7 Jan '25

• Generative AI

NATURAL LANGUAGE PROCESSING

- subfield of AI that uses ML to understand & recognise human language.
- enables digital system to understand / recog and also generate text & speech by using computational linguistics and statistics / ML / DL.
  - └ techniques / rule to understand language.

- Generative AI
  - LLM / Img recog or ability to understand prompts (LLM)
- Benefits
  - powerful content generation
  - automation of repetitive tasks [customer supp
  - enhanced ~~speech~~ search doc. classificatio
  - improved data analysis. : recommendation / analysis
- Applications
  - Sentiment analysis
  - chatbot
  - etc. eg. 5-6 eg.

- Levels of NLP.

1) Lexical Analysis.
2) Syntactic
3) Semantic
4) Discourse integration
5) Pragmatic.

(1b) - Phonological / (1c) Morphological

- independent existence — free morphemes — Eg. care, travel
- bound (dependent) — Eg. caring, traveller.

1a) Tokenization → Stop words — identify common insignificant words Eg: am, is, but, etc

- break word to manageable chunks.
- Lemmatization — scan / segregate text into sentence / para / word
- like short kama } manageable chunks } token. — to understand text at the word level.

takes more time (Caring → break karke)

Lemmatization / Stemming — better.

- Lemmatization — care banao.
- Stemming — does not check relevance.
  - └> Caring se break karke Car
  - closest root value.
  - ↓ irrelevant.

2 Syntactic : • parsing ; • Part of Speech tags (POS)

understand the structure of the word & relationship b/w the words.

• grammar
  └ structure | set of rules
  { using grammar / POS
  by labeling each part as such

POS: verb, noun, adverb etc
• identifiers.

3 Semantic : meaning to be conveyed

└ actual msg

Word Sense Disambiguation. — same word ke diff. meanings
↓ connected using sentences.
Eg bank — river bank
money

Name Entity Recognition
combine many entities for wider recognition.

4 Discourse Integration => Sequence.

- each sentence has some significance / relevance
- Contextual Reference → same type ki cheezein.
- Anaphora Resolution
  pronouns type ka
  pehle sentence mei NLP bola.
  baadmei, this subject bolda

NLP ≡ DS ✓
NLP ≡ History ✗

5 Pragmatic Analysis - change in tone.

voice modulation | └ sentence sarcastic / wonder / sad
└ Sentiment analysis

actually kya bolna chah rahe ho.

Feature Extraction

reduce dimension.

1) Term frequency (TF) — find significance of a word in doc.
   → no. of times a word appears in a doc / No. of total words in doc
2) Inverse Doc. frequency (IDF) — rarity of word in corpus
3) TF - IDF — statistical measure.
   └ TF × IDF
   • significance of a word AND rarity of word in doc

DOMS

Corpus — D_1 : Natural Lang
D_2 : Lang Processing
D_3 : Natural Lang. Processing.

combination

Vocabulary - #1 : Natural
#2 : Lang
#3 : Processing

ek ek karke

D_n → Documents
↳ corpus is a collection of documents

Term frequency ↓ of individual document	D_1	D_2	D_3
Natural	1/2	0	1/3
Lang	1/2	1/2	1/3
Processing	0	1/2	1/3

IDF : entire corpus : same documents

$$IDF = \log\left(\frac{\text{Total no. of doc in corpus.}}{\text{Total no. of appearance of a word across all doc}}\right)$$

Natural — $\log(3/2)$
Language — $\log(3/3)$
Processing — $\log(3/2)$

3 doc. { Total D_1 D_2 D_3

9/01/25

TF - IDF

of	D_1	D_2	D_3
Natural	$\log(\frac{3}{2}) \times 1/2$	$\log(3/2) \times 0$	$\log(3/2) \times 1/3$
Language	$\log(\frac{3}{3}) \times 1/2$	$\log(3/3) \times 1/2$	$\log(3/3) \times 1/3$
Processing	$\log(\frac{3}{2}) \times 0$	$\log(3/2) \times 1/2$	$\log(3/2) \times 1/3$

DOMS

N - Grams

- how many words together : contiguous sequence
  - N = 4, 5 : most effective.

Eg NLP

N=1 => {N , L , P}

N=2 => {NL, P}

N=3 => {NLP}

Eg Find the probability of <s> I I am </s> in the given corpus

1 : <s> I am a human </s>

2 : <s> I am not a robot </s>

3 : <s> I am not </s>

I I am not

Prob of (<s> and I together / just <s>)

=> P (I / <s>) x P (I / I) x P(am / I) x P (</s> / am)

P (I and I tog / just I)

I and am tog / just I.

