Name: ID:

Birla Institute of Technology & Science (BITS), Pilani
2nd SEMESTER 2022-23,
Natural Language Processing for Business MPBA G519
Mid-Semester Examination (Closed Book) – Part A

Max. Time: 30 Minutes Date: 14-03-2023 Max. Marks: 45

Instructions:

I. Every wrong answer will result in deduction of 0.5 marks due to negative marking.
II. Please write your answer in the response space as provided in the second page of the question paper. The answers should be neatly written and free of any over writing/cutting. Otherwise it will not be evaluated.
III. A question may have more than one correct response

---

1. Is Phonetics analysis a part of NLP?

A. True B. False

2. What is Morphological analysis?

A. focuses on how the components within a word (stems, root words, prefixes, suffixes, etc.) are arranged or modified to create different meanings
B. is to draw exact meaning, or you can say dictionary meaning from the text.
C. the main focus always on what was said is reinterpreted on what is intended.
D. deals with how the sounds are produced when we talk and how words are related to sounds.

3. How is syntactic analysis different from semantic analysis?

A. Syntactic analysis deals with syntax and semantic analysis deals with words and phrases to determine relationship b/w independent terms
B. Semantic analysis deals with syntax and structure of language, syntactic analysis deals with relationship b/w words and phrases
C. Semantic analysis deals with sequential arrangement of words and syntactic analysis deals with grammatic dependencies to form correct structure.

4. Which modelling approach is best suited for less data and more complexity in information?

A. Rule-based B. ML-based
C. DL-based D. Statistics-based

5. Which modelling approach is best suited for huge training data and less complexity in information?

A. Rule-based B. ML-based
C. DL-based D. Statistics-based

6. Which modelling approach is best suited for huge training and high complexity in information?

A. Rule-based B. ML-based
C. DL-based D. Statistics-based

7. In which scenarios rule-based methods of NLP works best?

A. Less complexity and fixed environmental conditions for modelling
B. Frequency based analysis for modelling purposes.
C. Less complexity and semantic information utilization
D. Initial step for data filtering

8. In which scenario ML models work best?

A. Less data, high complexity
B. Huge data, less complexity
C. Huge data, huge complexity
D. Less data, less complexity

9. In which scenario you should consider transfer learning for model development?

A. less data, high complexity
B. huge data, less complexity
C. huge data, high complexity
D. less data, less complexity

10. In which scenario you should consider deep learning LSTM model for model development?

A. Contextual data learning
B. sequential information learning
C. relationship learning from data.
D. new feature learning from data

11. What are the different levels of morphology information provided by linguistics?

A. Verbs. Noun. Determiners
B. Noun phrase, Prepositional phrases, Verb phrases
C. inflection, derivation, compounding
D. Open language, lexical language, Closed language

12. What is the difference between stemming and lemmatization?

A. stemming is trimming the words with an algorithm, lemmatization is resetting to root word.
B. stemming is resetting to root words, lemmatization is

trimming the words with an algorithm

13. What is the difference between co-referencing and correlation?

A. Coreference resolution is the task of finding all expressions that refer to the same entity in a text
B. Correlation is used to find the relationship between two variables which is important in real life
C. Coreference is used to find the relationship between two variables which is important in real life
D. Correlation resolution is the task of finding all expressions that refer to the same entity in a text

14. What is the difference between co-location similarity and positional similarity?

A. Co-location: Relative position of different entities. Positional: relative position of an entity in the contextual data
B. Positional: Relative position of different entities. Co-location: relative position of an entity in the contextual data
C. Colocation: used to create Associated n-gram frequency. Positional: used to create normalized features in data
D. Positional: used to create Associated n-gram frequency. Colocation: used to create normalized features in data

15. What is context free grammar method of modelling?

A. CFG is a list of rules that define the set of all well-formed sentences in a language.
B. CFG uses POS tags to develop rules
C. CFG uses Bag of words approach to develop model
D. CFG is a list of expressions that defines set of all well-formed words in sentences in a language

16. What is the difference between feature engineering and feature extraction?

A. feature engineering is SME based feature designing; feature extraction is a mathematical tool based information transformation
B. Convolution is DL is feature engineering.
C. Convolution is DL is feature extraction.
D. feature extraction is SME based feature designing; feature engineering is a mathematical tool-based engineering approach

17. What is the advantage of using Convolution is the deep learning models?

A. new feature engineering for enhanced dimensional space
B. new feature extraction for enhanced dimensional space
C. mathematically configured features to enhance deeper learning of patterns
D. learning the latent space based patterns

