Unit 1
Introduction to Natural Language Processing

Natural Language Processing (NLP)
MDS 555

Objective

- Introduction
- Terminologies
- Challenges of NLP
- History
- Some Application overview

What is Language?

- **Britannica**: Language is a system of conventional spoken, manual (signed), or written symbols by means of which human beings, as members of a social group and participants in its culture, express themselves.

- The functions of language include

  - communication, the expression of identity, play, imaginative expression, and emotional release.

- Based on this definition we can divide the language data into

  - Written (Text)

  - Spoken (Speech)

What do we use language for?

- We communicate using language

- We think (partly) with language

- We tell stories in language

- We build Scientific Theories with language

- We make friends/build relationships

What is NLP?

- It is a sub-field of artificial intelligence that is concerned with the human computer interaction using natural language.

- It is interdisciplinary field

  – Computer and Linguistics

- We can 1) Analyze and 2) Produce the natural language using NLP techniques

What is NLP?

- If we take any form of the language (written or spoken), the format is always unstructured

- NLU: Natural Language Understanding

- NLG: Natural Language Generation
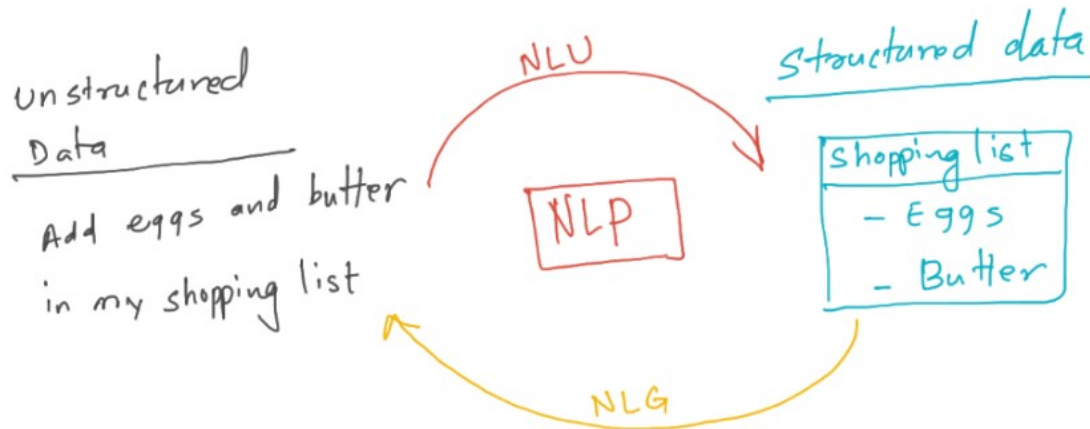
What is NLP?

- The NLP can be divided into two category
  - Text Processing
    - Machine translation
    - Spam email detection
    - Document classification
    - Text summary generation
    - Sentiment Analysis
  - Speech Processing
    - Text-to-speech Generation (Speech Synthesis)
    - Automatic Speech Recognition

Why NLP?

- Access Knowledge
  - search engine, recommend system

- Communicate
  - Translation, synthesis, recognition

- Linguistics and Cognitive Sciences
  - Analyse Languages themselves

Why NLP?

- Amount of Online Data
  - 70 billion web pages online
  - 55 million wiki articles
  - 156 million hours of video on youtube alone
  - 9000 tweets / second
  - 3 million email / second – 60% are spam
- These data made NLP important

Users of NLP

- 7.9 billion people use some sort of language (January 2022)

- 4.7 billion internet users (January 2021) (~59%)

- 4.2 billion social media users (January 2021) (~54%)

What Product in NLP?

- Search: +2 billion Google users, 700 millions Baidu users

- Social Media: +3 billion users of Social media (Facebook, Instagram, WeChat, Twitter...)

- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)

- Machine Translation: 500M users for google translate

Five Level of Linguistics

- Phonology   study of how sounds function within languages.

- Morphology   study of the structure of words and the formation of word forms.

- Syntax   study of the rules and principles for constructing sentences in a language.

