UNIT III
— Language Modeling —

★ **Probabilistic language modeling** –
The probablistic language model is to compute a probability distribution of ~~a sentence~~ sequence of words in sentence. It essentially tries to estimate how likely a particula sequence of words is to occur in a natural language It is a statistical approach

> Ngram model – One common approach to language modeling is the n-gram model – (considering the probability of a word based n preceeding words).

> The main task is to estimate the probabilities of word sequence based on training corpus. This involves occurrences of word sequences in the ~~Cor~~ Corpus

> Language models are evaluated using metrics such as perplexity, which measures how well the model predicts, Lower perplexity better performance.

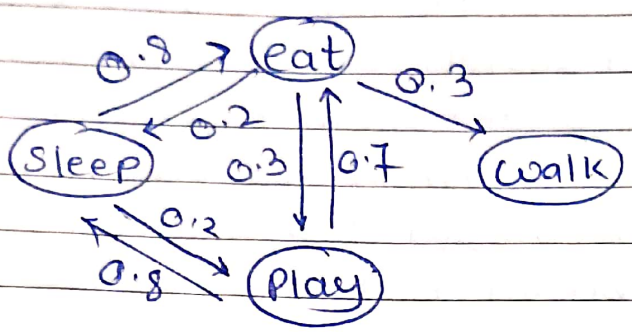
> Application –
Speech Recognition
Machine Translaⁿ
Text prediction
Summarizaⁿ

★ **Markov Models** – Markov model is a Stochastic model that describes a sequence of events where the probability of each event depends on the state of the previous event

> **Markov chains** – Simplest type of Markov model and are used to represent systems where all states are observable, It shows all possible states and between states, they show the transition rate, which is probability of moving from one state to other.

Teacher's Sign.

Consider you have a pet cat in morning Cat goes for walk, eat, play or sleep These are the states



Suppose 1) probability of going @ sleep to eat is very high
2) Probability of going from eating to play is moderate.
3) probability of going from eating to walk is low

Key idea is the probability of next state depends only on Current state & not on history.

> Hidden Markov model - Used to represent Systems ~~where all states~~ with Some unobservable States In addition to showing States and transition rates. Hidden states represent underlying Coditions or Situations in the System. hidden Markov models ~~are used for let's~~ also represent observation and @ likelihood ~~for~~ for each state.

Lets consider example - Suppose your friend lives in another city and you want to predict his/her mood based on city's weather So the ~~hidden~~ States are so the weather - Sunny, cloudy and rainy and ~~ab hidden~~ the moods happy & sad.

Teacher's Sign.: _____

observable

weather ~~cod~~ condition are ~~hidden~~ states which are ~~not~~ and his/her mood is hidden state

~~If your friend's mood transitioned from happy to sad then their is ,~~

If your friend's mood is happy then there's is the high probability that weather is sunny and if mood changed to sad that means there's high probability of weather changed from sunny to rainy.

Application -
- Speech recognition
- Pos-tagging

Generative models of language -

Generative models of language are statistical model used in NLP to generate new text that resembles ~~that~~ patterns and structure of a given language & use that knowledge to generate new text.

> Generative models are trained on large amount of text like books, articles, websites, etc., these models are based on probabilistic modeling techniques.

They learn probability distribⁿ of words phrases in language from a large corpus of text data.

> They assign probability to sequences of word based on their occurrence in training data.

> Ngram models popular type of generative model that predicts probability of a word based on previous n-1 words

> Deep learning based language models like RNNs LSTM, BiLSTM that captures long dependencies and learn to predict new sequence.

Teacher's Sign.: _____

Log linear models -

* N gram model - These are type of probabilistic model used in NLP & ML, They predict the probability of a word given the previous n-1 words in a sequence of text, widely used for tasks such as Speech recognition, Machine translation

> N gram can be defined as the contiguous sequence of n items from a given sample of text or speech.

> The probability of word given in its context is estimated using maximum likelihood estimation.

$$P(w_n | w_{n-1}, w_{n-2}, \ldots w_{n-N+1}) = \frac{\text{no. of time item appears}}{\text{Total no. of ngrams observed.}}$$

> Unigram mode ($N=1$) - each word in the text is considered independently of its surrounding words.

ex - Cat Sat on mat

['cat', 'sat', 'on', 'mat'] each token considered independently for predicting next sequence.

ex - Probability distribution -

$$P(W) = \prod_{K=1}^{n} P(w_K)$$

P (The man from jupyter came)

P (the) P (man) P (from) P (jupyter) P (came).

> Bigram - Probability of word is conditioned on the previous word in sequence

ex Probability distribution = $P(W) = \prod_{K=1}^{n} P(w_K | w_{K-1})$

ex P (The man from jupyter came)

P (The |<>) P (man | The) P (from | man) P (jupyter | from) P (came | jupyter)

Teacher's Sign.:

> Trigram (N=3) = Probability of word conditioned on two previous words. It captures more context.
Probability distribu^n -
$P(W) = P(W_K | W_{K-1} W_{K-2})$
P (the man from jupyter came)
P (the|<>) P (man|<>the) P (from|the man)
~~P (jupyter|fromman from~~
P (jupyter|from man) P (came|jupyter from).

> Estimation parameters -
1) Maximum Likelihood Estima^n (MLE) -
It is a basic approach to estimate the probabilities of N-grams based on their observed frequencies in the training data
$P(W_N | W_{N-1}, W_{N-2}, \ldots W_1) = \frac{\text{Count of item observed.}}{\text{Total no. of grams observed}}$
However, MLE can lead to zero probabilities for unseen N-grams which can cause model to perform poorly on unseen date.

