End to End GEN AI pipeline

① Data Acquisition:

→ Available Data (CSV, txt, PDF, DOCS, XLSX)

→ Other Data (DB, Internet, API, Scrapping)

→ NO Data → Create your own Data
- LLM

Note: Need — 10,000 , Collection — 5000

Data Augmentation

① Replace with Synonyms

→ I am a Data scientist

→ I am a AI Engineer

② Bigram flip

I am Bappy ←

Bappy is my name ← ←

③ Back Translation

① add Additional Data / noise

(I am a Data scientist, I love to work here)

② Data preparation :

① Cleanup : HTML, emoji, spelling, connection

② Basic preprocessing

③ Advance preprocessing →

Basic preprocessing

→ Tokenization → sentence
→ word

Optional preprocessing

① stop word removal

② Steaming —— less used

③ lemmatization — more used

④ punctuation removal (? ! , . $ @)

⑤ lower case

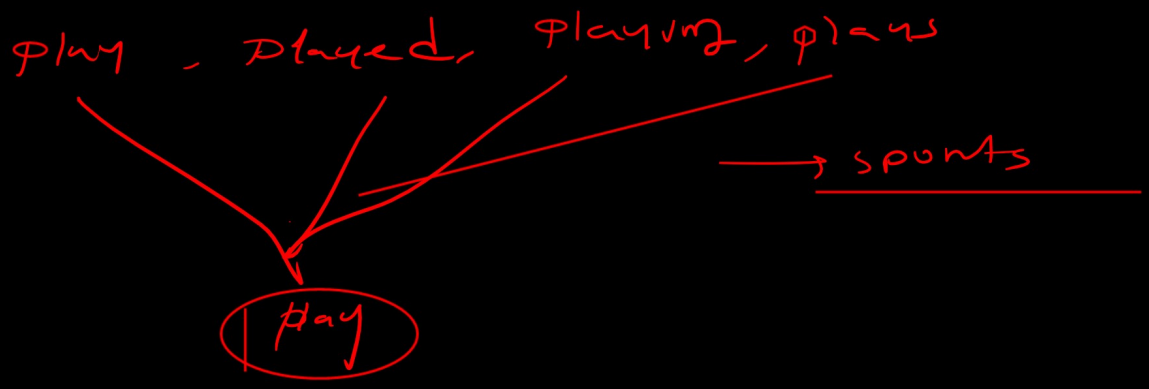
⑥ Language Detection —

My name is Bunny

word = ["My", "name", "is", "Bunny"]

My name is Bunny. I am a good boy

sen = ["My name is Bunny", "I am a good boy"]



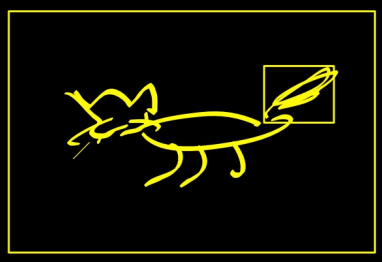
bunny is a python sentence

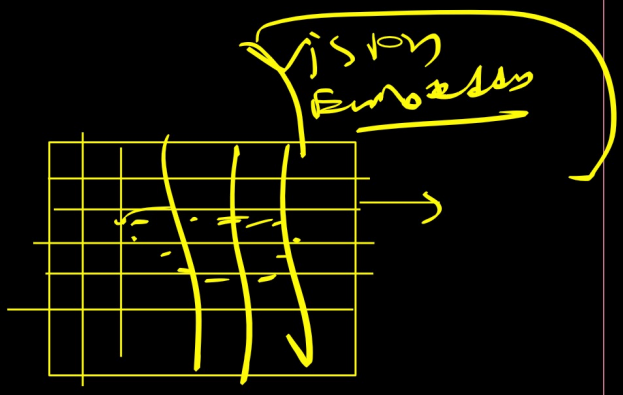
bunny is a good boy → ASCII

③ Feature Engineering

→ Text Vectorization

- → TFIDF
- → Bag of word (BOW)
- → Word2Vec
- → one hot encoding
- → Transformer Embedding →





④ Modeling

→ ① Open source

→ ② Paid model → API (OpenAI)

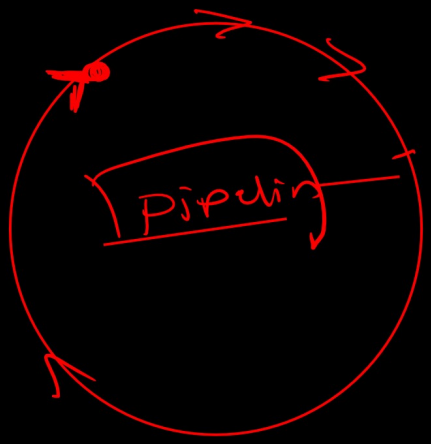
GPT

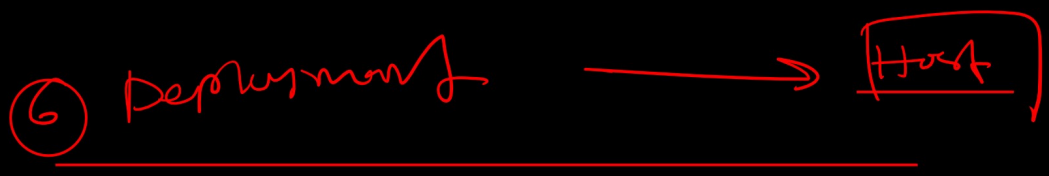
⑤ Model evaluation

① Intrinsive Eval → Metrics: P, R, F1, ...

② Extrinsive Eval →

↳ ...





AWS, Azure, GCP

common terms

1. corpus (Entire text)
2. Vocabulary (unique words)
3. Documents (single line)
4. words (single words)

Image augmentation generates random images based on existing training data to improve the generalization ability of models. In order to obtain definitive results during prediction, we usually only apply image augmentation to training examples, and do not use image augmentation with random operations during prediction.