Nepal College of Information Technology

Programme: BE | Year: 2025

NATURAL LANGUAGE PROCESSING

Chapter-wise Question Bank

CHAPTER 1: INTRODUCTION TO NLP (4 Hours)

Topics: What is NLP, NLU vs NLG, Turing Test, Human language and intelligence, Applications of NLP, Major challenges (ambiguity, context dependence, language diversity)

S.N.	Question	Marks
1.	Define Natural Language Processing (NLP). Explain the difference between Natural Language Understanding (NLU) and Natural Language Generation (NLG) with suitable examples.	8
2.	What is the Turing Test? Explain its relevance to Natural Language Processing and how it relates to the concept of human language and intelligence.	8
3.	Discuss the major challenges faced in Natural Language Processing. Explain ambiguity, context dependence, and language diversity with examples.	7
4.	Explain different types of ambiguity in NLP with examples: lexical ambiguity, syntactic ambiguity, and semantic ambiguity.	7
5.	What is language diversity and why is it a challenge for NLP systems? Discuss with reference to morphological complexity, word order, and resource scarcity.	7
6.	Compare and contrast Natural Language Understanding (NLU) and Natural Language Generation (NLG). Provide examples of systems that use each.	8
7.	What are the key components of an NLP system? Explain the NLP steps with an appropriate example.	7
8.	What is parsing? Create a syntactic tree for the given sentence. “I have selected NLP as an elective subject for this semester”.	7

CHAPTER 2: TEXT PREPROCESSING AND FEATURE ENGINEERING (8 Hours)

Topics: Data collection, Text cleaning, Tokenization, Stop word removal, Stemming, Lemmatization

S.N.	Question	Marks
1.	Describe text preprocessing pipeline in NLP. Apply each step to the following text: *"The running cats were quickly running towards the better hiding places!"* *"Dr. Smith's research paper on A.I. is fascinating! It explores N.L.P. applications."* *"Can you believe it?"*	8
2.	What is tokenization? Explain one hot encoding method with example and its major drawback	8
3.	Explain stemming and lemmatization. Compare their differences with examples and discuss when to use each technique.	8
4.	What is data cleaning in NLP? Explain the steps involved including lowercasing, punctuation removal, and whitespace handling with examples.	8
5.	What are stop words? Why are they removed in NLP preprocessing? List common stop words and explain scenarios where stop word removal might be harmful.	7

CHAPTER 3: TRADITIONAL TEXT REPRESENTATION (7 Hours)

Topics: Label encoding, One-hot encoding, Bag of Words, TF-IDF, BM25, Vector space models, Cosine similarity

S.N.	Question	Marks
1.	What is TF-IDF? Calculate the TF-IDF scores for each unique term in the following documents (use raw TF and log base 10 for IDF): *D1: "machine learning is a subset of artificial intelligence"* *D2: "deep learning is a subset of machine learning"* *D3: "artificial neural networks enable deep learning"*	5+10
2.	Explain the Bag of Words (BoW) model. What are its limitations? How does TF-IDF address some of these limitations?	7
3.	Explain the BM25 algorithm for improved term weighting. How does it improve upon basic TF-IDF?	8
4.	What is a Vector Space Model? Explain cosine similarity and calculate it for the following document vectors: *Document A: [2, 1, 3, 0, 1]* *Document B: [1, 2, 0, 1, 3]*	2+5
5.	Compare label encoding and one-hot encoding for text representation. When would you use each method?	7
6.	Why is cosine similarity preferred over Euclidean distance in the vector space model for text analysis? Explain with examples.	7
7.	What is Zipf's Law? How does it relate to term frequency in text documents? Explain with an example.	8
8.	Compare TF-IDF and BM25 scoring methods. Create a comparison table highlighting their differences.	8

CHAPTER 4: WORD EMBEDDINGS AND NEURAL METHODS (6 Hours)

