

**CSE (AI & ML)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Course code:** | **MR20-1CS0249** | **Course Name** | **NATURAL LANGUAGE PROCESSING** |

**QUESTION BANK**

|  |  |  |  |
| --- | --- | --- | --- |
| **Qno** | **Question** | **Marks** | **Section** |
| 1 | What word tokenization? What are the major challenges to tokenize words of a sentence | 12 | Section-I |
| 2 | What is minimum edit distance between two words? Calculate minimum edit distance between two words: “small” and “smell” using dynamic programming algorithm. Consider cost for insertion, deletion and substitution is 1,1,2. | 12 | Section-I |
| 3 | What is the difference between non-word and real-word spelling correction? What is perplexity? Estimate the perplexity of the corpus based on unigram language model: “**the man is a thief but the man is a good man**” | 12 | Section-I |
| 4 | What is Maximum Likelihood Estimate? How is it used Language Model? Given the Corpus    Calculate the following:   1. Find all the possible bigrams from the given corpus. 2. Find frequencies of all the bigrams. 3. Find the frequencies of all unigram. 4. Calculate the Maximum Likelihood Estimate for all bigrams. | 12 | Section-I |
| 5 | What is morphology in NLP? What is morphemes? What is bounded and free morphemes explain with example. What is stemming and how it is different from lemmatization? | 12 | Section-I |
| 6 | Write about Evaluation of Language Models and Basic Smoothing | 12 | Section-I |
| 7 | Explain about Noisy Channel Model for Spelling Correction-Gram Language Models | 12 | Section-I |
| 8 | Develop a comprehensive text processing pipeline that includes tokenization, stemming, normalization, and spelling correction. | 12 | Section-I |
| 9 | Apply different smoothing techniques to a language model and analyze their impact on performance. | 12 | Section-I |
| 10 | Design an algorithm to correct spelling errors in a given text document. | 12 | Section-I |
| 11 | What is morphology in NLP? What is morphemes? What is bounded and free morphemes explain with example. What is stemming and how it is different from lemmatization? | 12 | Section-II |
| 12 | What is the difference between inflectional and derivational morphology explains with example. What is morphological analysis explain with example. |  |  |
| 13 | What is POS tagging? Find the POS tag for the phrase “the light book”, using Viterbi algorithm in Hidden markov tagging model with the following information. | 12 | Section-II |
| 14 | What is difference between real words and non words? What is FSA and how inflections in words can be represented using FSA, explain with example. | 12 | Section-II |
| 15 | What are the problems of Hidden Markov model to predict the POS tags for a given sentence or phrase? Explain how Baum Welch algorithm learns the parameters – transition matrix, observation matrix and initial state distribution. | 12 | Section-II |
| 16 | What is smoothing in language model? What are the advantages smoothing? Find the Good turing smoothing for the following sentence:  “he is he is good man” | 12 | Section-II |
| 17 | Explain the different categories of affixes in morphology with examples. What are the differences between content and functional morphemes? What is the difference between regular and irregular forms of verbs and nouns respectively? | 12 | Section-II |
| 18 | Established why maximum entropy model is better than hidden Markov model. How POS tagging is achieved in maximum entropy model. What is beam search explain in detail. | 12 | Section-II |
| 19 | How is the uniformity maintained in maximum entropy model? Write the maximum entropy model principles. | 12 | Section-II |
| 20 | Consider the maximum entropy model for POS tagging, where you want to estimate *P*(*tag*|*word*). In a hypothetical setting, assume that *tag* can take the values *D*, *N* and *V* (short forms for Determiner, Noun and Verb). The variable *word* could be any member of a set *V* of possible words, where *V* contains the words *a*, *man*, *sleeps*, as well as additional words. The distribution should give the following probabilities  *P*(*D*|*a*) = 0.9  *P*(*N*|*man*) = 0.9 *P*(*V*|*sleeps*) = 0.9  *P*(*D*|*word*) = 0.6 for any word other than *a*, *man* or *sleeps P*(*N*|*word*) = 0.3 for any word other than *a*, *man* or *sleeps P*(*V*|*word*) = 0.1 for any word other than *a*, *man* or *sleeps*  It is assumed that all other probabilities, not defined above could take any values such that  *ΣP*(*tag*|*word*) = 1 is satisfied for any word in *V*.   1. Define the features of your maximum entropy model that can model this distribution. Mark your features as *f*1, *f*2 and so on. Each feature should have the same format as explained in the class 2. For each feature *fi*, assume a weight λ*i*. Now, write expression for the following probabilities in terms of your model parameters   *P*(*D*|*cat*)  *P*(*N*|*laughs*)  *P*(*D*|*man*)   1. What value do the parameters in your model take to give the distribution as described above. (i.e.*P*(*D*|*a*) = 0.9) and so on. | 12 | Section-II |
