Xavor AI BootCamp

1-Processing

2-Generation

3-Understanding

Processing is super set of both things – (Generation and understanding)

Data majorly in Text - major source: google

# NLP

e.g: Machine Language Transformation

Text data forms : sports, political, news-data etc.

## History of NLP:

Prime example: Machine Language Transformation (solved)

1950- Chomsky – concept of finite state machine (FSM) language understand and generate

Chatbot- 1950-1980 : SHRDLU: Pick the green ball pick the black box

Text > Video > Audio

Sentiments Recognition: **Reviews**

Entity Extraction**: Subject, objective, Adjective**

Topic Classification: **Labeling the docs**

**Chat-GPT by Open AI**

* **ChatGPT**
* **GEMINI**
* **COPILOT**

**Stages of Language Processing:**

* Phonetics – Way of speaking/ accents/ phonen(basic unit of speech)

**Challenges:** bear/ bare (Homophones same sound)

* Morphology- morphine (basic unit of writing)

Go went gone, bank / banks

**Challenges:**

* Lexical Analysis- Lexicon – find the dictionary of words
* Atlas- **wordNET dictionary – words/ defination**

**Challenges:**

* Syntax Tree- Parsing:

**Challenges:**

* Scope

Higher Level Knowledge:

**Semantics:**

**Pragmatics:** statement ma given na ho- Lakin want krty ho- lift 50 manzil joty

**Discourse:**  Relation ship of words

## Basic Text Processing :

Regular Expression

Disjunction OR

Negation in Disjunction

? us ky pichly 0 and 1ho gi like

\*1 or more

+ atlleast 1

Text ka multiple docs (Corpuses- Corpra)

- Email verification

## Stop Words/ KEY Words

TOKENS

## Normalization- Pre Processing

* Duplicates remove
* Missing – Spelling galat
* Stop Words
* Punctuation etc

Word Count – **Term Frequency**

**Issue in Tokenization:**

Same Abbrivition and different words

Case Folding- Lower Case converstion

## LEMATIZATION

Base form ma convert krna

Go went gone--- GO

Dictionary needed

Words convertion

## Stemming

Words cutting

No Dictionary needed

Cut the words up to 3 letters

## Sentence Segmentation:

Sentence end on . words

## Libraries

* NLTK – research
* Spacy- work

# **Coding part:**

Data Acquisition

Pre-processing

NLP- stop words,

## PANDAS

Head

Tail

Sample – randomly

# Case Folding

.str.lower()

# Remove Tags

Regular Expression-re

* Re.compile(‘<,\*?>’)
* Pattern.sub(r’’,text)

.apply()

# Remove URL

Same

# Remove Punctuation

.translate(.str.maketrans(‘’,’’,exclude))

# Replace Short Hands

Split in dictionary

# Correct Incorrect Words

* TextBlob – Library

# Stop Words Removal

NLTK

NLTK.CORPUS

Stopwords

# Emoji Removal

* Demojize
* Emoji

# Tokenization

Breaking of sentence into words.

* Nltk - research
* Spacy- accurate- fast

# Stemming

* Porter Stemmer
* Snowball

# Lemmatizing

* WORD-NET
* Lema

# Spacy vs NLTK