> Smoothing Techniques -
1) Add one Smoothing - Simple technique that adds one count to all ngrams including unseen ones. These ensures there's no zero probability distribu^n.
2) Laplace Smoothing - Adds fixed amount to the ngram and divides this new count by total number of n-grams plus the no. of possible n-grams.

Teacher's Sign.:

> Evaluating N-gram model -

1) Perplexity - common measure used to evaluate performance of N-gram model held out ~~era~~ on a test dataset. measures how well the model predicts. lower value better performance.

2) Cross validaⁿ - Dividing dataset in training, testing & validaⁿ and evaluate final model on test data.

Q Suppose you have a text corpus of 10,000 words, and you want to build a bigram model. Vocabulary size is 5000. After counting bigrams in corpus, you found that the bigram "the cat" appears 50 times while unigram 'the' occurs 1000 times and unigram 'cat' appears 100 times. Using the add k smoothing method with k = 0.5 what is probability of sentence "the cat sat on mat"?

— "the cat" = 50 times, the = 1000 times
"Cat" = 100 times vocabulary size = 5000

Probability of each Bigram - ~~P(the cat)~~ =

$$P(\text{the cat}) = \frac{\text{count}(\text{the cat}) + k}{\text{count}(\text{the}) + k \times \text{vocab-size}}$$

$$= \frac{50 + 0.5}{1000 + 0.5 \times 5000} = \frac{50.5}{2750} \quad \frac{50.5}{3500}$$

$$P(\text{Cat sat}) = \frac{\text{Count}(\text{Cat sat}) + k}{\text{Count}(\text{Cat}) + k \times 5000} = \frac{0 + 0.5}{100 + 0.5 \times 5000} = \frac{0.5}{2750} \quad \frac{0.5}{2600}$$

$$P(\text{Sat on}) = \frac{\text{Count}(\text{Sat on}) + 0.5}{\text{Count}(\text{Sat}) + k \times 5000} = \frac{0 + 0.5}{0 + 0.5 \times 5000} = \frac{0.5}{2500}$$

Teacher's Sign.: _____

$$P(\text{on the}) = \frac{\text{count}(\text{on the}) + 0.5}{\text{Count}(\text{on}) + 0.5 \times 5000} = \frac{0 + 0.5}{0 + 2500} = \frac{0.5}{2500}$$

$$P(\text{the mat}) = \frac{\text{count}(\text{the mat}) + 0.5}{\text{Count}(\text{the}) + 0.5 \times 5000} = \frac{0 + 0.5}{1000 + 0.5 \times 5000} = \frac{0.5}{2750}$$

$$= \frac{0.5}{3500}$$

$P(\text{Cat Set on the mat}) = P(\text{the cat}) * P(\text{cat Sat}) * P(\text{Sat on}) \times P(\text{on the}) \times P(\text{the mat})$.

$$= \frac{50.5}{3500} \times \frac{0.5}{2600} \times \frac{0.5}{2500} \times \frac{0.5}{2500} \times \frac{0.5}{3500}$$

$$= 0.0168 \times 0.00019 \times 0.0002 \times 0.0002 \times 0.00014$$

Bag of words - (BoW)

BoW model is common approach used in NLP for representing text data. In this model, a document is represented as a bag (unordered collecⁿ) of words, where each word's presence or absence in the document is used as a feature. order of words is ignored & only their frequency matters

Example - Review 1 - The movies is very Scary & long

Review 2 - This movie is not Scary and is slow

Review 3 - This movie is Spooky and good.

First step tokenization of all unique words in 3 reviews

	This	movie	is	very	Scary	and	long	not	slow	Spooky
Review 1	1	1	1	1	1	1	1	0	0	0
Review 2	1	1	2	0	0	1	1	1	1	0
Review 3	1	1	1	0	0	0	0	0	0	1

Teacher's Sign.: _____

In this Vector representation frequency of each unique word token from Review are shown. Since most documents contain Smaller Subset of words in the vocabulary resulting in Sparsity in vector presentation.
> It ignores order of ~~Cont~~ & Context of words.
Introduces Sparsity
Doesn't capture semantic relationships.

TFIDF (Term Frequency Inverse Document Frequency) -
TFIDF is a statistical weighting Scheme used in Info retrieval & NLP. Aims to ~~assist~~ asses the importance of a word to a document in collection.
> Term Frequency - measures how often word appears in a specific document. higher TF indicates the word is more frequent in that document.
> IDF - ~~He~~ considers how common a word is across all documents in the corpus. Words that appear in many documents will have lower IDF weight. Words that are more specific & distinguish document will have higher IDF weight.

$$TF = \frac{\text{Freq of term 't' in document 'd'}}{\text{Total terms in document 'd'}}$$

$$IDF = \log\left(\frac{\text{total no. of documents}}{\text{total document with term 't'}}\right)$$

$$TF\text{-}IDF = TF \times IDF.$$

Teacher's Sign.: _____

① Given document - term matrix.

	DOC 1	DOC 2	DOC 3
T_1	10	5	0
T_2	2	0	8
T_3	1	3	6

Calculate TF-IDF for Term T_1 in DOC 1.