- Semantics   study of meaning in language.

- Pragmatics (context)   study of how context influences the interpretation of meaning

- Extra-linguistic – other material along with language

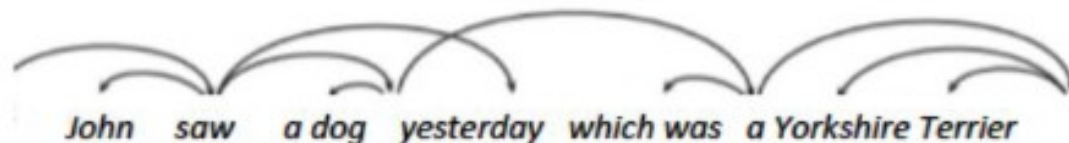|		
Analysis in context	Extra-linguistic context	*Found **him** in the street inside a bag. I think **he** is happy with his new life* http://9gag.com/gag/aVrhSwp/found-him-in-the-street-inside-a-bag-i-think-he-is-happy-with-his-new-life
	Linguistic context	— *You know what? John gave Peter a Christmas present yesterday* — *Wow, was **he** surprised? What was **it** like?* — *Surprisingly good. **He** spent quite a bit on **it**.*
	Semantic level	The landlord$_{SPEAKER}$ has not yet **REPLIED**$^{Communication\_response}$ in writing$_{MEDIUM}$ to the tenant$_{ADRESSEE}$ objecting the proposed alterations$_{MESSAGE}$.$DNI_{TRIGGER}$
Sentence-level analysis	Syntactic level	*John saw a dog yesterday which was a Yorkshire Terrier*
	Morphological level	*brav+itude, bio+terror-isme/-iste, skype+(e)r* *mang-er-i-ons* = MANGER+cond+1pl
	Phonological level	International Phonetic Alphabet [aɪ pʰiː eɪ]
	Graphemic level	*enough, cough, draught, although, brought, through, thorough, hiccough*

Phonology

- study of the sounds of a language
  - Every language has its own inventory of sounds and logical rules for combining those sounds to create words.
  - The phonology of a language essentially refers to its sound system and the processes used to combine sounds in spoken language.

Morphology

- Study of the internal structure of the words of a language
    - There are many words to which a speaker can add a suffix, prefix, or infix to create a new word
    - The morphology of a language refers to the word-building rules speakers use to create new words or alter the meaning of existing words in their language

Syntax

- Study of sentence structure
    - Every language has its own rules for combining words to create sentences
    - Syntactic analysis attempts to define and describe the rules that speakers use to put words together to create meaningful phrases and sentences

Semantics

- Study of meaning in language
  - Linguists attempt to identify not only how speakers of a language distinguish the meanings of words in their language, but also how the logical rules speakers apply to determine the meaning of phrases, sentences, and entire paragraphs
  - The meaning of a given word can depend on the context in which it is used, and the definition of a word may vary slightly from speaker to speaker

Pragmatics

- Study of the social use of language
  - All speakers of a language use different registers, or different conversational styles, depending on the company in which they find themselves.
  - A linguistic analysis that focuses on pragmatics may describe the social aspects of the language sample being analyzed,
  - Such as how the status of the individuals involved in the speech act could affect the meaning of a given utterance.

Challenges of NLP

1) Productivity

2) Ambiguous   Semantic ambiguity
                Pragmatic ambiguity

3) Variability

4) Diversity   Syntactic diversity

5) Sparsity

Productivity

Definition: "property of the language-system which enables native speakers to construct and understand an indefinitely large number of utterances, including utterances that they have never previously encountered." (Lyons, 1977)

- New words, senses, structure are introduced in languages all the time

- Examples: social distance were added to the Oxford Dictionary in 2021

- Why **Sanskrit** is not in common use?

