Need For NLP -
  - Communicate to machine like human being.
- Real World Applications -
  - Contextual Advertisements
  - Email Clients : Spam filtering, smart reply
  - Social Media
  - Search engines
  - Chatbots
- Common NLP Tasks -
  - Text/Document Classification
  - Sentiment Analysis
  - Information Retrival
  - Parts of speech Tagging
  - Language Detection & Machine Translation
  - Conversational Agents

- Knowledge Graph and QA System
- Text Summarization
- Topic Modelling (LTA)
- Text Generation
- Spell checking and grammer correction
- Text Parsing
- Speech To Text

- Approaches to NLP -

1. Heuristic Methods
2. ML Based Methods
3. DL Based Methods

1. Heuristic Approaches -

- Select some specific word and show results on base of that word. Rule based answers.

  e.g. - Regular Expression
  - Wordnet : Relation b/w two or more word.
  - Open Mind Common sense

2. ML Approaches -

- Text Vectorization : Text → Number
- Algorithms Used :
  - Naive Bayes
  - Logistic Regression
  - SVM
  - LDA (Topic Modelling)
  - Hidden Markov Models.

3. DL Approaches –
   - Architectures used
     - RNN
     - LSTM
     - GRU/CNN
     - Transformers (Attention to some words)
     - Autoencoders

- Challenges in NLP –
  - It is difficult to implement.
  - Ambiguity : Dual meaning of sentense
    - I have never tasted a cake quite like that one before!
  - Contextual words : Different meaning of single word
    - I ran to the store because we ran out of milk.
  - Colloquialisms and slang : Different meaning of sentence
    - Piece of cake, pulling your leg.
  - Synonyms
  - Spelling Errors
  - Creativity : Poems, dialogue, scripts
  - Diversity

△ End to End NLP Pipeline -

○ NLP Pipelines - set of steps followed to build an end to end NLP software.

1. Data Acquisition
2. Text Preparation
   - Text cleanup - Remove/correct spelling, emoji
   - Basic Preprocessing - Remove punctuations
   - Advance Preprocessing - POS tagging, chunking
3. Feature Engineering
4. Modelling
   - Model Building
   - Evaluation
5. Deployment
   - Deployment
   - Monitoring
   - Model Update