TF for Term (T_1) in DOC 1

$$TF = \frac{(\text{term } T_1 \text{ in Doc1}) \text{ freq}}{\text{Total terms in doc1}} = \frac{10}{18} \approx \frac{10}{13} = 0.769$$

$$IDF = \log\left(\frac{\text{total no. of docs}}{\text{total docs with term } t_1}\right) = \log\left(\frac{3}{2}\right) = \log 1.5$$

$$TF\text{-}IDF = 0.77 \times \log 1.5 \quad (\log(1.5) \approx 0.405)$$

$$= 0.3119$$

Word2vec – Popular word embedding technique used in NLP to represent words as dense vector in continuous vector space. These Word embedding eventually help in establishing the associa^n of a word with another similar meaning word. These

> These vectors capture semantic similarities between words, allowing more efficient and meaningful processing of textual data by ML algos

> The key idea is to learn word embedding by training a shallow neural network on a large corpus of text data resulting word embedding encode semantic relationships between words based on their co-occurrence patterns in the text.

> Word2vec utilizes 2 architecture

1) Continuous BOW - predicts a target word based on its surrounding context words. Observes sequence and tries to predict the central word based on words before & after.

2) Skip gram - Predicts surrounding context words based on given word. Model observes a single word and tries to predict the surrounding words.

* Doc2vec - Extension to Word2vec, used to learn distributed represent" of entire documents, allowing represent" of documents as continuous valued vectors.

> Word2vec focuses on individual words, Doc2vec aims to capture semantic meaning & topic of an entire document.

> Doc2vec adds an additional vector called a document vector to the word2vec model's architecture

> 2 main Architecture -

1) Distributed memory (DM) - incorporated document vector alongside word vector

Similar to word2vec's skip gram or CBOW model predicts surrounding words based on given word or predict target based on context.

2) Distributed BOW - DBOW doesn't involve predicting surrounding words. It focuses on predicting the document a given set of words belong to word vector. Helps to understand how words collectively contribute to the overall document meaning.

To be continued.....

Teacher's Sign.:

Contextualized representaⁿ (BERT) -

BERT which stands for Bidirectional Encoder Representaⁿ From Transformer, is a powerful technique for generating contextualized word embedding in NLP. BERT considers the surrounding word ~~regardd~~ to create context-aware representaⁿ

> Contextual word Embedding - BERT generates word embedding that captures the meaning of a word based on its surrounding words in a sentence. This allows the model to understand language and how words can have different meanings depending on the context.

> Bidirectional learning - Unlike traditional methods that analyze words sequentially (left to right), BERT utilizes bidirectional approach. Considers both the left and right context of a word simultaneously.

> working -

1) Input representaⁿ - words in sentence are converted into numerical tokens

2) Positional Encoding - BERT analyzes words bidirectionally it's crcial to understand relative position of each word in sentence. Positional Encoding adds INFO about t word's position to its embedding

3) Transformer Encoder - Core of BERT is multilayered transformer encoder each layer takes previous layer's output and perform 2 subtasks -

1) Multi head attention - Allows model to attend different parts of the input sentence simultaneously

2) Feed Forword Network - adds Non-linearity to the model, allowing it to learn complex relationship.

Teacher's Sign.: _____

> Benefits of BERT -

1) Improved Performance - (Bidirectional)
2) Contextual understanding
3) Transfer learning

> limitaⁿ

1) Computational cost
2) Data requirement.
3) Black Box Nature.

★ Latent Dirichlet Allocation (LDA) -

> Topic modeling - It is a method for unsupervised classificaⁿ of documents. Similar to clustering on numeric data, which finds some natural groups of items (topics) even when we're not sure what we're looking for

> Topic modeling helps to uncover latent theme or topics present in a collection of documents by identifying common patterns of words across doc.

> LDA automatically identifies latent topics automatically. LDA uses probabilistic approach to model documents as mixtures of these latent topics.

> LDA estimates the probability distribuⁿ of words for each topic.

> Working -

1) Data preprocessing - Involves tokenizaⁿ, removal of stopword
2) Modeling document as Topic mixtures - Each doc is assumed to be mixture of topics.
3) Probabilistic Distribution - 2 main probability distribuⁿ.
   a) Document - Topic distribuⁿ
   b) Topic - word distribuⁿ

Teacher's Sign.: _____

> ~~LDA applies a probabilistic distribu^n~~
and genera After preprocessing we get document word matrix LDA converts these matrix into 2 other matrix Document Topic matrix & Topic-word matrix.
> LDA involves estimating these probability distribu^n using an iterative algo, algorithm refines the topic assignments for each word and topic distribu^n.

> Benefits - Topic Discovery, Doc explora^n, summariza^n Doc classification.

Latent Semantic Analysis (LSA) is a NLP technique that analyzes the relationships between documents and their terms.
> It uses a mathematical technique called Singular value decomposition (SVD) to scan unstructured data & and find hidden relationships between terms & concepts.
> Text data can be high-dimensional due to large vocabulary LSA uses SVD to reduce dimensionality while preserving important ~~semanti~~ Semantic relationships.
> The first step involves creating a document-term matrix. It represents frequency of each term.
> Singular value Decomposi^n - SVD applies to the document term matrix, this decomposes the matrix in 3 components -
U - A left Singular vector representing documents.
Σ - ~~Dig~~ Diagonal matrix containing singular values
V^T - A right singular vector matrix representing terms.

Teacher's Sign.: _____

> LSA retains only the top k singular values and corresponding columns of U and V, effectively reducing dimensionality of the data.
> LSA calculates the similarity between documents or terms based on the cosine similarity between their vector representation.
> Benefits -
Improved Info retrieval
Summarization
Doc clustering
> limitation
Interpretability
Computational cost
polysemy and Homonymy

UIV

Info retrieval using NLP

Date
Page

* Information Retrieval (IR) -

IR in NLP involves the retrieval of relevant info from a large corpus of text document in response to user queries. It's a fundamental task in NLP and used in various applicaⁿ including web search, document retrieval system & text mining.

