NLP – Natural Language processing

Agenda:

1. Introduction to NLP
2. Application Of NLP
3. Various NLP Tasks
4. What is Language
5. Approaches to NLP
6. Challenges in NLP
7. NLP pipeline

① My name is Barry and I am a Data Scientis.....!

① There was a tiger. He was very cruel in his youth. He was a man-eater. In his old age, he became very weak. He could no longer hunt down any animal. So he suffered much for want of food. ad a gold bangle in his possession. He sat in an open place in the forest with that bangle.

Tomorrow's Agenda: (practical)

1. Text preprocessing Technique
2. Text representation
3. Word embeddings

NLP pipeline

1. **Data Acquisition:**
   - → Available Data (CSV, Txt, PDF, XLS)
   - → others Data (DB, Internet, API, Scarping)
   - → No data (create your own data)

Note: if you have less data then we Data Augmentation
→ replace with synonyms

ex: I am a Data Scientis ✓
I am a AI Engineer ✓

→ biagram flip

exp: (i) I am Bappy

(ii) Bappy is my name

→ back translate

→ add Additional noise

exp: I am a Data scientist,

I love this job → extra

(2) Text preparation:

(i) cleanup: HTML Tags, emoji, spelling correction

(ii) Basic preprocessing

(iii) Advance preprocessing

Basic preprocessing:

* Tokenization → sentense / word

I am Bappy,
I am a DS

list = ['I', 'am', 'Bappy']

list = ['I am Bappy;', 'I am a DS']

'I' 'am' 'Bappy'
[0 1 2]

* optional steps:

(1) stop word removal

(2) stemming — less use

(3) lamatization — good

(4) Removing punctuation (,!?$)

(5) lower case

(6) language Detection

[playing, played, plays]
↓
[play]

Bappy, bappy

bappy

Text summar in Eng
Bengli →

Advanced preprocessing:

(i) Parts of speech tagging

(ii) parsing

(iii) coreference resolution

(3) Feature Engineering

→ Text Vectorization

→ TFIDF, Bag of word, BoW, one Hot, word to vec

→ Encoding →

(4) Modelling

(i) Hunistic

(ii) ML

(iii) DL

(iv) Cloud API (AWS, GCP, Azure) → paid one

(5) Evaluation:

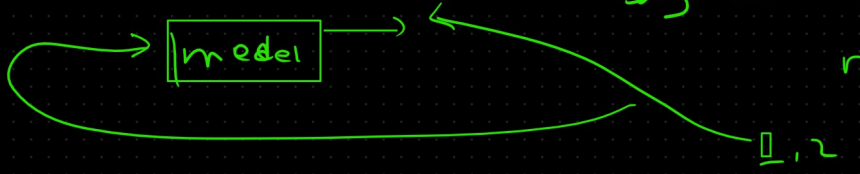
Data set → ML — classification

test → accuracy, P, R, F1, AUC, confusion

(i) Intrinsive Evaluation

(ii) extrinsive Evaluation → Host app to production

by users → A/B testing



(6) Deployment:

monitoring, retraining

common terms:

1. corpus (Entire text)
2. Vocabulary (unique word)
3. Document (one row)
4. word (simple word)