Day 2 , 10 Dec 2023

Natural Language Processing (NLP)

— by mentor Dr. Ryan Debnath, IIT Delhi + Harvard Alumni

Recap
1. What is NLP
2. NLP → application
3. Roadmap
4. Level -1

Tokenization

terminology

Corpus – paragraph
Document – sentences
vocabulary – unique words
words –

books, websites, different

tokenization → splitting text into smaller chunks / parts

paragraph → " sent1. sent2. ..... "

eg " I love NLP. It is very interesting topic. .... "
1 2

Sentence Tokenization
paragraph → sentences

sentence → I love NLP
1 2 3

word Tokenization
Sentence to words

Why Tokenization is needed ?
easy computation
easy processing

Stopwords

I am a boy and I love to eat chicken → boy love eat chicken

Q) Why we remove stopwords ?
easy ① computation, ② faster processing, ③ Redundant

text → no.

10 — — — — — → — — — — — — → Computation resource ↑
3 — — — → — — — → " " ↓

Stemming → Process of converting inflected word to stem/base word

eg ① history, historical → histori — Stem word / base word

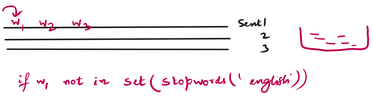
② loving, love, loveable → love

③ final, finalization, finally → fina

Lemmatization → convert the word to base word with some meaning to it.

eg history, historical → history
final, finalization, finally → final

Stemming	Lemmatization
① Convert to base word	① .... but with meaningful word
② Faster process	② Slower process wrt stemming
③ Application	③ Application
Sentiment analysis	chatbot / Q&A
Restaurant review	
— → +ve / -ve	




```
if w1 not in set(stopwords('english'))
    for w1 in word
rootword = stemmer.stem(W1)
```

Tokenization, Stopwords, Stemming , Lemmatization

Text processing : I Love NLP
① ↓
Love NLP
② ↓
numerical representation — 12 256

Bows , TFIDF

Bag of words (BOW)

Sent1 : He is intelligent boy
Sent2 : She is intelligent girl
sent3 : Both of them are intelligent boy and girl

	intelligent	boy	girl	
Sent1	✓	✓		
" 2	✓		✓	→ model
" 3	✓	✓	✓	
test	✓	✓		→