Process -

> Document collection - Involves collection ~~may~~ of web pages, articles, books, emails etc.

> Query processing - Interpretaⁿ & analysis of user queries to extract relevant info. This may involve. parsing, tokenizaⁿ & pos tagging. to understand user's query

> Indexing - Creation of data structure (indexes) that facilitate efficient retreival of documents based on their content. It represents docs in structural format for fast lookup & retrieval.

> Retrieval - Involves Searching the document collection based on the query ~~of~~ & ranking of the retrieved documents based how well they meet user's info need

Working -

> User Submit query
> Query understanding
> Document search
> Relevance ranking
> Result representaⁿ

Teacher's Sign.:

Vector Space Model (VSM)

VSM is a fundamental framework used in NLP and IR to represent text documents and queries as vectors in a high dimensional space. It provides a flexible & efficient way to calculate similarities between documents & queries.

Process -

1) Document preprocessing
2) Document Term matrix - matrix is created with rows representing docs and columns representing terms containing weight of term eg TF-IDF
3) Query vector - user's query also represented into vector space considering weight of each term.
4) Similarity calculation - It measures cosine similarity to compare query vector with each document vector. Doc with higher score considered to be more relevant.
5) Document Ranking - Based on similarity score documents are ranked.

Named Entity Recognition (NER) -

NER is sub-task of info retrieval in NLP that classifies Named entities into predefined categories such as persons, names, organizan, location

> NER serves as bridge between unstructured text and structured data, enabling machine to shift through vast amounts of textual info & extract valuable data in categorized form

> NERs objective is to go through unstructured text to identify specific chunks as named entities subsequently classifying them into predefined categories.

Teacher's Sign.: _____

> Working -

1) Tokenization -
2) Entity identificaⁿ - using various linguistic rules or statistical methods, potential named entities are detected. Involves recognizing patterns such as Capitalizaⁿ in names or specific format
3) Entity classificaⁿ - entities are identified, they are categorized into predefined classes such as person, organizaⁿ or Location. Often acheived using ML models trained on labeled dataset.
4) Contextual analysis - NER Systems often consider surrounding context to improve accuracy
5) Post-processing - Post processing might applied to refine results. This could involve resolving ambiguities, merging multi token entities.

Evaluating NER Systems -

1) Precision - Measures the proportion of predicted named entities that are actually correct. Calculates True positive divided by the sum of true positives & false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

2) Recall - Measures proportion of correctly identified named entities among all true entities

$$\text{Recall} = \frac{TP}{TP + FN}$$

3) F1 Score - Harmonic mean of precision & recall

$$\frac{2\,(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Teacher's Sign.: _____

4) ~~Entity level metrics~~ Accuracy - Evaluate system's ~~performance prediction~~ overall correctness in predictions. Calculate no. of correct predicⁿ divided by ~~sum of~~ total no. of predictions.

5) Entity level - Evaluate system's performance at the entity level, considering both span and type of named entities.

6) Macro vs Micro Average - Calculate metrics per each type individually ~~or ta~~ (macro) or take the overall averages across all entity types (micro).

(RE)

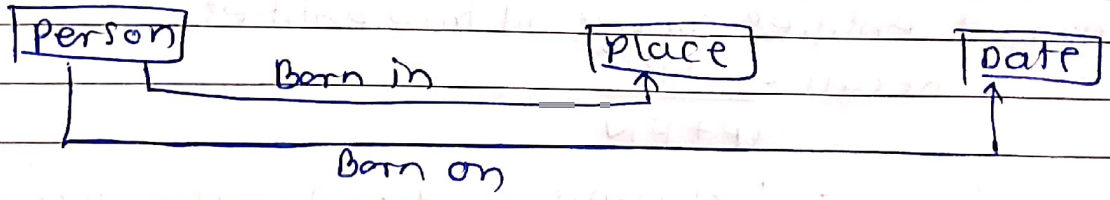
* Relation Extraction - It is a technique that helps understand the connection between entities mentioned in text

> RE is an imp. process in NLP that automatically identifies and categorizes the connections between entities within natural language.

> These entities can encompass individuals, organizaⁿ, locaⁿ, dates or any other nouns. The relationships denote how these entities are related to each other like Founder, location, works at, etc.

> example.

Steve jobs was born in San Francisco on 24th feb



Teacher's Sign.: _____

> Approaches in RE

1) Rule based – Relies on predefined ~~report~~ rules or patterns based on syntactic & or semantic structures in text to identify relationships.
Analyzes grammatical structure of sentence to identify depencies between words.

2) Supervised RE - RE is treated as a classification problem, where entities & their relations are annoted in a training entities
Uses Supervised techniques like logistic regression, random forest, RNNs.

3) Unsupervised RE – Aims to identify relationships in text without labeled training data, Represents relations as string of words extracts them directly from text without relying on prior knowledge or annoted examples.

★ Reference Resolution – It is a task in NLP that involves identifying and resolving references to specific entities or concepts in text. The goal of reference ~~resolu~~ resolution is to clarify the meaning of pronoun demonstrative or other referring expressions by identifying entity or concept

> Types of References –
1) Pronoun (eg. he, she, it)
2) Demonstrative (eg. this, that, those)
3) Definite noun phrase (eg. the book, the car)

> The primary objective of reference resolu^n is to disambiguate and identify the entity or concept to which a reference refers. By resolving references, the meaning of the text becomes clearer.

