Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

NLP MSE-1 solution

Q1(a) Define affixes & its list its types

Ans Affixes are a grammatical element that is considered with a word, stem or phrase to produce derived or inflectional form

Types of affixes
- prefix
- infix
- suffixes

Eg unwillingness

Here there are three morphems (affixes)

un - prefix

willing - stem word

ness - suffix

Q1(b) What is lexical ambiguity in natural language processing. Give one example.

Ans Lexical ambiguity occurs when a word carries a different sense ie having more than one meaning & sentence in which it is combined can be interpreted differently depending on its correct sense.

Lexical ambiguity can be resolved using part of speech tagging technique

Eg - The chicken is ready to eat.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Q1(c)	List the application of NLP
	Ans	1) Text summarization 2) Recommendation engine 3) Opinion Mining 4) Sentiment Analysis 5) QuillBot - Paraphrasing tool 6) Chat Bot / Chat GPT 7) Alexa / Siri
	Q1(d)	For a corpus, the maximum likelihood estimate (MLE) for bigram "Battery life" is 0.27. frequency of "Battery" is 800. After applying laplace smoothing the MLE for battery life 0.025. What is the vocabulary size of the corpus.
	Ans	$P_{MLE}(life/Battery) = \frac{f(Battery\ life)}{f(Battery)}$ $0.27 = \frac{f(battery\ life)}{800}$ $f(Battery\ life) = 0.27 \times 800$ $= 216$ with laplace smoothing $P_{MLE}(life/Battery) = \frac{f(Battery\ life)+1}{f(Battery)+V}$ $0.025 = \frac{216+1}{800+V}$ $V = \frac{197}{0.025}$ $\boxed{V = 7880}$

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	8(c)	Differentiat between Stemming & Lemmitization with example
	Ans	1) Stemming is an elementary rule based process for removing inflectional forms from token & output are the stem/root of the word. Eg - Laughing, Laughed, Laughs will become Laugh ② Changing → chang
		2) Lemmatization is a systematic step by step process for removing inflectional form of a word. It makes use of vocabulary, word structure, part of speech tags, and grammar relation. Output of Lemmatization is the root word called a lemma. Eg - ① Running, Run, Ran → Run. ② Changing → Change.
	Q1(f)	Differentiat between Inflectional & Derivational Morphology with example

Ans	Inflectional Morphology	Derivational Morphology
①	The study of the modification of words to fit into different grammatical contexts.	The study of the formation of new word that differ either in syntactic category or in meaning from their base.
②	Create new forms of same word.	Creates new word
③	Run - Running VB VB	Danger → Dangerous Noun Adj. (JJ)

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

Q1(g) Apply Porter's Algorithm —

(1) Running → Run
(vowel)ing → ϕ

Rule used
(*V*)ing → ε
motoring → motor
sing → sing

(2) King → King

Q1(h) Define Regular Expression & its usage

Ans A regular expression is a language for specifying text search string. RE helps us to match strings or set of strings, using a specialized syntax held in a pattern. Usage in NLP — A RE search corpus function will search through the corpus returning all texts that contain the pattern.

Q2(a) Explain the stages of NLP.

Ans Stages of NLP —

(1) Morphological Analysis — In this phase individual words are analysed and distinct words are identified depending on the morphemes. Stemming & lemmitization are used as part of text preprocessing to identify the tokens (words)

(2) Syntatic Analysis — In this phase linear sequence of words are transformed into structure

Total Marks of Question no.		Examiner 1	

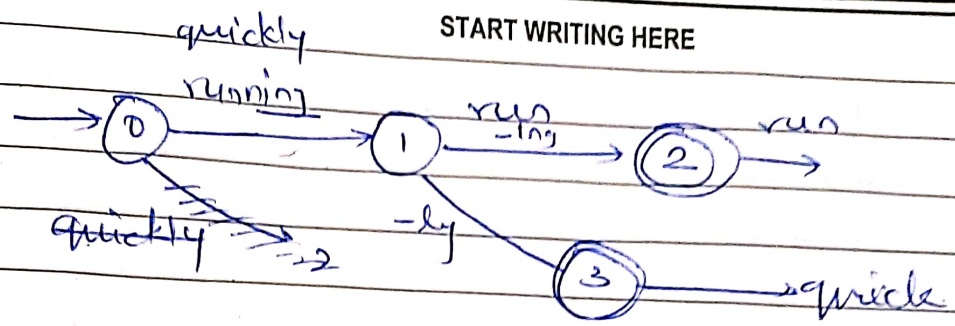
Space for Marks	Question No.	START WRITING HERE
		that shows how the words relate to each other. In this phase Part of speech tagging takes place.
		3) Semantic Analysis - In this phase a transf... is made from the input text to an internal representation that reflects the meaning. Here word sense disambiguation using wordnet / babelnet is performed.
		4) Pragmatic Analysis - To reinterpret what was said to what was actually meant.
		5) Discourse Analysis - Using Hobb's Algorithm resolves references.
	Q2(b)	Explain FST in detail with example.
	Ans	Finite State Transducer (FST) is a sophisticated tool that is used for understanding & transforming language.
		# FST is used in various applications like auto correcting, search engine. It is computational model used for representing & manipulating finite state machine (FSM) that map input sequence to output sequences.
		- Key component of FST are - state, transition input symbol, output symbol (annotation), & FSM

