Name: ID:

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

Max. Time: 30 Minutes Date: 09-05-2023 Max. Marks: 30

Instructions:

I. Every wrong answer will result in a deduction of 0.5 marks due to a negative marking.
II. Please write your answer in the response space as provided on the fourth page of the question paper. The answers should be neatly written and free of any overwriting/cutting. Otherwise, it will not be evaluated.

---

1. Which model is generative:
   a. SVM
   b. VAE
   c. RF
   d. KNN

2. ML ops is necessary for:
   a. Continuous integration – Continuous development
   b. Continuous training - Continuous testing
   c. Continuous analysis – Continuous feature selection
   d. Continuous processing - Continuous modelling

3. Definition of large language models:
   a. Large Language Models (LLMs) are a type of artificial intelligence that's been trained on a massive corpus of text data to produce human-like responses to natural language inputs.
   b. Large Language Models (LLMs) are types of deep learning models having a parameter count on the order of millions.

4. Which model can be used for dimensionality reduction:
   a. LSTM
   b. CNN
   c. GAN
   d. AE

5. What sentence defines EPOCHs:
   a. Number of data points passing through the model
   b. Number of times individual data point passing through the model
   c. Batches in which the data passes through the model.
   d. Epoch = data size / batch size

6. What is the meaning of embedding in NLP in deep learning:
   a. Contextual representation of words/ sentences in form of vectors where semantic similarities are represented by closeness in vectors.
   b. Cosine similarity of words/ sentences in form of vectors where lemmas and stemming based similarity are represented by closeness in vectors.

7. Which is correct method of embedding for huge textual data:
   a. CBOW
   b. SKIP-GRAM
   c. BOW
   d. C-GRAM

8. What is the definition of diameter in a graph:
   a. Average distance between the nodes
   b. Distance between the 2 extreme nodes in the graph
   c. Longest distance between any 2 nodes
   d. Shortest distance between 2 extreme nodes

9. Which one is the property of a graph:
    a. Betweenness
    b. Centroid
    c. Eigen values
    d. Coupling

10. What is padding in pre-processing:
    a. Adding null values to the start of each sentence
    b. Adding a constant value to the end of each sentence to make them all equal.
    c. adding null values to start and end of each sentence to notify the change in sentence.
    d. adding constant values to the start and end of each sentence

11. properties of convolution operation in deep learning:
    a. enhance features.
    b. extract features
    c. engineer features
    d. select features.

12. in NLP transformers can be used for:
    a. Sentiment analysis
    b. Machine translation
    c. Next work prediction
    d. All of the above

13. LSTMs are used for:
    a. Learning word to vector embedding
    b. Learning sentence to vector embedding
    c. Learning sequence mapping
    d. Learning cosine similarity of words

14. CNN can be used in NLP for:
    a. Extract features
    b. Classification
    c. Sequence learning

    Options:
    1. a, b
    2. a, c
    3. b, c
    4. none

15. GANs in NLP can be best used for:
    a. Sentiment analysis
    b. Sentence completion
    c. Topic modelling
    d. Aspect detection

16. Beautiful Soap package in python is used for:
    a. Extraction from web pages
    b. Extraction from pdf.
    c. Extraction from word document
    d. Extraction from excel workbook.

17. What is name entity recognition:
    a. Detection of noun
    b. Detection of verbs
    c. Detection of adverbs
    d. Detection of conjunctions

18. Which python package is used for classical ml modelling:
    a. Sklearn
    b. Pandas
    c. Numpy
    d. Tensorflow

19. Which python package perform NER better:
    a. NLTK
    b. Spacy
    c. Stanford NLP
    d. Steamlit

20. Which step is a part of data cleaning:
    a. Stop word removal.
    b. Lemmatization
    c. Stemming
    d. TF=IDF

21. Which of the following can be considered as stop word:
    a. And
    b. That
    c. Agra
    d. The

22. N-grams are useful in:
    a. Association mining
    b. Context capturing
    c. Features generation
    d. Embedding

Options:

1. A, b
2. B, c
3. C. d
4. A, d

23. Which methods is suitable for topic modelling in smaller dataset:
    a. Latent Dirichlet Allocation
    b. Latent semantic Analysis
    c. K-means clustering
    d. Hierarchical clustering

24. How to determine the need for re-training?
    a. Monitoring the data drift
    b. Monitoring the concept drift
    c. Monitoring the feature drift
    d. Monitoring the model accuracy in test environments
    e. All of the above

25. Which type of algorithms are most suited for recommender system development:
    a. Deterministic
    b. Generative
    c. Discriminative
    d. Association mining models

26. Graph based modelling is suited for which scenarios:
    a. Sentiment analysis
    b. Document classification
    c. Recommender system
    d. Knowledge based development.