18. What is the advantage of using Auto encoders?

A. dimensionality enhancement
B. dimensionality reduction
C. generative learning
D. discriminative learning

19. What are the advantages of using word-to-vector embedding in deep learning?

A. contextual learning
B. global positioning of words in the embedding space
C. local positioning of words in the embedding space
D. non-contextual learning

20. Select correct comments regarding PCA?

A. non-linear dimensionality reduction
B. linear dimensionality reduction
C. components decide the information to decide the reduced space
D. components do not decide the information to decide the reduced space

21. Select correct comments regarding variational models?

A. generative in nature
B. deterministic in nature
C. discriminative in nature
D. capture the data distribution in relation to parametric statistical models

22. What is the difference between generative models and discriminative models?

A. Generative: learn the joint probability distribution. Discriminative: learn the conditional probability distribution
B. Discriminative: learn the conditional probability distribution. Generative: learn the joint probability distribution
C. Generative: capability to generate new information from the same space. Discriminative: cannot generate new data
D. Discriminative: capability to generate new information from the same space. Generative: cannot generate new data

23. What is the difference between generative models and deterministic models?

A. deterministic model are fixed in nature and produce same results at every inference step with same input
B. generative model are variational in nature and produce different data at every inference step with same input
C. deterministic model are variational in nature and produce different data at every inference step with same input
D. generative model are fixed in nature and produce same results at every inference step with same input

24. What is the difference between deterministic and discriminative models?

A. discriminative models have probabilistic nature and learn the conditional probability distribution
B. deterministic model are fixed in nature and produce same results at every inference step with same input
C. discriminative models learns the boundaries between classes or labels in a dataset
D. deterministic model learn the boundaries between classes or labels in a dataset

25. Select the use cases where you can select dictionary-based NLP model approach?
A. Clustering B. sentiment analysis
C. machine translation D. language generation

26. Which are all the correct comments regarding Latent Dirichlet Allocation algorithm?
A. Unsupervised in nature
B. supervised in nature
C. semi-supervised in nature
D. self-supervised in nature

27. Which are the correct comments regarding Latent Semantic Analysis?
A. Unsupervised in nature
B. supervised in nature
C. semi-supervised in nature
D. self-supervised in nature

28. What are the correct methods for capture semantic relationships?
A. CBOW/ Skip-gram B. N-grams
C. colocation and positional similarity
D.TF/IDF

29. What are the existing libraries for grammar analysis for sentences?
A. NLTK B. StanfordCoreNLP
C. Pandas D. Numpy

30. Which processes are considered as part of data cleaning?
A. stop word removal, alpha-numerical data removal, relevant information filtering
B. tokenization, stemming, lemmatization.
C. data formatting, data conversion, image to text extraction
D. TDM, TF-IDF, N-gram

31. Which methods are considered as part of statistical NLP approach for modelling?
A. Mean, median, mode analysis
B. frequency of co-location and positional frequency analysis
C. TDM, TF-IDF, N-gram
D. context free grammar modelling

32. What is conjugate prior relationship?
A. Likelihood and Prior belong to the same family of distribution.
B. Posterior and likelihood belong to same family of distribution.
C. Prior and Posterior belong to belong to same family of distribution.
D. Prior, Posterior and likelihood belong to same family of distribution.

33. What is the correct order of complexity handing in terms of contextual learning?
A. embedding representation | deep neural nets | optimization of loss | testing.
B. embedding representation | optimization loss | deep neural nets | optimization of loss | testing.
C. embedding representation + deep neural nets | optimization of loss | testing.
D. embedding representation | optimization of loss | testing | deep neural nets.

34. What is hyperparameter tuning in machine learning?
A. update the external parameters of a ML algorithm to improve the performance.
B. cross validation is a part of hyperparameter tuning.
C. hyperparameters are the same for ML algorithms.
D. hyperparameter tuning suggest different methods and approaches for the same base method

35. What is the correct mapping of algorithm in terms of operations?
A. SVM: coordinate based algorithm | random forest: tree-based algorithm.
B. SVM: tree-based algorithm | random forest – coordinate based approach.
C. DL: contour based algorithm | KNN : Coordinate-based approach
D. SVM: binary linear classification

36. What is difference between self-supervised and unsupervised learning?
A. Self-supervised: same data is used as Input and Output
B. unsupervised: internal properties of the data are explored to identify similarity in data points
C. Self-supervised: used mostly to learn joint probability distribution of data
D. unsupervised: used to understand the border conditions of different classes in the data

37. Select the correct mapping of the algorithms to corresponding setup?
A. SVM: Supervised learning | random forest: Supervised learning
B. SVM: unsupervised learning | random forest – unsupervised learning
C. DL: supervised, self-supervised | KNN : supervised learning
D. SVM: binary linear classification