Ambiguous

- Most linguistic observations (speech, text) are open to several interpretations

- We (Humans) disambiguate
  - i.e. find the correct interpretation
  - using all kind of signals (linguistic and extra linguistic)

- Ambiguity can appear at all levels
  - phonology, graphemics, morphology, syntax, semantics

Semantic Ambiguity

- Polysemy: e.g. set , arm, head
  - Head of New-Zealand is a woman

- Name Entity: e.g. Michael Jordan
  - Michael Jordan is a professor at Berkeley
    basketball player and businessman

- Object/Color: e.g. cherry
  - Your cherry coat

Pragmatic Ambiguity

- Two Soviet ships collide, **one dies**
- Dealers will hear **car talk** at noon

Ambiguous

- Disambiguating can requires Discourse Knowledge
  - Where can I find **a vegetarian restaurant** in **Paris**

    Here is a list of restaurant in Paris: ….

  - Give me the **top ranked** **ones**, **in the 14th arrondissement**

    Here are the top ranked restaurant in the 14th arrondissement in Paris

  - How far is **the closest** **one** from my current location?
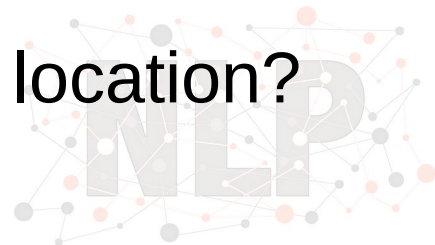
Variation

- Language Varies at all levels
  - Phonetic (accent)
  - Morphological, Lexical (spelling)
  - Syntactic
  - Semantic

Variation Determiners

- Who is talking?
  - To Whom?
  - Where? Work, Home, Restaurant
  - When? 19th century, 2008, 2022…
  - About what? Specialised domain, the Weather,…

- Essentially, the Variability of a language depends on
  - Social Context
  - Geography
  - Sociology
  - Date
  - Topic

Diversity

- About **7000 languages** spoken in the world
- About 60% are found in the written form

- Phonologic Diversity
- Graphemic Diversity (latin, arabic, devanagari, greek)

Syntactic Diversity

- A key characteristics of the syntax of a given language is the word order

  - Word order differs across languages

  - Word order degree of freedom also differs across languages

  - We characterize word orders with: Subject (S) Verb (V) Object (O) order

What is Natural Language Processing?

What is NLP?

- In a nutshell, NLP consists in handling the complexities of natural languages "to do something"
  - Raw Text / Speech → Structured Information
  - Raw Text / Speech → Speech / Text

  In this course we will focus on **textual data**

Terminologies

Tokenization

- splits longer strings of text into smaller pieces, or tokens

- Larger chunks of text can be tokenized into sentences

- sentences can be tokenized into words, etc.

- Further processing is generally performed after a piece of text has been appropriately tokenized

Normalization

- Normalization generally refers to a series of related tasks

  - converting all text to the same case (upper or lower),

  - removing punctuation,

  - expanding contractions,

  - converting numbers to their word equivalents, and so on.

- Normalization puts all words on equal footing, and allows processing to proceed uniformly.
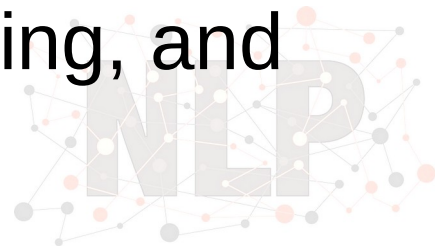
Stemming

- Stemming is the process of eliminating affixes
  - suffixed, prefixes, infixes, circumfixes
- Stemmer: Tool to obtain a word stem.

  - Running – run
  - Unmanage - manage

Lemmatization

- Lemmatization is related to stemming, differing in that lemmatization is able to capture canonical forms based on a word's lemma.