Teacher's Sign.: _____

> Expression consists of pronouns (he, she, it), noun phrase or distrip^{n}.
> Coreference - Relationship bet^{n} these expressions when they refer to the same entity.
> Anaphora - Referring expression comes later in sentence eg. John went to store, He bought milk
> Cataphora - Le referring expression comes before the 1st mention of the entity.
eg He seemed happy. John finally won the game.

Coreference Resolution - It is a task in NLP that involves identifying and clustering mentions of the same entity and replaces pronouns with noun phrases.
example - "I gave my laptop to Andrew", peter said.
first it groups the words into several groups by considering entities.
Andrew
peter
peter's laptop.
After it replaces all the pronouns in the sentence with relevant nouns.
'Peter gives Peter's laptop to Andrew."
> Types of Coreference -
> Pronominal - Refers to pronoun that stands in for previously mentioned entity
2) Nominal - Refers to common noun phrases that refer to the same entity
3) Anaphora
4) Cataphora

Teacher's Sign.:

* Cross lingual info retrieval - (CLIR) -
Subfield of IR deals with retreiving info written in different language from the user's query language
> CLIR is useful in situation where the information is not in the user's native language.
> Challenges - Fundamental challenge in CLIR is bridging gap between the user's query language and the document language. This can be difficult because languages have different vocabularies, grammer & semantics.
> Techniques -
1) Dictionary based - This approach uses bilingual dictionaries to translate query term into the document language
2) Parallel corpora based - Collections of documents that exist in both the query language and the document language. By analyzing these parallel texts the system can learn how words phrases correspond across language.
3) Machine Translator based - Machine translation can be used to translate entire query or document to bridge the language gap.

> challenges -
1) language gap
2) Data scarcity - training CLIR requires large amount of data.
3) Machine transla^n limitation.

Teacher's Sign.: _____

Unit V

NLP Tools & Techniques

NLP Tools

1) Natural Language Tool kit (NLTK) -

NLTK is a popular open source library for python that provides a powerful suite of tools for working with human language data. It provides tools, algorithms and resources for tasks such as tokenization, part of speech tagging, Parsing, Sentiment Analysis.

> NLTK supports a wide range of languages, not just English. It provides tokenization, stemming and morphological analysis tools for languages such as Arabic, chinese, Dutch, French, japanese, Russian & more.

> NLTK can also be used in machine learning libraries such as Scikit learning & Tensorflow

> Features of NLP -

1) Morphological processing - (Tokenizaⁿ) splitting up large input blocks into smaller groups of tokens
2) Text processing - Involves Functionalities for common text processing tasks such as Tokenizaⁿ, Pos tagging, Stop word removal, Stemming, lemmatizaⁿ.
3) Parsing - Analyzing grammatical structure of Sentences to determine their Syntactic relationships
4) Corpus Access - Provide Access to rich collection of pre-existing corpora (large collecⁿ of text data) that can be used for training and evaluating NLP models.
5) ML integraⁿ - NLTK integrates with popular ML libraries like Sci-kit learn, allowing to build your own NLP models for tasks like classificaⁿ, Sentiment analysis and NER.
6) Statistical Model - offers Functionality for statistical language processing tasks like n-grams.

Teacher's Sign.: _____

Spacy - Spacy is an open source NLP library designed for efficient and production ready NLP Task. It provides robust, fast and accurate processing for 8 tasks such as tokenizaⁿ, Pos tagging Named Entity recognition, parsing, etc.

Strengths -

1) Performance - Optimized for speed and memory usage allowing you to handle large datasets efficiently.
2) Pre-trained models - Spacy provides pre-trained models for various languages used to train on large corpus of data for tasks like NER, POS tag, dependency parsing
3) Flexibility - Allows you to fine tune pre-trained models or train your own custom models for specific NLP needs.
4) Language Support - Spacy supports multiple languages. This includes tokenization rules, Pos tagger and syntactic parser.
5) Named Entity Recognin - Identifying named entities Spacy's NER component. Spacy's NER can be customized to recognize domain specific entities.
6) Dependency parsing - Spacy's dependency parsing analyzes grammatical structure of a sentence & represents it as a dependency tree.
7) Tokenizaⁿ - dividing sentence into smaller units called tokens Spacy's Tokenizerⁿ is customizable for specific domain.

Teacher's Sign.: _____

TextBlob - It is a python library disigned to simplify NLP tasks. It offers a user friendly API that allows you to perform common NLP operations on text data with minimal code.

> It performs NLP tasks like tokenizaⁿ, POS-Tagging Noun p- Noun phrase extraction, lemmatization, N-grams and Sentiment analysis. It has more features such as Spelling correction, creating Summary Translation & language detection.

> Tokenization

> POS Tagging - Tagging is a kind of classificaⁿ that may be defined as automatic assignment of descripⁿ to the tokens. labeling each word in sentence with it's appropriate PoS tag such as Nouns, verbs, adverbs, adjectives, Pronoun, conjunction.

I	Like	to	read	Books
PRP	VBP	TO	VB	NNS

> Noun phrase extraeⁿ - TextBlob can extract noun phrase from text, identifying phrases that function as subjects or objects in sentences.

> Language Translation - It supports language translaⁿ allowing you to translate text from one language to another.

> Spelling correction - one of the unique features of the TextBlob library. With the correct method of TextBlob object you can correct all the spelling mistakes.

> Lemmatizaⁿ - Word object from TextBlob library used for performing lemmatizaⁿ.

Teacher's Sign.: _____

* Gensim - It is a popular open source library in python specifically designed for working with unstructured digital text. It excels in tasks related to topic modeling and semantic similarity.

> features

1) Vector space model - Gensim represents documents as vectors in a high dimensional space, where each dimension corresponds to a unique term or word in the vocabulary.
2) Topic Modeling - Gensim provides implementa" of popular topic modeling algos including LDA & LSA. These algos allows users to discover hidden topics within collection of documents.
3) Document Similarity - Gensim allows users to calculate similarity between documents based on their vector representa". Used for document clustering info retreival.
4) Word Embedding - Creates vector representa" of words that capture their semantic meaning and relationships with other words. Useful for tasks like sentiment analysis, machine transla".

> Gensim provides utilities for text preprocessing, including tokeniza", stopword removal and stemming
> Gensim is designed to be ~~scalabe~~ scalable, allowing users to process large datasets efficiently using techniques such as online training & distributed computing.
> Integra" with other libraries such as Numpy, Scipy, ~~sm~~, TensorFlow, Scikit-learn.

Teacher's Sign.:

* Lexical Knowledge Network – (LKN) –
These are type of knowledge resource that focuses on the relationships between words in language. They aim to capture complex web of meaning within a language by representing words, their concepts, and the connection between them.

> LKNs are typically conceptualized as graphs or networks where nodes represent words or terms and edges represent the relationship between them

> Relations may include Synonymy, antonymy, ~~hypoh~~ hyponymy and more.

> LKNs typically structured as graphs or networks ~~whe~~ ~~com~~ Connections relects different aspects of lexical knowledge.

> Benefits –
1) Disambiguaⁿ – LKNs can help computers disambiguate meaning of similar words depending on text.
2) Info retrieval
3) Machine Translaⁿ
4) Question Answering.

> Examples
WordNet, HowNet, FrameNet

* WordNet – WordNet is a lexical database & Semantic network of words in Natural language, organized into sets of synonyms, called Synsets and their Semantic relationships. used for tasks such as Word sense disambiguaⁿ, info retreival, text mining & Translaⁿ.

Teacher's Sign.: _____

> Features

1) The basic building blocks of WordNet are synsets which are sets of words that ~~are~~ ~~synonymous~~ Synonymous

2) Semantic relations - WordNet captures various Semantic relations including
   Hyponymy - heirarchical relationships
   Antonymy - opposite meaning.
   Synonymy - similar meaning

3) WordNet distinguishes betn different POS tags and provides seperate sets of Synsets for each POS category.

4) Word Sense Disambiguation (WSD) - WordNet is often used in ~~co~~ WSD tasks to determine the correct meaning of a word in a given context By leveraging semantic relationships

> Applicatn

1) Info retrieval, Machine Translation
   lexical database.

~~A~~ VerbNet - It is lexical database and heirarchical classification system for English verbs, focusing on their Syntactic & semantic property.

> VerbNet organizes verbs into hierarchies or classes based on their Syntactic & semantic property. Each class represents a group of verbs that share ~~same~~ similar syntactic frames.

> Syntactic ~~rule~~ frame - Each verb class is associated with a set of Syntactic ~~frame~~ frame, specifying the grammatical roles that verb can take. These include info about Verb's Subject, object, phrases etc.

Teacher's Sign.:

Semantic Role - VerbNet defines standardized set of semantic roles that verbs can assign to their arguments. Roles includes agent, theme, experiencer, location and other.

> Semantic classes - VerbNet also organizes verbs into semantic classes based on their roles.

> lexical Entry - Includes details like its class ~~to~~ its Syntactic Frame, Semantic Role and its meaning to understand how a verb works and how its used.

> Benefits

1) Verb Sense Disambiguation
2) Sentence understanding
3) Machine Translation

PropBank - Short for Proposition Bank, is a resource in NLP that annotates sentences with information about the semantic roles of words and their relationships within the sentence.

> Semantic roles - PropBank annotates words in sentence with semantic roles, such as agent, patient, theme location. These roles indicates the different functions that words play in expressing an event.

> Annotated sentences - Each sentence in propBank is annotated. for eg. "The dog chased the cat", "dog" might be annotated with role agent and cat might annotated as theme.

> Verb specific info - PropBank provides verb-specific info about the roles associated with each verb.

> Corpus of Text - PropBank is built ~~upon~~ on a corpus of text These allows for the extraction of patterns & statistics about verb.

Teacher's Sign.: _____

> Applicaⁿ – PropBank used for info extracⁿ, Question answering, & machine Translation.

✗ TreeBank – Treebanks are collection of Syntactically annotated sentences in NLP. They consist of sentences from a Corpus that have been annotated manually with ~~Syntac~~ Syntactic Structures, typically represented as parse Tree.

> These trees visualize the grammatical relations between words in a sentence

> Each Sentence is parsed and broken down into Constituents parts creating heirarchical ~~tr~~ tree that illustrates relation between words

> 2 main types of treebanks –

1) Syntactic – Focus on grammatical structure of a sentence, using labeles~~s~~ noun phrase, verb phrase, adjective phrase.

2) Semantic – Annotates sentences with a meaning representaⁿ, aiming to capture deeper semantic relations betⁿ words & concepts.

> Applicaⁿ – Parsing, Translation, text generaⁿ.

Teacher's Sign.: _____

Word Sense Disambiguation - (WSD) -

It is a task in NLP that aims to determine the correct meaning (sense) of word in given context. Many words in natural language have multiple meanings and determining intended meaning of a word in particular context is crucial.

> Importance
1) Improved accuracy
2) Reduced ambiguity.

> Approaches
1) Knowledge based
2) Supervised
3) semi supervised.

Lesk Algorithm - This Algo is one approach to WSD that relies on comparing the meaning of words in the context of the surrounding words in sentence.

> The algo first identifies the word in the sentence with multiple possible meaning
> It retreives the dictionary definitions for the ambiguous word and each word in a specific surrounding window
> The algo calculates the overlap between the definitions. The sense (meaning) of the ambiguous word whose definition share the most overlap with definitions of its neighbouring words is considered the most likely meaning in that context.
> It Replaces the ambiguous word in the context with its selected sense to disambiguate it.

> Advantage
Simplicity
Efficiency.

Teacher's Sign.:

Walker's Algorithm -

It uses Thesaurus as its primary tool, find the Thesaurus category to which the target word in each of its sense belongs
If context word's thesaurus category matches that of the sense, the sense score will increase by one.

WordNets for WSD -

WordNets act as lexical knowledge base that captures the semantic relations between words, making them valuable for understanding the different meanings a word have in different contexts.

> WordNets groups words with similar meaning called Synsets.
> Wordnets captures various semantic relations such as synonymy, antonymy, hyponymy and more. By leveraging these relations WSD algos can infer the likely sense of a word based on its connection.
> In Supervised learning approaches to WSD, WordNet can be used to create labeled training data by annotating sentences with their correct word senses.

UVI

NLP Application

✱ Machine Translation - (MT)

MT is the process of automatically converting text from one natural language to another. It involves using computational algorithms and model understand and generate equivalent expression in different languages.

> Approaches -

1) Rule based MT - (RBMT) It relies on linguistic rules and dictionaries to translate text from one language to another. Manually created rules governs the translation process including rules for grammar, syntax, and vocabulary RBMT analyzes the structure of the input sentence and apply the relevant rules to generate translation.

- Advantages - 1) High accuracy in specific domain
  2) Requires less computational resources
  3) offers transparency
- Disadvantage - 1) Limited scalability
  2) Struggles with handeling ambiguity
  3) Difficulty in capturing richness & variability in language

2) Statistical Machine Translation (SMT) -

- SMT uses statistical models ~~to learn~~ make model to learn pattern from large bilingual corpora, which is collection of paralled texts in the source and target language. SMT analyzes these corpora to derive poobablistic mapping between words, phrases or sentences in the source and target languages. It algns source and target language segment to learn translation pattern.

Advantage -> can handle coide range of language pairs and domains 2) Incorporate context from large bilingual corpora.
3) Provides a probablistic framework that allows for flexibility.

Disadvantage - 1) Relies heavily on the quality of corpora
2) May struggle with translating rare or unseen data.
3) Limited ability to capture long-range dependencies.

>SMT has 3 main components -
1) Language model - Captures probability of sequence of words
2) Translation Model - Estimate probability of translating language
3) Decoding Algorithm. - Searches for most probable ~~translating~~ transla" by combining probabilities.

3) Cross lingual Translation (CLT) - It refers to the process of translating text between languages that are not directly related or similar. Unlike bilingual translation which involves translating betn two specific languages. CLT deals with the translating across different languages. This can be challenging as languages have different structures, vocabularies.

>Techniques -
1) Both Rulebased MT and SMT can be used for multiple language Translation.
2) Neural Machine Transla" uses deep learning models to analyze bilingual text data. It can capture complex relationships between languages leading more accurate translation.
3) Lexical Mapping focuses on identifying corrensponding words or phrases with similar meaning across languages.
4) Transfer learning utilizes knowledge gained from translating one language pair to improve transla" for a different pair.

>Application -
1) Machine to Machine communica"
2) Multilingual info retrieval
3) Sentiment analysis in multilanguage.

> SMT has 3 main components -

Sentiment Analysis -

> Sentiment Analysis is also known as opinion mining is a NLP technique that involves extracting subjective info from text to determine the sentiment of the text.

> It aims to understand the attitude, opinions and emotions of individuals or groups towards specific topics by analyzing textual data.

> Process of sentiment analysis -

1) Text preprocessing - First step involves cleaning and preparing text for analysis it involves tasks such as removing punctuan, stopwords ~~and~~ removal, tokenizan and stemming, lemmatizan.

2) Feature Extracn - Relevant features are extracted from text data representing sentiment bearing content. This includes words, phrases, n-grams, syntactic pattern or semantic features.

3) Sentiment classifican - Extracted features are then used as input to machine learning ~~technique~~ models to classify sentiments in positive Negative. Supervised learning technique like SVM, Naive Bayes, Logistic regression. Deep learning architectures like RNNs can be used for training.

4) Sentiment Analysis Output - Once classifican is performed output can be categorized into sentiment classes such as positive, Negative or neutral. Additionally, sentiment analysis output may include sentiment scores probabilities.

> Applications - 1) Market Research
2) Customer service Improvement
3) Social media management.
4) Movie review.

> challenges -
1) Irony & Sarcasm
2) Context dependence
3) Multilingual Sentiment Analysis.

Question Answering - (QA)

QA is a field of AI and NLP that focuses on developing systems capable of understanding and responding to natural language question. These systems aim to extract relevant info from large amounts of unstructured data. Such as articles, books, news articles.