38. What is the requirement of normalization in model development process?
A. to reduce the big number effect
B. to scale the entire information space
C. to bound the range in which model gets trained
D. to keep the data distribution consistent while neglecting big number effect

39. What are the drawbacks of noisy data, select correct options?
A. bias in model
B. poor test performance on model

C. un-wanted introduced complexity in data
D. high variance in model

40. What are the advantages of using POS tagging in solution development?
A. patterns rule generation.
B. words agnostic modelling
C. frequency-based modelling.
D. context free grammar modelling

41. What are the differences between CBOW and Skip-gram approaches?
A. CBOW learn the single word context with regard to associated words.
B. Skip-gram uses single word to learn the context of associated words.
C. CBOW uses single word to learn the context of associated words.
D. Skip-gram learn the single word context with regard to associated words

42. What is the difference between AE and VAE?
A. AE is used for dimensionality reduction and VAE is used for data generation
B. AE do not use parametric pdf function while training and VAE uses parametric pdf function while training
C. VAE is used for dimensionality reduction and AE is used for data generation
D. VAE do not use parametric pdf function while training and AE uses parametric pdf function while training

43. What is correct mapping of loss functions with problem type in Deep learning?
A. RMSE -> Regression modelling
B. categorical cross entropy -> classification modelling
C. RMSE -> Autoencoder modelling
D. categorical cross entropy -> regression modelling

44. What are the best use cases for Generative models?
A. Machine translation
B. Natural language generation
C. Summery generation
D. Sentiment classification

45. Select the correct statements for word-to-vector approach?
A. tokenization -> one-hot encoding -> CBOW/Skip-gram -> weights to get vectors.
B. CBOW/Skip-gram -> weights to get vectors -> tokenization -> one-hot encoding.
C. one-hot encoding -> tokenization -> CBOW/Skip-gram -> weights to get vectors.
D. tokenization -> CBOW/Skip-gram -> one-hot encoding -> weights to get vectors.

RESPONSE SPACE

Question No.	Answer	Question No.	Answer
1		24	
2		25	
3		26	
4		27	
5		28	
6		29	
7		30	
8		31	
9		32	
10		33	
11		34	
12		35	
13		36	
14		37	
15		38	
16		39	
17		40	
18		41	
19		42	
20		43	
21		44	
22		45	
23			

Birla Institute of Technology & Science (BITS), Pilani
2nd SEMESTER 2022-23,
Natural Language Processing for Business MPBA G519
Mid-Semester Examination (Closed Book) – Part B

Max. Time: 60 Minutes | Date: 14-03-2023 | Max. Marks: 55

---

1. PROBLEM SOLVING SKILLS [30]

Problem statement:

OLIVER GREEN Pvt. Ltd. Is a FMCG company which has lot of farm-based products for daily use. The company wants to identify the exact improvement points of its products from the feedback shared by customers as well as what exactly customers like about the products. So, the company decided to collect the data from different e-commerce web sites where its products are available to the customers to purchase. Following are the websites:

1. Amazon
2. Walmart
3. D-mart

Other specification:

1. The total number of different products available are approximately 50.
2. Price range of the products range from 100/- to 5000/-
3. Sale of the cheaper products is more in comparison to costly products.
4. Feedback is provided in a limit of 250 letter.
5. Feedbacks have spelling mistakes.
6. There are utilization of special characters and emojis in the text

e-commerce platform capabilities:

1. They provide APIs for data collection.
2. Real time feedbacks are available

Requirement:

1. Develop a mechanism/ architecture to extract information about the products:
    a. Which products are doing good in market.
    b. Which products need improvement.
        i. What specific improvements are targeted?
    c. What platform is good in which product.
    d. Generate recommendations based on the specifications of different products
        i. Cross sell
            1. Definition: recommend different products based on their selection pattern of users
        ii. Up-sell
            1. Definition: recommend upper versions of the same product type based on the user purchase patterns

Python related questions:

1. Please provide a pseudo code for the approach in terms of functions and classes
2. List down different python packages/ libraries needed for the solution development with reasoning

2. BUSINESS USE CASE SOLUTION APPROACH [25]

Max Profit Pvt. Ltd. Is a social media marketing and advertising company which helps its customers to build advertisements, run campaigns, manage social media handles etc. to boost sales. The company wants to introduce NLP capabilities using AI to built robust models to help its customers.

Problem Statement:
Please recommend three different use cases related to NLP which company can build to help improve performance and sales of customers.
In each use case, following details are required:

1. Data source
2. Data type
3. Complexities in data e.g., cleaning requirements, stop word removal requirements etc.
4. Solution approach architecture diagram
5. Reason behind selection of different components in the solution architecture
6. Output the user gets from the model.
7. Performance metrics to measure the quality of output.