Eg <s> I

count I

=> $\frac{3}{3} \times \frac{0}{3} \times \frac{3}{3} \times \frac{0}{3} = 0$

Q Predict : not ______ ?

=> P (/ not)

P (I / <s>) => <s> and I tog / just <s> = 3/3 = 1

P (am / I) => I and am / I = 3/3 = 1

P (a / am) => am and a / am = 1/3

P (human / a) => a and human / a = 1/2

P (not / am) -> am and not / am = 2/3

* P (a / not) => not and a / not = 1/2

P (robot / a) => a and robot / a = 1/2

P (</s> / human) => human </s> / human = 1/1 = 1

P (</s> / robot) -> robot </s> / robot = 1/1 = 1

* P (</s> / not) => not </s> / not = 1/2 = 1/2

P(a / not) = P (</s> / not)

↳ end of sentence not considered

agar occurrence along hai, jiski prob zyada vali consider.

So, ans: a

13 Jan 2025

Transformation based learning → Rule + Stochastic.

Syntactic level of NLP — noun / verb / qtq.

Parts of Speech Tagging.

① Dissemination of the World Class

- I am about to book a flight.
- I like to read book.

open
- Noun
- Adverb
- Adjective
- Verb.
  → required constant updation

closed.
- Preposition
- Pronoun
- Determinant
  → fixed.
  → appear frequently in a sentence
  → shorten words.

② Basis

Rule based

Stochastic based
- Word frequency measurement
- Tag sequence probability

Rule based:

- When a word is preceded by a determinant and followed by a noun: the word is an adjective
  → higher time complexity, inefficient, inaccurate

Stochastic based: better than rule based.

I like to play cricket.

Word frequency
- counting the no. of occurence of word
- 8 → verb 80%
- 2 → noun 20%

Ton verb wale ki mein probability

maano ki out of 10 sentences, 8 mein (play) is a verb, ⇒ 80%. 2 mein (play) is a noun ⇒ 20%

I like to participate in a play

play → V

10

PVP (V) — zyada reliable

Tag Sequence — Identifying a pattern.

look for patterns

W_1	W_2	W_3	W_4
P	V	P	V

Ki agar most sentences mein pronoun – verb – prepo. – verb hai, toh (PVPP) DOMS ke baad fir V hi aayega.

15 Jan 2025

Transformation Based Learning

decision making factors

Rule + Stochastic

i/p (Training Data)
- Tagged corpus
- List of dictionary words with their most frequent tags.

o/p
- sequence of transformation rules → change the tags
- accuracy

20 Jan 2025

Smoothing → used when scarce data or missing info
- (1) Log based ← Underflow condition ← not used a lot
- (2) Laplace
- (3) Add - K / Lidstone / Addictive. → redistribute vocabulary masses → negligible data

---

Training : Natural Language Processing

$P(\ ) = 1/3 * 1/3 * 1/3$

Testing : Natural Understanding.

$P(\ ) = 1/2 * 0/2 = 0$

Vocabulary = { N, L, P, U }

disadvantage

kyuki yeh

This is why we do Smoothing

still probability gives 0.

"Understanding" is in vocabulary

DOMS

Laplace Smoothing:

Explanation:
- add 1 to numerator
- add total no. of words to denominator

$\frac{1}{2} \times \frac{0}{2}$ becomes $\frac{1+1}{2+4} \times \frac{0+1}{2+4}$

How to actually solve:

Final Ans:
- $P(N) \times P(U)$

$$\frac{\text{total occurrences of N in training} + 1}{\text{no. of words in testing \& training} + \text{no. of words in vocabulary}}$$

{ final / count the word in training. if (training = true) check testing else 0

So,

Training : Natural Lang Processing

Testing : Natural Understanding

Vocab = {N, L, P, U}

$P(N) \times P(U)$

$$\Rightarrow \left(\frac{2+1}{5+4}\right) \times \left(\frac{0+1}{5+4}\right)$$

{ if word is in ~~testing~~ training, check in testing

$$= \frac{3}{9} \times \frac{1}{9} = \frac{3}{81} = \frac{1}{27}$$

if word not in training then seedha 0

22/01/25

Add K / Lidstone — more complex than laplace

NATURAL LANGUAGE PROCESSING

keeps on changing

count(N) — α or k

(1) + 0.5	1 + 0.5	1 + (0.5)
(3) + (0.5 × 3)	3 + (0.5 × 3)	3 + (0.5 × (3))

