_/_/___

JUNAID · GIRKAR

NATURAL LANGUAGE PROCESSING 60004190057

ASSIGNMENT - 2 BE COMPS A2

Q1 Bag of words and its types

ANS → Bag of words (BoW) is a commonly used technique in NLP for representing text data as numerical vectors. In BoW, a text document is represented as an unordered collection or "bag" of words, disregarding grammer and word order. The frequency or presence of each word in the document is used to create a numerical representation.

There are 2 main types of BoW :-

1] BINARY BOW

In this type of BoW model, the presence or absence of a word is represented by binary values, typically '1' for presence and '0' for absence. It only consider whether a word appears in the document or not, disregarding the frequency of occurance.

2] COUNT BoW

In this model, the frequency of each word in the document is considered. It represents the document as a vector of word counts, where each element in the vector corresponds to the count of a specific word in the document.

Q2 Explain pragmatic analysis using a real life example.

ANS Pragmatic analysis, in the context of linguistics, involves studying the ways in which language is used in real-life situations and the effects it has on communication. It focusses on the intended meaning behind utterances, taking into account context, speaker intention, and the shared knowledge between the speaker and the listener.

EXAMPLE: Imagine a group of friends go out to dinner. "Let's go to that new Italian restaurant tomorrow'

ANALYSIS:

In this example, pragmatic analysis would involve examining the utterance in light of the context and -shared knowledge among the friends

1. CONTEXT: The context includes the current situation and previous conversation. What led to the suggestion of going out for dinner? Have they been discussing different types of cuisines and restaurants.

2. SPEAKER INTENTION: The speaker's intention might be to propose a specific option for the group to consider. They may have prior knowledge or heard positive reviews about the new italian restaurant downtown, which they find appealing and want to share with the group.

_/_/_

3. INFERENCE: The listener would engage in inference to interpret the utterance.

4. SHARED KNOWLEDGE: Pragmatic analysis takes into account shared knowledge between the speaker and the listener. This could include cultural norms, background information and assumptions about the group's preferences

Q3 what is lesk Algorithm?

ANS Lesk algorithm is a word sense disambiguation algorithm developed by Michael lesk in 1986. word sense disambiguation is the task of determining the correct meaning of a word in a given content

The lesk algorithm works by comparing the definations of words in a given content with the definations found in lexical database, such as wordNet. It assumes the meaning of a word can be inferred based on the overlapping words in its defination and the content in which it appears.

Outline of lesk Algorithm:-

1. Identify the target word that needs disambiguation and obtain its surrounding content. The content usually consists of a fixed window of words, such as the words within a certain number of words before and after the target word.

__/__/____

2. Retrieve the definations of the target word and its surrounding words from a lexical database, such as WordNet. WordNet provides synsets, which are sets of synonymous words representing different senses of a word.

3. Calculate the overlap between the words in the definations and the words in the content. The lesk algorithm typically uses a simple overlap measure, such as counting the number of shared words or using a more sophisticated measure like the Jaccard coefficient.

4. Select the sense of the target word with the highest ~~senses~~ overlap as the disambiguated sense. If there are multiple senses, with the same highest overlap, additional heuristics can be applied to make the final decision.

__/__/____

Q4 HMM VS MEMM

	FEATURE	HMM	MEMM
1	Model Representation	Generative model, assumes observed data is generated by hidden process.	Discriminative model directly models the conditional probability of the output sequence given the input sequence.
2	Model Dependencies	Assumes Markov property, current state depends on previous state.	captures more complex dependencies betn input & output sequences.
3	Training & Inference	Training involves estimating transition & emission probabilities. Inference is performed using Viterbi algorithm.	Training involves estimating the parameters of the models using max entropy principles, often through techniques like iterative or gradient-based optimization. Inference = Viterbi
4	Handling features	Feature set is fixed & predefined.	Feature set is typically manually engineered.
5	Overcoming label bias	Prone to label bias, where it may assign high probabilities to incorrect labels due to its generative nature and simplified modelling assumptions.	can alleviate label bias problem as it directly models the conditional probability of the output sequence