SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY, SURAT
DEPARTMENT OF ARTIFICIAL INTELLIGENCE
B. Tech-3rd Year (5th Sem)
Natural Language Processing (AI357)
Mid Semester Exam

Date: 25/09/2025 Time: 11:15 AM To 12:45 PM Marks: 10*3=30

1. You need to pre-process each sentence as per the following: **[CO2][2+2+2+2+2=10]**
    a. Tokenize each sentence, handle numbers, punctuations, and URLs.
    b. For every number appearing in the sentence, you need to convert it into a special token **NUMBER**
    c. For any URL appearing in the sentence, convert it into a special token **URL**
    d. For any punctuation symbols appearing in the sentence, convert it into a special token **PUNCT**
    e. Lowercase the complete text.

After preprocessing the sentences, define the following functions for computing TF-IDF scores for each term.

```
def preprocess(sentence):
# