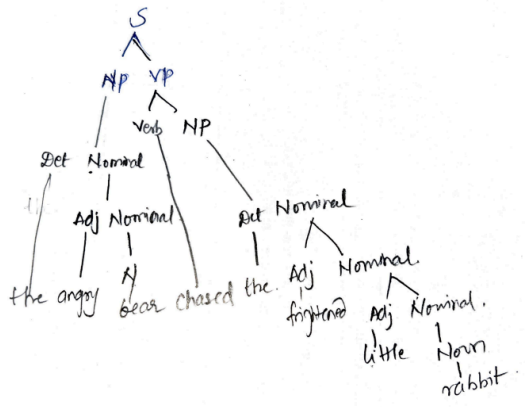
Vivekanand Education Society's Institute of Technology, Chembur, Mumbai,
Department Of Computer Engineering,
Year:2023-24 (ODD Sem)
MID TERM TEST

Class : BE	**Division: A,B.C**
Semester: VII	**Subject: Natural Language Processing**
Date: 6/9/2023	**Time: 1 hr**

Course Outcome	CO1	CO2	CO3	CO4	CO5	CO6
Percentage %	30%	35%	35%	-	-	-

Q.1)		(Attempt any five of the following)
	a)	Why is smoothing required? List different techniques for smoothing. Smoothing is designed to address the issue of words or events that are part of the vocabulary but appear in an unseen context, where traditional statistical models might assign them a zero probability. This situation often arises when dealing with language models or other probabilistic models of text.When words or events occur in a test set but in a previously unseen context, a standard probability model might fail to provide meaningful predictions. To mitigate this issue, smoothing, also referred to as discounting, is employed. Smoothing techniques redistribute a small fraction of the probability mass from more frequent events to less frequent or unseen events. This redistribution helps ensure that no event is assigned a zero probability, which is essential for the robustness and practicality of probabilistic models. Types of smoothing techniques, ● Add-one (Laplace) smoothing, ● Interpolation smoothing ● Good-Turing smoothing, and ● Katz smoothing
	b)	Use the following grammar rules: S → NP VP Det → the NP → Det Nominal Adj → little | angry |frightened VP → Verb NP N → rabbit | bear Nominal → Adj Nominal |N V → chased **Create a parse tree for the following sentence** **"The angry bear chased the frightened little rabbit."**

		
	c)	POS tag the following sentences **1.She enjoys running in the park.** She - PRP (Personal Pronoun) enjoys - VBZ (Verb, 3rd person singular present) running - VBG (Verb, gerund/present participle) in - IN (Preposition) the - DT (Determiner) park - NN (Noun, singular or mass) So, the POS-tagged sentence is: "She (PRP) enjoys (VBZ) running (VBG) in (IN) the (DT) park (NN)." 2.The cat is sleeping on the mat. The" - DT (Determiner) "cat" - NN (Noun) "is" - VBZ (Verb, 3rd person singular present) "sleeping" - VBG (Verb, gerund/present participle) "on" - IN (Preposition) "the" - DT (Determiner) "mat" - NN (Noun)
	d)	List the ambiguities in natural language processing?. Identify the ambiguity present in the following sentences: i) Police were chasing the man with a bat. **Lexical Ambiguity:** Explanation: In this interpretation, the word "bat" is lexically ambiguous. It could mean a baseball bat or a flying nocturnal mammal. **Syntactic Ambiguity:** Explanation: This sentence is syntactically ambiguous because it's unclear whether "with a bat" is an adverbial phrase describing how the police were chasing (using a bat) or if it's an adjectival phrase describing the man (the man had a bat). **Semantic Ambiguity:** Explanation: In this case, the ambiguity is semantic. It's unclear whether "bat" refers to a baseball bat or a flying mammal, which can drastically change the meaning of the sentence.

