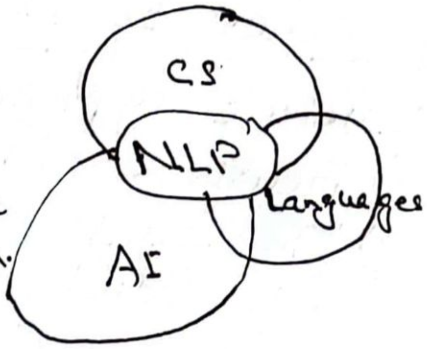
Natural Language Processing (NLP)

Lecture 12

Introduction

- → NLP is subfield of linguistic & CS, & AI concerned with the interactions betw computers & human language. In particular how to program computers to process & analyze large amounts of natural language data.
- → New field, emerged in last 5 years.
- → With DL (especially tranformers), last 10 years have seen explosive growth.



Need

→ In neuropyshology, linguistics and language philosophy language evolves naturally without concious planning or premediation. It can take different forms such as speech or signing and are different than formal languages (C++, logical programming).

Real World Applications

- → Chatbots
- → Contextual Advertisements
- → Email Clients (Spam filtering, smart reply)
- → Social Media (removing adult content, opinion mining).
- → Search engines.

Common NLP Tasks

→ Need to master these.

1) Text/Doc Classification
2) Sentiment Analysis.
3) Information Retrieval.
4) Parts of Speech tagging.
(5) Language detection & Machine translation.
(6) Conversational Agents (Speech based).

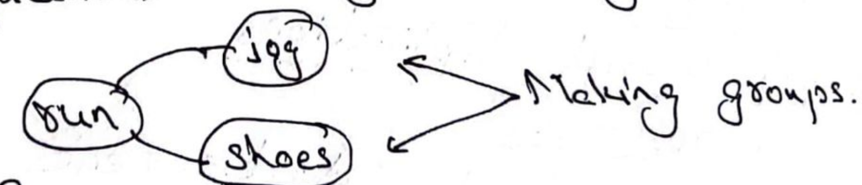
7). Knowledge graph & QA Systems.
8). Text Summarization
9). Topic Modelling
10). Text generation
11). Spell checking & Grammer Correction.
12). Text parsing.
13). Speech to text.

Approaches to NLP

→ Different approaches to implement NLP.
1) Heuristic Methods.
2) ML Based Methods.
3) DL Based Methods.

1) Heuristic Method (HM)
is any approach to problem solving or self-discovery that employs a practicle method that is not guerented to be optimal, perfect or rational but is sufficient to reach an approximation.

Example:
→ Regular Expressions (C++, Python finding salutations).
→ Wordnet



→ Open Mind Common Sense
↳ Open source effort to list all universal facts etc.

→ It is quick and is used as help to ML & DL.

ML Methods:
→ Advantage on ML is that, ML is rule based.
→ Can work on open-ended problems.
→ Text is converted to numbers. (Naive Bayes, LR, SVM, LDA, Hidden Markov Models)

DL Approaches:

→ In ML, text conversion to numbers loses the sequential order. (e.g. This is my house). In DL, this shortcoming is overcome.

→ In DL, models generate features.

→ RNN (not optimal for long sentences), LSTM, GRU, CNN, Autoencoders, Transformers.

Challenges

→ NLP work on human languages so inherently challenging.

1). Ambiguity (I saw the boy on the beach with my binoculars).
2). Contextual Words. (I ran to the store becz we ran out of milk).
3). Slang (Piece of Cake).
4). Synonyms
5). Irony, Sarcasm & tonal difference.
6). Spelling errors.
7). Creativity.
8). Diversity.

NLP Pipeline

→ NLP is a set of steps followed to build an end to end NLP software.

NLP consists of following steps

- Data Acquisition
- Text preparation
  - → Text cleanup
  - → Basic preprocessing
  - → Advance preprocessing
- Feature Engineering
- Modelling
  - Model Building
  - Evaluation

- Deployment
  - Deployment
  - Monitoring
  - Model Update

→ This pipeline is not universal (change according to application) and non-linear.

→ DL pipelines are slightly different.

Data Acquisition

→ Example of customer sentiment Analysis

Dat Acquisition

- Available
  - .csv
  - Less Data
    - ↳ Data Augmentation
      - ↳ Synonyms
      - ↳ Bigram flip (I am walking on the road)
      - ↳ Back translation
    - using tools
- Others
  - public dataset
  - web scraping
  - API
  - Pdf
  - Images & Audio
- Nobody
  - ↳ Survey.

Text Preparation

- Cleaning
  - ↳ HTML tag removal <p> Hi <p>
  - ↳ Emoji
  - ↳ Spelling Check
- Basic Preprocessing
  - Basic
    - ↓ Tokenization
      - Word
      - Sentence
  - Optimal
    - ↳ Stop word removal
    - ↳ Steming (dance, dancing, danced) bring to dance
    - ↳ Removing digits, punctuations
    - ↳ lower case
- Advance Preprocessing
  - ↳ POS (Part of speech tagging)
  - ↳ Parsing (syntax understanding)
  - ↳ Coreference resolution

→ My name is $\underbrace{\text{Hamza.}}$ $\underbrace{\text{I am}}$ a student at UET P

→ POS tagging

Mr. $\underbrace{\text{Charles}}_{\text{Noun}}$ $\underbrace{\text{wrote,}}_{\text{verb}}$ directed and composed the music.

Feature Engineering

→ Changing text to numbers.

→ Sentiment Analysis. (review of movies)

→ Example

50,000 rows & 2 columns

Text	Sentiment
Review 1	0
Review 2	1

Convert to Numbers

# of +ve words (e.g. happy, good, great)	# of -ve words (e.g. pathetic, bad)	# of neutral	Sentiment
3	1	6	

→ Very basic technique (text vectorization)

→ Advanced
- → Bag of words
- → Tdiaf
- → OneHot Encoding
- → WordtoVec

→ F.E. technique depends on problem (e.g. sentiment analysis, summarization).

Modelling:

- Modelling
  - Modelling — Depends on
    - → Amount of data
    - → Nature of problem
    - Heuristics
      - → Spam classification (based on email address, if sale word is used)
    - ML Algo
    - You can use both, if a lot of emails.
  - DL Algo (Needs a lot of data to train)
    - → Transfer Learing is where you bring in already trained model (BERT which is already trained on 40GB data) for your problem.
  - Cloud API
- Evaluation
  - Intrinsic (Purely technical, based on model's accuracy)
  - Extrinsic (Business Model)
    - Writing email), it gives you suggestions.
    - → If customer is choosing 1 in 3, vs 0 in 100 suggestions.

# of +ve words	# of -ve words	Spamer
3	6	0
4	6	1

Deployment

- Deploy
- Monitoring
- Update