  - For example, stemming the word "better" would fail to return its citation form (another word for lemma); however, lemmatization would result in the following:

    better → good

- Implementation of a stemmer would be the less difficult

Stemming vs Lemmatization
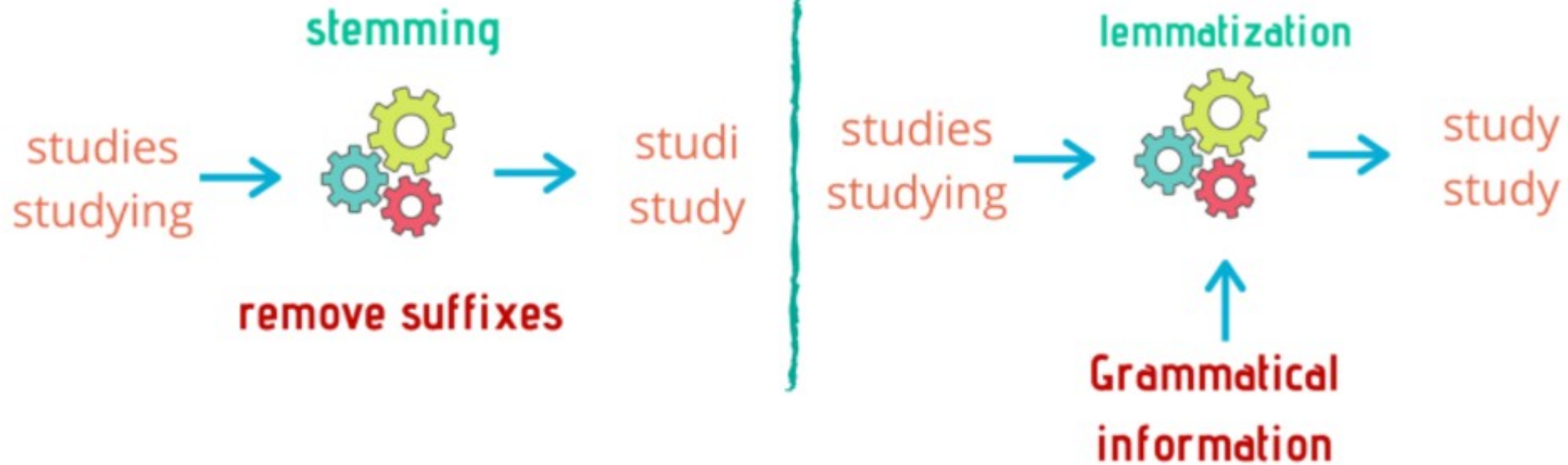
Corpus / Corpora

- In Latin corpus means **body** refers to a collection of texts

- Sources
  - Single
  - Multiple
  - Multilingual

- Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

Stop Words

- Stop words are those words which are filtered out before further processing of text, since these words contribute little to overall meaning, given that they are generally the most common words in a language.

    – A, an, the, in , on etc.

    – **The** quick brown fox jumps over **the** lazy dog.

Parts-of-speech (POS) Tagging

- POS tagging consists of assigning a category tag to the tokenized parts of a sentence.

- The most popular POS tagging would be identifying words as nouns, verbs, adjectives, etc.
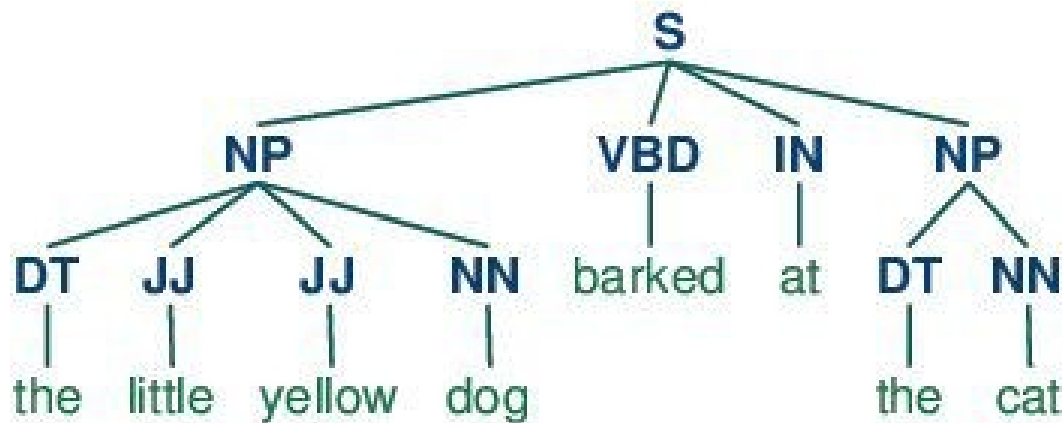
POS Tagging

- POS Tag:
  - Verb, Particle, Noun, Adverb, Adjective, Pronoun, numeraletc.