Topics: Word2Vec (CBOW, Skip-gram), GloVe, FastText, Evaluating word embeddings

S.N.	Question	Marks
1.	Explain the Word2Vec model. Describe and compare the CBOW (Continuous Bag of Words) and Skip-gram architectures with diagrams.	8
2.	What is GloVe (Global Vectors for Word Representation)? How does it differ from Word2Vec in terms of training methodology and capturing word relationships?	7
3.	Explain FastText and how it handles subword information. What advantages does it have over Word2Vec for handling out-of-vocabulary words?	8
4.	How do word embeddings capture semantic relationships? Explain with the classic example of 'king - man + woman = queen'.	7
5.	How does CBOW architecture work? Explain the network structure, input representation, and training process with a diagram.	8
6.	What are the limitations of traditional word embeddings like Word2Vec? How do contextual embeddings (like BERT) address these limitations?	8
7.	What is n-gram model? Explain different n-grams model with example	7

CHAPTER 5: TEXT CLASSIFICATION AND LANGUAGE MODELS (5 Hours)

Topics: Naive Bayes classifier, Laplace smoothing, N-gram language models, Neural language models, Evaluation metrics

S.N.	Question	Marks
1.	What is a language model? Explain unigram, bigram, and trigram language models with examples.	7
2.	Explain the Naive Bayes classifier with Laplace smoothing. Using the training data below, predict the class for the test document "abba": *Class X: "aab", "abaa", "baaa"* *Class Y: "bba", "bbab", "abb"*	3+10
3.	Using the following corpus, calculate the bigram probabilities with add-1 smoothing for the sentence "<s> I like Sam </s>": *<s> I am Sam </s>* *<s> Sam I am </s>* *<s> I like green eggs </s>*	3+12
4.	Explain the Naive Bayes assumption. Why is it called 'naive'? Discuss its advantages and limitations for text classification.	8
5.	What is perplexity in language models? How is it calculated and what does a lower perplexity indicate?	7
6.	Explain the problem of data sparsity in language models. How do smoothing techniques like Laplace (add-1) smoothing address this problem?	8
7.	Explain how neural language models differ from traditional n-gram models. What are the advantages of neural approaches?	8
8.	Using the Naive Bayes algorithm, predict the class of the test document based on the training documents provided: *Training: Class A: "abac", "baabaaa", "abbaa" | Class B: "bbaabbab", "abbb", "bbbaab"* *Test: "aabc"*	7

CHAPTER 6: TRANSFORMERS AND MODERN ARCHITECTURES (6 Hours)

Topics: Attention mechanism, Self-attention, Multi-head attention, Transformer architecture, Positional encoding, BERT

S.N.	Question	Marks
1.	Explain the Transformer architecture in detail. Describe the encoder-decoder structure, self-attention mechanism, and the role of positional encoding.	8
2.	What is BERT? Explain how Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are used in BERT pre-training.	7
3.	Explain the self-attention mechanism and multi-head attention in Transformers. Why are they important for capturing long-range dependencies?	8
4.	Describe the architecture of BERT. Compare BERT-Base and BERT-Large in terms of layers, hidden units, and parameters.	7
5.	What is positional encoding in Transformers? Why is it necessary and how is it calculated using sinusoidal functions?	7
6.	How does the Transformer encoder differ from the Transformer decoder? Explain the role of masked self-attention in the decoder.	8
7.	Compare the Transformer architecture with RNN/LSTM architectures. What advantages do Transformers have for sequence modeling?	8
8.	How does multi-head attention allow the model to attend to different representation subspaces? Explain with an example.	8

Short Notes:

S.N.	Question	Marks
1.	Stemming vs Lemmatization	8
2.	POS tagging	7
3.	BERT	8
4.	Multi-Head Attention	7
5.	GloVe	8
6.	One-Hot Encoding	8
7.	Cosine Similarity	7
8.	BOW vs TF-IDF	7

***** END OF QUESTION BANK *****