(3) → Total words

(3) → 3 vocabulary

Good turing Smoothing

No. of words in a bin

	Count	N	C = 2
Natural	2	2 } N_2	[N, L]
Language	2	2	
Processing	1	1 N_1	
Understanding	0	$N_0 = ?$	$P = \frac{2}{1}$ [P]
	5		

N_2 : [N, L] — no. of elem 2 isliye N, L ke 'N' column mei 2 aajayega

N_1 [P] — no. of elem = 1 isliye P ke 'N' column mei 1 aajayega

N_0 [U] — no. of elem = 0

$$C^* = (C+1)^* \frac{N_{c+1}}{N_c}$$

$$P^*( ) = \frac{C^*}{N}$$ — N → total no. of words

$$P(\text{unseen}) = \frac{N_1}{N}$$

DOMS

	N_2 [N, L]	N_1 [P]	N_0 [U]
Count	2	1	0
	5	5	

C^* $(2+1)^* \frac{N_3}{N_2}$ → ∵ no N_3 → can't calculate

$C^* : (1+1)^* \frac{N_2}{N_1}$: $2^* \frac{2}{1}$: 4

$C^* = (1)^* \left(\frac{N_1}{N_0}\right)$ = $1^* \frac{1}{1}$ = 1 — unseen probability

$P^* = \frac{C^*}{N}$ x can't calculate

$P^* : \frac{4}{25}$ → total no. of words

$P^* = \frac{1}{5}$ ← $N_1 = 1$ ← from P(unseen) vala formula.

agar 2 unseen words, then divide by 2

23/01/25 **Regular Expression.**

- ab^* → b is n no. of times
- $abc?$ → c can occur 0 or 1 times
- $abc\{2\}$ → c can occur max 2 times
- $abc\{2,\}$ → c's range starts with 2
- abc^+ → 1 or n no of c
- $\wedge a$ → string starts with a
- $b\$$ → string ends with b
- $a \wedge b$ → a and b
- $\wedge(A-Z)$ → Mean ^ acts as negation, A−Z range is not at all present (a−z) present.

Example questions

- Phone Number (starts with 8 or 9)

[8 9] [0 - 9]{9}

Phone number in format

XX - XX - XXXX

[8 9]{1} [0-9]{1} [\ -]

[0-9]{2} [\ -]

[0-9]{4}

escape character

(\) ke baad jo likhoge as it is print hojayega.

- Email

[a-z A-Z 0-9 _ \+ \...]

+ [\@] [a-z] + [\.]

[a-z]{2,3}

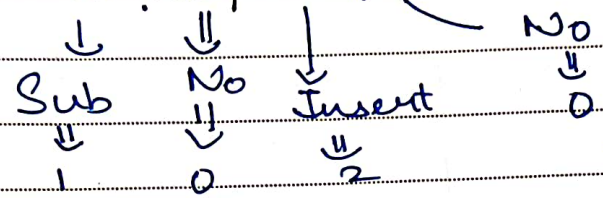
29/01/25

EDIT DISTANCE minimise the cost.

3 operations	Cost
→ Insert	→ 2
→ Delete	→ 2
→ Substitution	→ 1

S_1 : a | b | c

S_2 : d | b | e | c

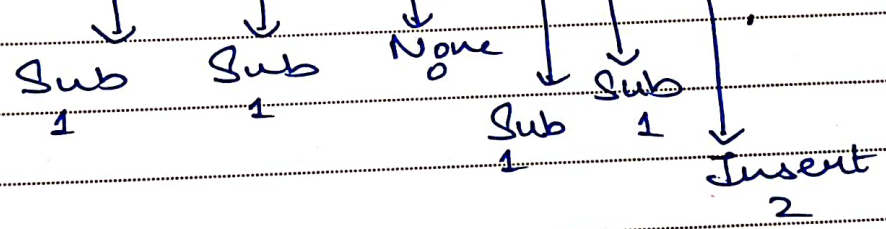


⇒ 3 cost

S_1 : I | N | T | E | N | T | I | O | N

S_2 : E | X | T | R | A | C | T | I | O | N

→ 0 cost



⇒ Cost : 6