		ii) My sister doesn’t use glasses. **Lexical Ambiguity:** Ambiguity: The word "glasses" can refer to eyeglasses (for vision correction) or drinking glasses (for beverages). Interpretation 1: "My sister doesn't wear eyeglasses for vision correction." Interpretation 2: "My sister doesn't use drinking glasses for beverages." **Semantic Ambiguity:** Ambiguity: The sentence can be semantically ambiguous because "use" can have multiple meanings. It could mean "wear" in the context of eyeglasses or "utilize" in the context of drinking glasses.
	e)	Why is lemmatization better than stemming ? Justify your answer with a suitable example. Compared to stemming, which simply removes suffixes from words to reduce them to a common form, lemmatization produces a more linguistically accurate and meaningful representation of words. Stemming can sometimes result in non-existent words or words that do not convey the intended meaning, whereas lemmatization retains the integrity of the word.
	f)	Discuss any two NLP applications and their challenges. **<u>Machine Translation:</u>** Machine translation involves the automatic translation of text or speech from one language to another. It's widely used for translating content on the internet, facilitating communication across language barriers, and supporting global business operations. Challenges: **Ambiguity:** Different languages have varying word orders, idiomatic expressions, and grammar rules. **Low-resource languages:** Machine translation systems tend to perform better for languages with ample training data. Low-resource languages, with limited available data, pose a significant challenge for achieving accurate translations. **Context:** Translation often depends on the context, and machine translation models might struggle to capture nuanced meanings, cultural references, or context-specific language use. **<u>Sentiment Analysis:</u>** Sentiment analysis, also known as opinion mining, aims to determine the sentiment or emotion expressed in a piece of text, such as a social media post, customer review, or news article. It's used in brand monitoring, customer feedback analysis, and market research. **Challenges:** **Sarcasm and Irony:** Detecting sarcasm, irony, or subtle nuances in text can be difficult for sentiment analysis models. These forms of expression often rely on context and tone, which may be challenging to capture. **Data Quality and Bias:** Sentiment analysis models heavily depend on the quality and representativeness of training data. Biases present in the data can result in biased sentiment predictions and reinforce stereotypes. **Multilingual Analysis:** Extending sentiment analysis to multiple languages is challenging, as the same words or phrases may convey different sentiments in different languages. Building models that work well across diverse languages is an ongoing challenge.

Q.2)	a)	Discuss various stages involved in NLP with suitable example. 1. **Lexical Analysis and Morphological** The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences, and words. **2. Syntactic Analysis (Parsing)** Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.**Example:** "Agra goes to the Poonam". In the real world, Agra goes to the Poonam, does not make any sense, so this sentence is rejected by the Syntactic analyzer. **3. Semantic Analysis** Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences. **4. Discourse Integration** Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it. **5. Pragmatic Analysis** Pragmatic analysis is a crucial aspect of natural language processing (NLP) and linguistics that focuses on understanding the context-dependent aspects of language use. It deals with the interpretation of language beyond its literal or grammatical meaning, taking into account factors such as speaker intentions, implied meanings, conversational implicatures, and the effect of context on language comprehension. **For Example:** "Open the door" is interpreted as a request instead of an order.
		OR
	b)	Justify the use of FST as a parser and as a generator with a suitable example. Formally, a finite transducer T is a 6-tuple $(Q, \Sigma, \Gamma, I, F, \delta)$ such that: • Q is a finite set, the set of *states*; • Σ is a finite set, called the *input alphabet*; • Γ is a finite set, called the *output alphabet*; • I is a subset of Q, the set of *initial states*; • F is a subset of Q, the set of *final states*; and • $\delta \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\}) \times Q$ (where ε is the empty string) is the *transition relation*. Recognizer: is_pasttense_verb(bought) → true Parser bought ⟶ buy + V
Q.3)	a)	Assume a bigram language model is trained on the following corpus of sentences . <s> My name is Merry</s> <s> Merry my name is </s> <s> A girl said that her name is Merry </s> <s> My daughter's name is Merry </s> **What is the estimated bigram probability of the following:** **i) P(<s> A girl name is Merry</s>)** **ii)P (name | my)** **Ans : (i) 0 ii)2/3**
		OR

	b)	Compare Rule based and stochastic POS taggers.. **Rule based POS taggers.** A rule-based POS (Part-of-Speech) tagger is a system for assigning POS tags to words in a text based on a set of predefined linguistic rules and patterns. Instead of relying on statistical probabilities or machine learning models, rule-based taggers use explicit rules and heuristics to make tagging decisions. **Stochastic POS taggers.** Stochastic POS (Part-of-Speech) taggers, also known as probabilistic taggers, are a type of POS tagging system that assigns POS tags to words in a text based on the statistical likelihood of a word having a particular tag, as learned from a training corpus. Here are some key characteristics and features of stochastic POS taggers: **Stochastic tagger applies the following approaches for POS tagging −** **1. Word Frequency Approach** In this approach, the stochastic taggers disambiguate the words based on the probability that a word occurs with a particular tag. We can also say that the tag encountered most frequently with the word in the training set is the one assigned to an ambiguous instance of that word. The main issue with this approach is that it may yield inadmissible sequence of tags. **2. Tag Sequence Probabilities** Here the tagger calculates the probability of a given sequence of tags occurring. It is also called the n-gram approach. It is called so because the best tag for a given word is determined by the probability at which it occurs with the n previous tags.