| 21 | What is syntax? What is parsing? What is the difference between derivation and parse tree? What is constituency? Write down different forms of constituency with example. What is the significance of “head” of a constituency, explain? | 12 | Section-III |
| 22 | What is the difference between top down and bottom up parsing? Apply CYK algorithm to parse the sentence “a pilot likes flying planes” with given grammar | 12 | Section-III |
| 23 | What is inside-outside probability? Apply CYK algorithm to parse the sentence “a pilot likes flying planes” with given probabilistic context free grammar to find most probable sparse tree. | 12 | Section-III |
| 24 | What is dependency parsing? What is difference between classical and dependency parsing? Exaplain the dependency structure in dependency parsing with suitable example. What is head and dependent and what are the criteria are set for them? | 12 | Section-III |
| 25 | What is dependency graph? What are the main characteristics of dependency graph? What is configuration in transition based dependency parsing and what is the initial value for configuration. Parse the following sentence with Arc-Eager algorithm. | 12 | Section-III |
| 26 | For the given grammar    Find the inside probabilities for each word for the following sentence;  “Astronomers saw stars with ears” | 12 | Section-III |
| 27 | Evaluate the effectiveness of the CKY algorithm in various syntax parsing tasks. | 12 | Section-III |
| 28 | Describe the inside-outside algorithm for calculating probabilities over parse trees. | 12 | Section-III |
| 29 | Explain how PCFGs assign probabilities to different parse trees for a given sentence. | 12 | Section-III |
| 30 | Discuss the evaluation of transition-based parsers using different metrics. | 12 | Section-III |
| 31 | What do you mean by distributional semantics? What is contextual representation and how we can we learn new words from contextual cues? Explain with examples. What do mean by Distributional Semantic Models(DSMs)? | 12 | Section-IV |
| 32 | What is word space? Write down the steps to create words space and explain it with example and how can it be useful to show the word similarities. | 12 | Section-IV |
| 33 | How weights can be measured based on context? Deduce the formulation for weight measurements. What is difference between attributional and relational similarity? | 12 | Section-IV |
| 34 | What one-hot encoding? How words can be represented using one-hot encoding explain with example? What are the limitations of one-hot encoding explaining with example? | 12 | Section-IV |
| 35 | What is CBOW? How CBOW is used to emebed word explain with example. What is the difference between skip-gram and CBOW? | 12 | Section-IV |
| 36 | Discuss the advantages and limitations of distributional semantic models compared to other approaches. | 12 | Section-IV |
| 37 | Discuss the application of distributional semantic models in sentiment analysis and topic modeling. | 12 | Section-IV |
| 38 | Discuss the different types of word embedding techniques, including word2vec, GloVe, and fastText. | 12 | Section-IV |
| 39 | Describe the application of word embeddings in various NLP tasks, including machine translation, sentiment analysis, and question answering. | 12 | Section-IV |
| 40 | Explain how WordNet is used for word sense disambiguation and lexical relation extraction. | 12 | Section-IV |
| 41 | What is summary? What is text summarization? What are the applications of text summarization give examples? | 12 | Section-V |
| 42 | What are the main stages of text summarization? How salient words can be defined? How sentence can be weighted? | 12 | Section-V |
| 43 | How sentences can be simplified, Explain with example. How summarization systems can be evaluated? What is ROUGE and how is it used for system evaluation | 12 | Section-V |
| 44 | What is text classification? What kind of problems can be solved using text classification? How text classification problems can be solved? | 12 | Section-V |
| 45 | Discuss the different types of text classification tasks, including binary, multi-class, and hierarchical classification. | 12 | Section-V |
| 46 | Discuss the evaluation of text classifiers using metrics like accuracy, precision, recall, and F1-score. | 12 | Section-V |
| 47 | Describe the application of sentiment analysis in social media analysis, product reviews, and customer feedback. | 12 | Section-V |
| 48 | Discuss the challenges of sentiment analysis, including handling sarcasm, irony, and ambiguity. | 12 | Section-V |
| 49 | Describe the application of machine learning algorithms like Naive Bayes, support vector machines (SVMs), and random forests in text classification. | 12 | Section-V |
| 50 | Describe the application of optimization algorithms like integer linear programming (ILP) and genetic algorithms in text summarization. | 12 | Section-V |