Natural Language Processing (315329)

MSBTE Diploma · K-Scheme · Model Question Paper — Set A

Time: 3 Hours **Max. Marks: 70**

Instructions:

1. All questions are compulsory.
2. Answer each next main question on a new page.
3. Illustrate your answers with neat sketches wherever necessary.
4. Figures to the right indicate full marks.
5. Assume suitable data, if necessary.
6. Mobile phone, pager and any other electronic communication devices are not permissible in the examination hall.

Q.1 Attempt any FIVE of the following: **10**

a) Define Natural Language Processing (NLP). State any two of its applications.

b) Define regular expression. Write a regular expression to match a valid 6-digit Indian PIN code (first digit cannot be 0).

c) Define stemming. Give any two examples.

d) Define the Bag-of-Words (BoW) model.

e) What is Named Entity Recognition (NER)? List any four types of named entities.

f) State the formula of cosine similarity. What is its range for term-frequency vectors?

g) Define ambiguity in NLP. Give one example of lexical ambiguity.

Q.2 Attempt any THREE of the following: **12**

a) Explain semantic network with a suitable example and neat diagram.

b) Differentiate between stemming and lemmatization (any four points). Give examples.

c) What is a text corpus? Explain corpus annotation with suitable examples.

d) Perform POS tagging on the sentence: ‘Virat Kohli scored a century in Mumbai yesterday’. Explain the tags used.

Q.3 Attempt any THREE of the following: **12**

a) Explain Byte-Pair Encoding (BPE). Apply BPE (5 merges) on the corpus: play (6), played (3), player (4), ray (2).

b) Explain the attention (self-attention) mechanism used in transformers.

c) Generate unigrams, bigrams and trigrams for the sentence: “Natural language processing helps computers understand human language”.

d) Explain transfer learning in NLP with a neat diagram.

Q.4 Attempt any THREE of the following: **12**

a) Draw the parse tree for the sentence “The old man sat on the bench.” Explain its components.

b) Explain chunking with example. Write NLTK code to extract noun phrases from “The young students read an interesting book in the big library”.

c) Calculate TF-IDF for all terms of the following documents (use TF = count / words in document and IDF = $\log_{10}$(N/df)):
D1: “NLP is fun and NLP is useful” D2: “Python is useful” D3: “NLP uses Python”

d) Explain synonymy, antonymy, hyponymy and polysemy with one example each.

e) Explain the Hugging Face Hub and the AutoTokenizer class with an example.

Q.5 Attempt any TWO of the following: **12**

a) Describe the steps of raw text preprocessing (special characters, contractions, case conversion, tokenization, stop words, lemmatization) with a Python example.

b) Explain the vector space model. Two documents over the vocabulary [data, science, python, model, learning] have term-frequency vectors D1 = [2, 0, 3, 1, 4] and D2 = [1, 2, 2, 0, 3]. Calculate the cosine similarity and interpret the result.

c) Describe Part-of-Speech tagging. Explain rule-based, stochastic and transformation-based tagging techniques with example.

Q.6 Attempt any TWO of the following: **12**

a) Explain the Transformer architecture with a neat encoder–decoder diagram and its working.

b) Write a Python program for sentiment classification of product reviews using TF-IDF features and a Naive Bayes classifier. Show the accuracy and predict the sentiment of new reviews.

c) Explain Word2Vec. Describe CBOW and Skip-gram models with diagrams. Generate skip-gram training pairs for “students learn natural language processing” with window size 2.

Note: This is a **model question paper prepared by Ananda Technologies** only for practice and exam preparation, based on the pattern of previous MSBTE question papers and the K-scheme syllabus. It is **not an official MSBTE paper**, and it does **not** mean that these questions will come in the exam. Answers follow standard textbook points; marks shown are indicative. Free for students — anandatechnologies.in