> Types of QA -

1) Retrieval Based ~~(QA)~~ QA - Focuses on finding relevant documents or passages contain the answer of question.
~~2) Reading comp~~
2) Generative QA - System generates answers by analyzing info from ~~not~~ multiple sources to generate new text.
3) Hybrid QA - Combine both retrieval based & Generative QA.

> Process of QA -

1) Natural Language Understanding (NLU) - QA systems must first understand user's question including Syntax, Semantics & Named Entity recognition. helps to extract relevant info
2) Info retrieval - Once user's question is understood, QA system retrieve relevant info from large text data. IR Techniques such as indexing, ranking, retrieval are used efficiently. QA systems analyze the retrieved text data to extract specific info relevant to user's question. These involves entity extraction, relation extraction.
3) Answer Generation - Based on extracted Info QA systems generate an answer that accurately addresses the user's question.

> Limitations

1) Ambiguity & understanding
2) Knowledge representation -
3) Scalability & Efficiency
4) Evaluation.

> Models - BERT, GPT, Ts can be trained on open source dataset like SQUAD QA dataset.

Natural Language Generaⁿ (NLG) -

NLG is a branch of AI and NLP that focuses on generating human like text from structured data or other forms of input. NLG systems aim to convert non linguistic representaⁿ such as data tables, graphs, or semantic representations, into ~~coherent and~~ contextually appropriate Natural Language text.

> NLG plays Crucial role in -

1) Text summarizaⁿ -
2) Data to Text Generaⁿ - Converting structured tabular data like data tables, spreadsheets into natural language.
3) Language translaⁿ -
4) Chatbots
5) Content Generaⁿ - NLG can generate personalized content such as articles, stories, poem etc.

> Working -

1) Data preprocessing - cleaning text, feature extraeⁿ, Semantic analysis understanding underlying meaning.
2) Content planning — NLG determines overall structure, content, style of generated text based on input. Decides which info to include and which to exclude.
3) Text generaⁿ - Generating Natural language based on approaches -
   1) Rule based generaⁿ - Follows predefined rules like grammar and syntax to generate the text.
   2) Statistical NLG - uses probabilistic models to generate text based on the statistical properties of input data.
   3) Neural language - Leverages neural net like RNNs LSTM or Transformer used for generating text.
4) Evaluaⁿ - NLG systems evaluated based on Criteria such as fluency, coherence, relevance, readability, Grammatically.

☆ Text Entailment - Also known as Natural language Inter~~Inter~~ Inference (NLI), is a task in NLP that involves determining whether one piece (hypothesis) logically follows another piece of text (the premise). It aims to assess the relationship between two sentences.

> The text entailment is typically framed as b classificⁿ problem, where given is hypothesis & premise sentence the goal is to predict whether the hypothesis is entailed by ~~(similar~~ (related) by premise (entailment), contradicts premise (contradicⁿ) or neutral.

For example. Premise - "The cat Sat on mat"

hypothesis 1 - "These is a cat on mat" (Entailment)

hypo 2 - "The cat is flying" (contradicⁿ)

hypo 3 - "The mat is blue" (Neutral)

> Text Entailment can be helpful in understanding relations and used in applicaⁿ like QA, summarizaⁿ, Info retrieval.

> challenges - Ambiguity, Sarcasm, irony

> Approaches -

1) Rule based
2) ML approach - Models are trained on large dataset
3) Deep learning approach - Advanced neural network models to capture complex relations between words.

☆ Discourse processing - it is a field within NLP that focuses on understanding and analyzing the structure and meaning of larger units of text such as paragraphs, conversations. Aims to understand how sentences work together.

> key tasks -

1) Discourse parsing - Involves analyzing the syntactic & semantic relationships betⁿ sentences within text.
2) Coreference resoluⁿ - Aims to identify & link expressions in a text that refer to the same entity. Includes Pronouns, phrases, etc

3) Anaphoric resolu^{n} - Specific type of coreference resolu^{n} that focuses on resolving references to previously mentioned entities or concepts.

4) Cohesion - Refers to the mechanisms that binds sentences together to create unified whole sentence. Pronouns used for refering previously mentioned entity.

5) Coherence - Signifies overall meaningfulness of discourse. Involves ensuring ideas presented are logically connected and flow smoothly for reader or listener.

6) Discourse relations - relationships that exist betn different parts of a discourse & generating text based on input Discourse units.

- else add the current item as much as we can and break out of the loop

5)Return res

CONCLUSION:-Implemented fractional knapsack using a greedy strategy successfully.

QUESTIONS FOR REVIEW:

1) Write realistic applications of this experiment in brief (at least two applications).?
2) Explain Kanpsack with example using greedy approach?

Item	Weight	Value
1	5	30
2	10	40
3	15	45
4	22	77
5	25	90

3)Analyse time complexity of knapsack ?

Natural language Generaⁿ (NLG) –

... NLP that focuses on generating

Dialog & conversational Agent –

Also known as AI or chatbots, AI systems designed to engage in natural language conversaⁿ with humans.

Characteristics.

1) Natural language generaⁿ (NLU)
2) Dialog Management - Involves maintaining & controlling the flow of ~~conse~~ conversation between user & Agent.
3) knowledge representaⁿ - Response generaⁿ technique may include template based response, rule based response or generative based response.
4) personalizaⁿ - may leverage user profile, historical interaction & contextual info to tailor response and provide more personalized assistance.
5) Evaluaⁿ & Improvement.

Applicaⁿ –

1) Customer Service.
2) A Virtual Assistants.
3) Education
4) Healthcare