- Pipeline is non-linear

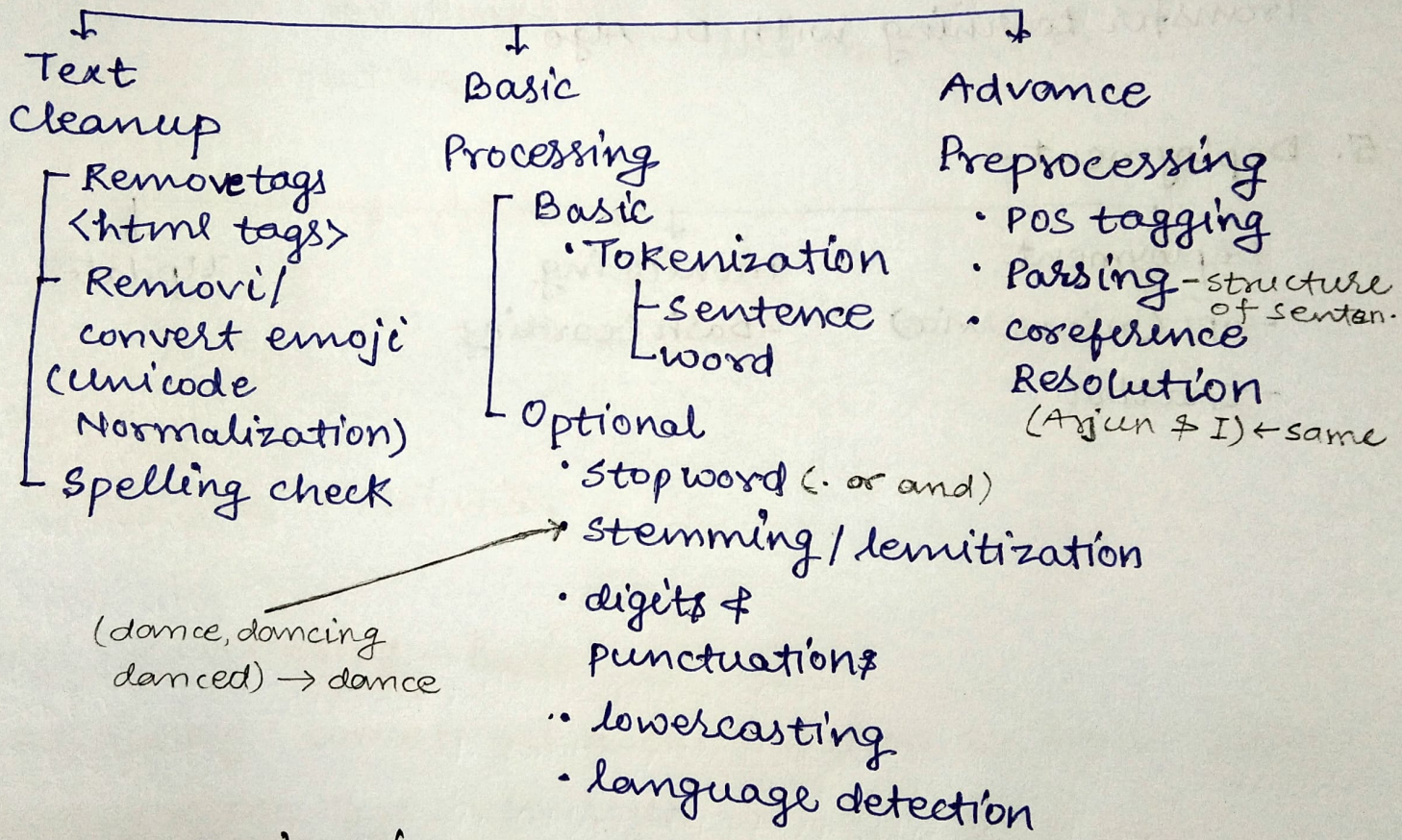
1. Data Acquisition -
   1. Data Available - table, database, less data
   2. Other (Externally Available)
   3. No data

- Data Augmentation : - Synonyms
  - bigram flip on the → the on
  - back translate lang1 → lang2
  - Add noise

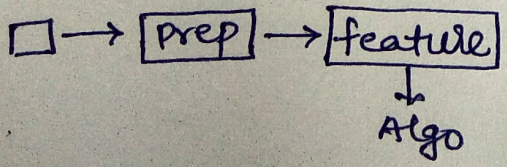
- Other (Externally data present)
  - public domain
  - web scraping
  - API – Rapidapi.com
  - PDF Format
  - Image
  - Audio – speech 2 Text

- No data

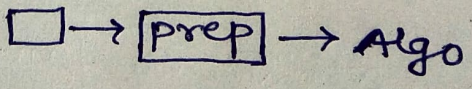
2. Text Preperation –



3. Feature Engineering –
- convert text to numbers. : Text Vectorization
- Bag of words, Tfidf, OHE, Word2Vec
- ML Algo



- DL Algo



4. Modelling –

Modelling

- Heuristic – less data
- ML Algo – sufficient Data
- DL Algo – Large data
- Cloud API – Paid API

Evaluation

- Intrinsic Evaluation
- Extrinsic Evaluation
  - Accuracy, Recal, Confusion mtx
  - Bussiness setting

- Amount of data
- Nature of problem
- Transfer Learning with DL Algo

5. Deployment –

Deployment

- API (microservice)
- Chatbot

Monitoring

- Dash boarding

Update.

Text Processing -

- Text Preprocessing → Basic
  → Advance

- Text Preprocessing - Basic :
  - Lowercasing
  - Remove HTML Tag
  - Remove URLs
  - Remove Punctuations
  - Chat word Treatment
  - Spelling Correction
  - Remove Stop words
  - Handling emojis
  - Tokenization
  - stemming
  - Lemmatization

- Text Preprocessing - Advance :
  - POS Tagging
  - Chunking
  - Parsing
  - Coreference Resolution

- Text Preprocessing - Basic
  - Lowercasing : convert all documentation to lower case
  - Remove HTML Tags : use regex
  - Remove URLs : http, https, www
  - Remove Punctuations :