- Use of PoS tagging
  - Chunking
  - Syntax Parsing
  - Information extraction
  - Machine Translation
  - Sentiment Analysis
  - Grammar analysis & word-sense disambiguation

Bag of Words

- Bag of words is a particular representation model used to simplify the contents of a selection of text.

- The bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.

"Well, well, well," said John. "There, there," said James. "There, there."

{ 'well': 3, 'said': 2, 'john': 1, 'there': 4, 'james': 1 }
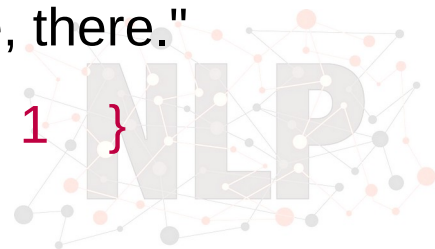
n-grams

- n-grams is another representation model for simplifying text selection contents

- As opposed to the orderless representation of bag of words, n-grams modeling is interested in preserving contiguous sequences of N items from the text selection

Terminologies

- Regular Expressions

- Similarity Measures

- Zipf's Law: describe the relationship between word frequencies in document collections

- Syntactic Analysis

- Semantic Analysis

- Sentiment Analysis

- Information Retrieval

Terminologies

- Named Entity Recognition

- Categorization

- Text to Speech (TTS)

- Automatic Speech Recognition (ASR)

- Text Mining

- Polarity

3-gram

("There, there," said James. "There, there.")

- Appears as a list representation below:

  3-gram model

  [

  "there there said",

  "there said james",

  "said james there",

  "james there there",

  ]

History of NLP

Symbolic - 1940-2000

- Focus on rule-based systems

- formal grammars

- Development of linguistic resources
  - Lexicon, ontologies, grammars

Statistical Learning - 1990-2010

- Statistical learning theory

- SVM, Random Forest

- Graphical Probabilistic Models

  - Latent Dirichlet Allocation (LDA),

  - Hidden Markup Model (HMM)

- Development of annotated datasets

Deep Learning – 2010 - today

- Deep Learning Architecture (Transformer)
- Transfer Learning in NLP
  - word2vec, BERT,CamemBERT, GPT, Wav2Vec
- More compute
  - larger (raw) dataset
  - Open Source Deep Learning Libraries

NLP Applications

Standard NLP pipeline
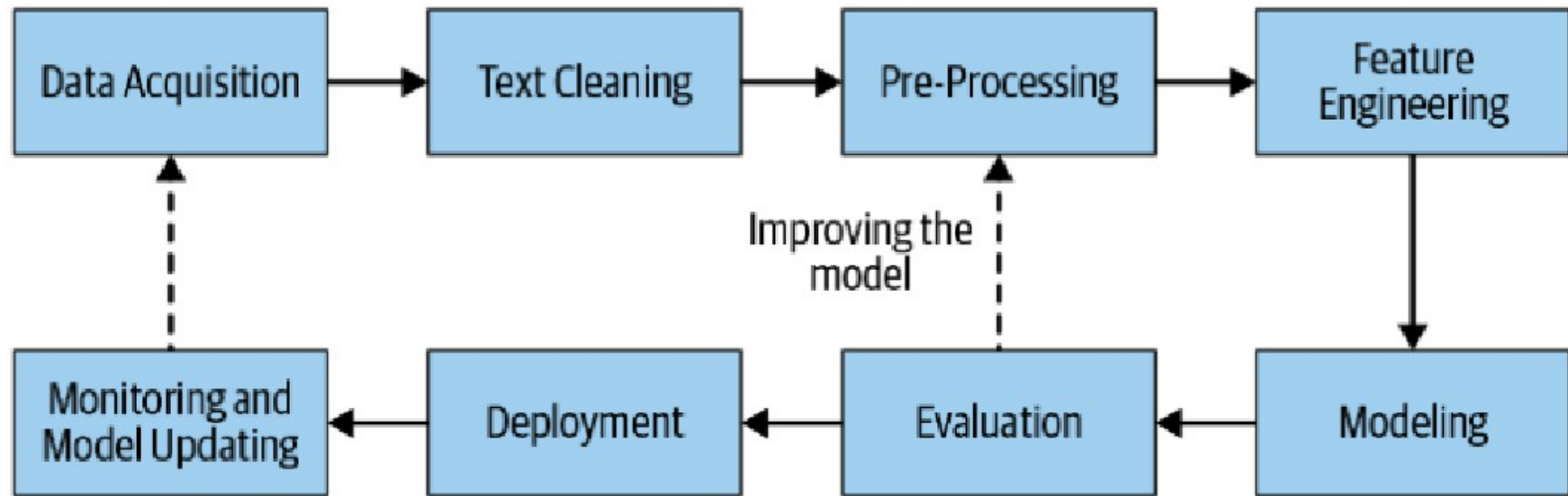
Document Classification

