NLP ⟹ Natural Language proccessing.

Application

* Auto complete.
* Spam / ham email filters.
* voice Assistant (google, Alexa, Siri)
* Language translation.
* chat bots.
* Document summarization.

NLP is a field of AI, that enables the machine to understand human language and assist accordingly.

Libraries:

→ spacy, Gensim, NLTK, sklearn, tensorflow, pytorch, hugging face.

Regular Expression → RegEx (python)

regex101.com ⇒ use this for practice (dynamic page)

Steps of NLP:

* Data Preprocessing - 1
  - → Tokenization
  - → stop words
  - → stemming.
  - → Lemmatization.

* Data preprocessing - 2:
  - → Bag of words (BOW)
  - → TFIDF
  - → Unigrams, Bigrams

* Data preprocessing - 3
  - → Word2Vec
  - → Average Word2Vec

* Traditional NLP models.
  - → Naive Bayes

{ Data preprocessing = Text preprocessing }

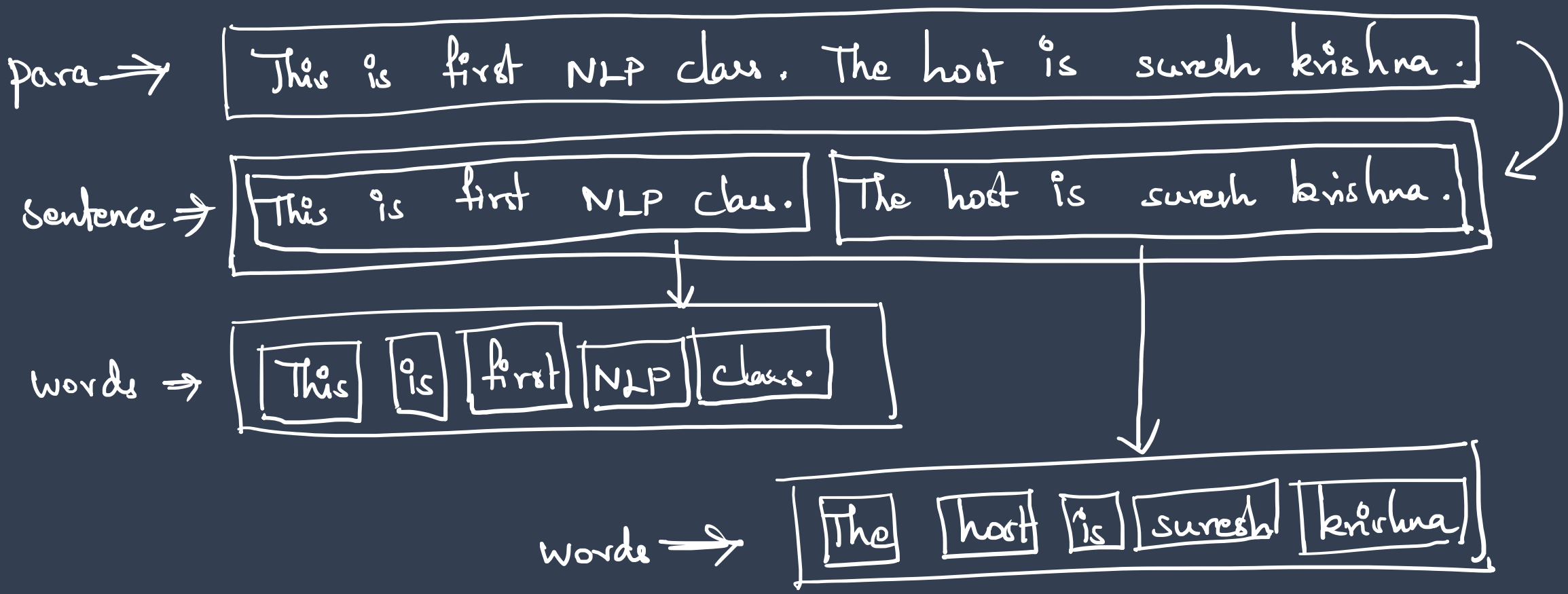
Spam / ham mail filter:

Mail subject	Mail body	class
NLP Doubts	Hi Suresh, I have a doubt...	ham
₹ hurray!! Lottery	WooW!!! You Won a lottery. Claim	spam

10,00,000 ↑ {size of the data}

Data preprocessing - 1 :

① Tokenization — split & convert paragraph ⇒ sentence ⇒ words.



⑥ stopwords:

⇒ Technique by which we eliminate the words that doesnot contribute to the prediction.

⇒ NLTK has predefined set of stopwords, but we can create our own stopwords.

⇒ Ex: How (is) (the) climate today ?

↳ stopwords

⑦ stemming ⇒ proccess of finding the root word.

programming, program, programs, programmed, programmer

↳ Rootword = Program

achieve, achieving, achieved

↳ Rootword = Achiev

↳ This doesn't make sense

} disadvantage of stemming

Advantage ⇒ Faster in computation.

(*) Lemmatization ⇒ overcomes the disadvantage of stemming.

{ Uses the dictionary of predefined words from the library to compare and find the meaningful word from its library. }

Advantage → It gives meaningful root word from dictionary

disadvantage → slow in terms of computation.

Application of stemming & lemmatization:

stemming	lemmatization
→ spam / ham.	→ language translation.
→ sentiment / review classification.	→ text summarization.
	→ chat bots.