# UNIT V
## NLP Tools and Techniques:-

* **Prominent NLP Libraries:-**

- **Natural Language Tool kit (NLTK):-**
  - NLTK is a very powerful tool.
  - It is most popular in education & research
  - It has led to many breakthrough in text analysis.
  - It has a lot of pre-trained models and corpora which helps us to analyze things very easily.
  - It is an excellent library when you require a specific combination of algorithms.
  - NLTK features:
    tokenization, pos, NER, classification, sentiment analysis, access to corpora, package for chatbots.
  - NLTK use cases:-
    - Remove stop words & persons names in a Recommendation system with NLTK.
    - Sentiment Analysis with NLTK - check if product review is positive or negative.
    - Build a chatbot using python NLTK - understand who is the target audience, the intent or desire of the user and provide responses that can answer the user.

Pros:-

- Most well-know and full NLP library with many 3rd Extensions.
- Supports the largest no. of languages compared to other libraries.

Cons :-

- Difficult to learn and use, slow
- only splits text by sentences, without analyzing the semantic structure.
- No neural network models.

Spacy:-

- It is an advanced NLP library available in Python and Cython.
- It is geared toward performance and operating together with deep learning frameworks such as Tensorflow or pyTorch
- It comes with pre-trained statistical models and word vectors.
- It features tokenization for 50+ languages. Convolutional neural network models for tagging, parsing & named entity recognition.

Spacy features:-

Tokenization, POS, NER, classification, sentiment analysis, dependency parsing, word vectors.

Spacy usecases:-

Search autocomplete (and autocorrect) is popular type of NLP that many people use on a daily basis.

- Analysis online review. Extract the key topics covered by the reviews without having to go through all of them.
  Help the sellers/ retailers get customer feedbacks in the form of topics.

- Automatic summarization of Resumes with NER Evaluates resumes at a glance to facilitate - Evaluation of resume at a quick glance, thereby simplifying the effort required in shortlisting candidate — pile of resumes.

- pros
  - Fast
  - Easy to learn & use
  - Uses neural networks for training models.
- Cons
  - Less flexibility Compared to NLTK

Text Blob:—

- It is based on NLTK & pattern
- It has great API for all the common NLP operations.
- It's a more practical library Concentrated on day-to-day usage.
- It's great for initial prototyping in almost every NLP project.
  Unfortunately, it inherits the low performance

from NLTK & therefore it's not good for large scale production usage.

**Features:**

Tokenization, POS, NER, classification, sentiment analysis, spellcheck, parsing, translation & language detection.

**Usecases:**

- Sentiment analysis
- Spelling correction.
- Translation & language detection

**Pros:**

- Easy to use & intuitive interface to NLTK library.
- provides language translation & detection which is powered by google translate.

**Cons:**

Slow, no neural network models, no integrated word vectors.

**Gensim:**

- It is one of the top python libraries for NLP.
- It was originally developed for topic modelling, but today it supports a variety of other NLP tasks, but it is not a complete NLP toolkit like NLTK or Spacy.

- It's primary usecase is working with word vectors.
- Word vectors improve our ability to analyse relationships across words, sentences & documents.

**Features:-**

parallelized implementations of fastText, word2vec & doc2vec algorithms, latent semantic analysis (LSA, LSI, SVD), non-negative matrix factorization (NMF), Latent Dirichlet allocation (LDA), tf-idf.

**Usecases:-**

Converting words & documents to vectors
finding text similarity
text summarization

**Pros:-**

- Intuitive interface
- Efficient implementation of popular algorithms.
- Scalable - can run latent semantic analysis & latent Dirichlet allocation on a cluster of computers.

**Cons:-**

- Designed primarily for unsupervised text modelling.
- Won't implement full NLP pipeline, should be used with other library like Spacy or NLTK.

* **Linguistic Resources:-**

Linguistic resources are essential for creating grammers in the frameworks of symbolic the approaches or to carry out the training of modules based machine learning.

**Corpus:**

A corpus is a large & structured set of machine readable texts that have been produced in a ~~neutral~~ to natural communicative setting. It's plural is corpora.