Germany's minimum wage hike will not cost jobs -labour minister

BERLIN, Jan 21 (Reuters) - Germany's planned minimum wage hike to 12 euros ($13.61) per hour from October means a pay rise for over 6 million people across the country and should not cost jobs contrary to critics, Labour Minister Hubertus Heil said on Friday.

Increasing the German minimum wage, currently 9.82 euros per hour and will increase to 10.45 euros per hour from July, to 12 euros per hour was one of the key election promises of Chancellor Olaf Scholz and his Social Democrats.
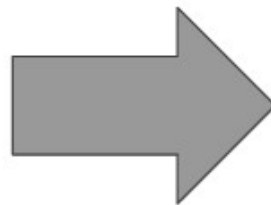
Politics

Economy

Travel

….

Geopolitics

Sentiment Analysis

- Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral.

  - **Provide objective insights:** accurate sentiment analysis tools sort and classify text to pick up emotions objectively

  - **Build better products and services:** A sentiment analysis system helps companies improve their products and services based on genuine and specific customer feedback.

  - **Analyze at scale:** Can be done on unstructured data that are mined for business analytics purpose

  - **Real-time results**

Document Ranking (Retriever)

Slot-Filling / Intent Detection

- Intent Detection is a sequence classification task that consists in classifying the intent of a user in a pre-defined category.

- Slot-Filling is a sequence labeling task that consists in identifying specific parameters in a user request.

Can you please play Hello from Adele ?

Intent: play_music

Slots: [Can, you, please, play, **Hello**, from, **Adele**, ?]

[O , O , O , O , SONG, O , ARTIST, O]

Name Entity Recognition

- NER: Find the Name-Entities in a sentence

[My , name, is, **Bob**, and, I, live, in, **NY**, !]

[O , O, O, **PERSON**, O, O, O, O, **LOCATION**, O]

Other Applications

- Document Summarization

- Question Answering

- Machine Translation

- Automatic Answering

Finally,
Performing NLP Research

Natural Language Processing WorkFlow

- Assume we have a Research, Engineering, Product Problem
    1) Define a NLP System to solve it Split into modules, each one performing a task
    2) Define Evaluation Metric(s) for your system and sub-modules
    3) Collect Data to build/train your models
    4) Build Baseline Models (i.e. most simple model you can think of that have a non trivial performance metric)
    5) Build Better Models using symbolic/statistical/DL methods
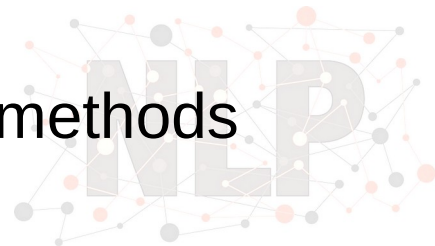
Thank you

Join Google Classroom

pbbksxd