Introduction to NLP (Natural language Processing)

Goal: For computers to process or "understand" natural languages.

Applications

1. Chatbots / Dialog systems
2. Assitive systems (alexa)
3. Auto-complete
4. Search engines
5. Spell Check
6. Sentiment Analysis
7. Machine Translator
8. Question Answering

Why is NLP difficult ?

1. **Complexity of Representation :**

   eg. bag of words

   I play cricket : [0 0 1 0 1 . . . . . 0 1 0]
   (1 → I, 1 → cricket, 1 → play)

2. **Human languages are inherently ambiguous**

   eg light ⟶ opposite of dark
   opposite of heavy

3. **Human language interpretation depends on real world, common sense knowledge and the context.**

   eg. Scientists study whales from space.

Basics of NLP

Syntax : Structure (grammar) in the language.

Semantics : Meaning in the language.

Word Tokenization : Divide a sentence into its component words.

type - An element of the vocabulary (unique)

token - An instance of that type in the text.

eg. they lay back on the San Francisco grass and look at the stars and their....

types : 12 or 13

tokens : 14 or 15.

Issues in Tokenization

(i) Finland's capital Helsinki.

Finland's
↓
Finland

What're
What are

I'm
I am

(ii) upper case vs lowercase.

(iii) Stemming / Lametization

- keep the root word.

worked, working → work

am, is, are → be

(iv) Stop-words

The, am, is, are, a,

Bag of Words

One hot representation

user interface management system.

user : [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]

interface : [1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

management : [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

System : [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0]

user interface management System.

[1 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0]

user interface interface managent system.

[2 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0]

Corpus:

$j=1$ w_{11} w_{21} w_{31}
- Human machine interface for computer applications

$TF = \frac{1}{7}$

$j=2$ w_{12} w_{22}
- User opinion of computer system response time

$j=3$
- User interface management system

$j=4$
- System engineering for improved response time

$TF = \frac{1}{6}$

V = [human,machine, interface, for, computer, applications, user, opinion, of, system, response, time, interface, management, engineering, improved]

machine:

0	1	0	...	0	0	0

TF-IDF (Term Frequency - Inverse Document Frequency)

weigh more : if a word is frequent in the current document.
also if word is infrequent in other documents.

(TF-IDF)

n_{ij} : no. of times word w_i appears in j^{th} document.

$$(TF)_{ij} = \frac{n_{ij}}{\sum_i n_{ij}}$$

hubble

$$TF(user) = \frac{1}{7}$$

total no. of words in document j.

Corpus:

- Human machine interface for computer applications
- User opinion of computer system response time
- User interface management system
- System engineering for improved response time

V = [human,machine, interface, for, computer, applications, user, opinion, of, system, response, time, interface, management, engineering, improved]

machine: | 0 | 1 | 0 | ... | 0 | 0 | 0 |

$$IDF(w) = \log_2 \frac{N}{N_w}$$

N : total no. of documents

N_w : total no. of documents that contain the word w.

$$IDF(user) = \log_2\left(\frac{4}{2}\right) = \log_2(2) = 1$$

$$(TF\text{-}IDF)_{ij} = TF_{ij} \times IDF(w_i) = 1 \times \frac{1}{7} = \frac{1}{7}$$

user

user