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
		- One common application of FST in NLP is Morphological analysis (which involves analyzing the structure & meaning of words at the morphem level)
		- Stemming with FST - Reducing words to their root form often by removing the affixes.
		- Eg - English stemming with an FST - Lets consider an English stemming. FST will have states representing the process of removing common english suffixes.
		Step 1 - Define the FST states & transitions states by defining the states of the FST representing different stages of stemming
		- Define transitions between states based on rules for removing suffixes
		- Eg transition - State 0 - Initial state Transition - If the the input ends with ing remove -ing & transition to state 1
		- State 1 : "ing" suffix removed Transition: if the input ends with ly remove "ly" & transition to state 2
		- State 2 : "ly" suffix removed Final state : Output the stemmed word.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE



Q2(c) Write a note on Language model

Ans: A language model is the core model component of NLP. It's a statistical model that is designed to analyze the pattern of human language & predict the likelihood of a sequence of words or tokens.

- Language model determine the probability of the next word by analysing the text in data. These models interpret the data by feeding it through algorithms.

- For training a language model, a number of probabilistic approaches are used. These approaches vary on the basis of the purpose for which a language model is created.

- Types of language model—

(1) Statistical Language model —
It includes the development of probabilistic models that are able to predict the next word in the sequence, given the word that preceds it.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	a)	N-Gram Model – In this a probability distribution for a sequence of n is created, where 'n' can be considered as size of gram. (a sequence of words). So if we use N-gram model, so it predicts the next words in the sequence based on N-1 words.
	b)	Neural Language Model – It overcomes of N-gram model & are used for complex task such as speech recognition or machine translation.
		Common Challenges of NLP Language Model
	(1)	Long term dependency – Capturing relationships between words that are far apart in a sentence. Traditional RNN faces vanishing gradient problems.
	(2)	Low Resource Language – Building effective model for indigenous language is challenging due to limited data.
	(3)	Difficult to handle sarcasm, irony, idioms.
	(4)	Handling noisy text
	(5)	Contextual Ambiguity

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Qs(A)	Find the probability of the following sentence <s> Michael & Zack played at the playground</s> from the following corpus. Assume a trigram language model (w/o laplace smoothing)
		<s> The school was open </s>
		<s> Michael & Zack went to the school </s>
		<s> The playground at the school was huge </s>
		<s> Bob, Michael & Zack were friends </s>
		Also find the perplexity of the sentence which has highest probability.
	Soln	Trigram Model Probability
		P(Michael | <s>) = 1/5
		P(& | <s> Michael) = 1/1
		P(Zack | Michael &) = 2/2
		P(Played | and Zack) = 1/3
		P(at | Zack Played) = 1/1

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

$P(\text{the} \mid \text{played at}) = 1/1$

$P(\text{playground} \mid \text{at the}) = 1/2$

$P(\langle /s \rangle \mid \text{the playground}) = 1/2$

$\therefore P(S) = \frac{1}{5} \times \frac{1}{1} \times \frac{1}{2} \times \frac{1}{3} \times 1 \times 1 \times \frac{1}{2} \times \frac{1}{2}$

$$\boxed{P(S) = \frac{1}{60}}$$

$\text{Perplexity} = P(S)^{-1/n}$ $\quad n = \text{Token} = 9$

$= \left(\frac{1}{60}\right)^{-1/9}$

$$\boxed{\text{Perplexity} = 1.576}$$

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Q3(4)	Discuss different Text Preprocessing techniques with example.
	Ans	Text preprocessing refers to a series of techniques used to clean, transform & prepare raw textual data into a format that is suitable for NLP or ML task. The goal of text preprocessing is to enhance the quality & usability of the text data for subsequent analysis or modelling
		- It involves following steps –
		1) Lowercasing 2) Removing punctuation & special character 3) Stop word removal 4) Removal of URLs 5) Removal of HTML tags 6) Stemming & Lemmatization 7) Tokenization 8) Text normalization
		① Lowercasing : text.lower()
		② Removing punctuation & special character using NLTK & RE library re.sub(punctuation_pattern, '', text)
		③ Stop word removal can be done by importing stopword from nltk.corpus import stopwords remove_stopword(en_text, "english")
		④ Removal of URLs using Regular Expression RE re.compile(r'https?://\S+|www\.\S+')

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	(5)	Stemming & Lemmatization — Stemming removes common suffixes from the end of word tokens, lemmatization ensures the output word is as existing normalized form of the word. Running — Run Smiling — Smile
	(6)	Tokenization — It is the process of dividing the text into smaller units known as tokens. The word tokenization divides the text into individual words. Sentence Tokenization: This is useful for tasks requiring individual sentence analysis or processing.