They can be derived in different ways like text that was originally electronic, transcripts of spoken language & optical character recognition etc--

* **Lexical Knowledge Networks:-**

Lexico-Semantic networks such as the princeton WordNet are being considered vital resources.

- Wordnets are being constructed in different languages
- Competing lexical networks such as ConceptNet, HowNet, mindNet, VerbNet & FrameNet are also Emerging as alternatives to wordnets.
-
- users are interested in knowing not only the relative merits from among a selection of choices but also the intrinsic value of such resources.

- How do you disambiguate 'web' in the spider spun a web' from 'go surf the web!'

* How do you summarize a long paragraph?
* How do you automatically construct language phrasebooks for tourists?

* many of these issues can be resolved just by knowing more about the meaning of words lexical semantics theory.

- Need a lexicon that provides
  - dictionary or thesaurus-like information
  - more rich associations among words.
- No major evaluations proposed or tried or lexical nw
- LKNS Embedded a conceptualization of the world
- share universal properties across different languages.
- The properties are combinatorial & graph theoritic in nature & pertain to the path length, degree, density etc--
- They indicate the level of maturity of LKN.

* **WordNet:-**

Wordnet is created by princeton is a lexical database for English language.

It is a part of the NLTK Corpus.

- In WordNet, nouns, verbs, adjectives & adverbs are grouped into sets of cognitive synonyms called Synsets.

* All the synsets are linked with the help of conceptual semantic & lexical relations.

* In information systems, wordNet is used for various purposes like word-sense disambiguation, information retrieval, automatic text classification & machine translation.

- WordNet is to find out the similarity among words.

* Indian language wordNet (IndoWordNet) :—

* It is a linked lexical knowledge base of wordNet of 18 Scheduled languages of India viz, Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Marathi, Nepali, odia, Punjabi, Sanskrit, Tamil, Telugu, Urdu.

- These wordNets have been created using the expansion approach from Hindi wordNet & English.

- Each entry in the IndoWordNet consists of the following elements along with a related lexicon in other Indian languages, including English:

(1) Synonymy
(2) Gloss
(3) Example sentence

- IndoWordNet is highly similar to EuroWordNet.
- The pivot language is Hindi, which of course is linked to the English wordNet.
- IndoWordNet is publicly browsable
- The Indian language wordnet building efforts forming the subcomponents of IndoWordNet. project are:

North East wordNet project, Dravidian word Net project & Indradhanush project of which are funded by TDIL project.

* **VerbNet (VN):** —

It is the hierarchical domain-independent & largest lexical resource present in English that incorporates both semantic as well as syntactic information about its contents.

- VN is broad coverage verb lexicon having mapping to ther lexical resources such as wordNet, xtag, & frameNet.
- It is organized into verb classes extending levin classes by refinement & addition of subclasses for achieving syntactic & semantic coherence among class members.
- Each VerbNet class contains
  (1) A set of syntactic descriptions or syntactic frames.

(5) A set of semantic descriptions such as animate, human, organization.

* **PropBank:-**

PropBank more specifically called "Proposition Bank" is a corpus which is annotated with verbal propositions & their arguments.

- The corpus is a verb-oriented resource; the annotations here are more closely related to the syntactic level.
- Martha palmer et al, Department of linguistic, university of Colorado Boulder developed it.
- We can use the term propBank as a common noun referring to any corpus, that has been annoated with propositions & their arguments.
- In NLP, the propBank project has played a very significant role.
It helps in semantic role labelling.

* **Treebanks:-**

It may be defined as linguistically parsed text corpus that annotates syntactic or semantic sentence structure.

- Geoffrey Leech coined the term 'treebank' which represents that the most common way of representating the grammatical analysis is by means of a tree structure.

- Treebanks are created on the top of corpus, which has already been annotated with part-of-speech tags.

**Types of treebanks corpus:-**

(1) **Semantic treebanks:-**

* Use a formal representation of sentences semantic structure
- They vary in the depth of their semantic representation.
- Robot commands treebanks, Geoquery, Groningen meaning Bank, Robocup Corpus are Eg. of semantic treebanks

(2) **Syntactic Treebanks:-**

Inputs to the Syntactic treebanks systems are expressions of the formal language obtained from the conversion of parsed Treebanks data.