    Option:

    1. A, b
    2. B, c
    3. C, d
    4. A, d

27. Which all algorithms are self – supervised:
    a. VAE
    b. GAN
    c. Transformers
    d. CNN
    e. LSTM

    Options:

    1. A, B, C
    2. B, C, D
    3. D, E, A
    4. A, B, D

28. which are the latest large generative AI models:
    a. GPT
    b. Dalle
    c. BERT
    d. GAN

    Options:

    1. A, B
    2. B, C
    3. A, B, C
    4. D, A

29. Which all are the correct statements:
    a. Chatbots try to mimic human-like responses.
    b. Chatbots can be specific to domains/ tasks.
    c. Chatbots can be generic and can respond to any query.
    d. Chatbots do not hold contextual multi-turn capabilities.

    Option:

    1. A, b
    2. B, c
    3. C, d
    4. D, a

30. Select correct properties regarding classical ml modeling:
    a. Garbage in – garbage out
    b. No free lunch
    c. Curse of dimensionality
    d. Occam razor

    Options:

    1. A, B
    2. B, C
    3. A, B, C, D
    4. A, B, C

RESPONSE SPACE

Question No.	Answer	Question No.	Answer	Question No.	Answer
1		11		21	
2		12		22	
3		13		23	
4		14		24	
5		15		25	
6		16		26	
7		17		27	
8		18		28	
9		19		29	
10		20		30	

Response	Number	Weight	Total
Correct		+ 1	
Wrong		-0.5	
Total			

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

Max. Time: 150 Minutes Date: 09-05-2023 Max. Marks: 90

Section1: Short Answers (Each question carries 2 marks)

1. Suppose you are doing bag-of-words text classification on a document. The raw input is a single string containing the text of the entire document. Describe in one or two sentences the pipeline to go from the raw input to a feature vector.

2. Suppose you have a neural network that is overfitting to the training data. Describe two ways to fix this situation.

3. You are training a neural network with Adam and watching the negative log likelihood of the training set over epochs. Rather than decreasing, it seems to fluctuate around where it started. What is one change you could make to your training procedure that could fix this?

4. Describe context free grammar-based modelling with suitable example.

5. Describe different types of feature engineering techniques in classical ML considering document classification use case.

6. Draw the general architectural diagram of NLP Classical ML modelling.

7. Describe in short, diverse types of chatbots and their design dependencies on LLMs.

8. Describe which method of word embedding is suitable for larger datasets and why?

9. Describe the training mechanism of self-supervised deep learning models.

10. Draw the general framework of MLops working mechanism for general NLP projects.

11. List different POS taggers we used in class, also design an algorithm for Auditor identification from financial reports.

12. Draw in steps the working mechanism of Latent Dirichlet Allocation based topic modelling.

13. List down 5 NLP use cases in health care industry.

14. List down 5 NLP use cases in logistic industry.

15. Mention with a brief description of all the stages of language modelling where NLP can be used?

Section 2: Case Studies

1. Oliver Green Pvt. Ltd. is a news channel company. It aims to provide the latest news as quickly as possible on its channel, for this the IT cell of Oliver Green Pvt. Ltd. keeps track on multiple social media channels and handles capture the news and develops reports accordingly.
   a. Could you propose a design architecture that IT cell would be following to collect relevant information and data from different social media channels. (10)
      i. The architecture should contain detailed flow of information collection from different sources.
      ii. The architecture should contain required cleaning steps.
      iii. The architecture should contain different block-based symbols to detect relevant news using alarms.
      iv. The architecture should contain different use cases you can draw from news data.
   b. Information captured from different social media sources needs to be classified into different categories. Please provide a solution design to news classification using deep learning model. (10)
      i. The design should contain detailed cleaning steps.
      ii. The design should contain details about different categories in the news.
      iii. The design should contain proper features extraction/ engineering steps.
   c. Every news collected needs to be summarized and writing summary is a difficult manual task. Provide a design to develop summaries using GPT with correct Prompts and design architecture. (10)

2. Hotel Comfort is a 5-star Hotel. The Hotel is looking for strategies to boost performance using different social media platforms and marketing strategies. Please provide functional design for the following asks:
   a. The hotel wants to boost the restaurant's performance and wants the city local people to also visit. Please provide suggestions with solution design using NLP and different channels of marketing. (10)
   b. The hotel wants to boost room booking. Please provide suggestions with solution design using NLP and different channels of marketing. (10)
   c. The hotel wants to enhance employee satisfaction, and for this, it created a portal where employees can give feedback and suggestions. Please provide a design to collect and analyze these comments for the following reasons: (10)
      i. The overall sentiment of employees
      ii. The most important aspects represent positive and negative organizational operations.
      iii. Immediate actions are required.
      iv. Capture gender discrimination