- the outputs of such systems are predicate logic based meaning representations.

Eg: Penn Arabic Tree bank, Columbia Arabic treebanks created in Arabic language.

* **Universal Dependency Treebanks:-**

It is frequently abbreviated as UD, is an International cooperative project to create treebanks of the world's languages.

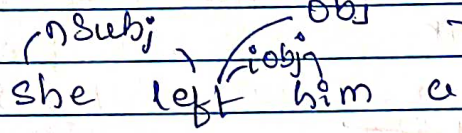
- These treebanks are openly accessible & available.
- UD is a project that is developing cross-linguistically consistent treebank annotation for many languages. With the goal of facilitating multilingual parser development,

Cross-lingual learning & parsing research from a language typology perspective.

- **Dependency structures:**

- The UD annotation scheme produces syntactic analyses of sentences in terms of the dependencies of dependency grammar.
- Each dependency is characterized in terms of a syntactic function, which is shown using a label on the dependency edge.

Eg:




This analysis shows that she, him and a note are dependents of the left.

- The pronoun she is identified as a nominal subject (nsubj), the pronoun him as an indirect object (iobj) & noun phrase a note as a direct object (obj)
There is a further dependency that connects a to note.

* **Word sense disambiguation:**

- In NLP, WSD is the problem of determining which sense of a word is activated by the sense in a particular context, a process which appears to be largely unconscious in people.

WSD approaches are categorized mainly into 4 types
1. Knowledge-based & Dictionary based
2. Supervised
3. Unsupervised
4. Semi-Supervised

* **Dictionary & knowledge-based methods:—**
These methods rely on text data like dictionaries, thesaurus. It is based on the fact that words that are related to each other can be found in the definitions.

* **Supervised methods:—**
In this type, sense-annotated corpora are used to train machine learning models. But a problem that may arise is that such corpora are very high & time-consuming to create.

* **Unsupervised Methods:—**
Unsupervised methods pose the greatest challenge to researchers & NLP professionals.
- A key assumption of these models is that similar meanings & senses occur in a similar context.
- They are not dependent on manual efforts, hence can overcome the knowledge acquisition deadlocks.

* **Lesk Algorithm:-**

Lesk Algorithm is a classical WSD algorithm introduced by michael E. Lesk in 1986.

- It is based on the idea that words in a given region of the text will have a similar meaning.

- In the simplified Lesk algorithm the correct meaning of each word context is found by getting the sense which overlaps the most among the given context & its dictionary meaning.

- Basic Lesk algorithm implementation involves the following steps:

  - Count the number of words in the neighborhood of the word & in the dictionary definition of that - Sense for each sense of the word being disambiguated

  - The sense to be picked is the one with the greatest number of items in this count.

* **Walker's Algorithm:-**

- A thesaurus based approach.

**Step1:**

For each sense of the target word find the thesaurus category to which that sense belongs.

**Step2:-**

Calculate the score for each sense by using context-words. A context words will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense.

**Eg:** The money in this Bank fetches an interest of 8% per annum.

Target word : Bank

Clue words from the context: money, interest, annum.

| | Sense1: Finance | Sense2: Location |
|---|---|---|
| money | +1 | 0 |
| Interest | +1 | 0 |
| Fetch | 0 | 0 |
| Annum | +1 | 0 |
| Total | 3 | 0 |

* **WordNets for words Sense Disambiguation:-**

- WordNet is the lexical database ie dictionary for the English language, specifically designed for NLP.

- Synset is a special kind of a simple interface that is present in NLTK to look up words in WordNet.

- Synset instances are the groupings of synonymous words that express the same concept. Some of the words have only one Synset & some have several.

- WordNet lists five senses for the word pen:
  - pen - a writing implement with a point from which ink flows.
  - playpen, pen - a portable enclosure in which babies may be left to play.
  - penitentiary, pen - a correctional institution for those convicted of major crimes.
  - pen - female swan.

- WordNet links words into semantic relations including synonyms, hyponyms & meronyms.

- WordNet includes lexical categories nouns, verbs, adjectives & adverbs but ignores prepositions, determiners & other function words.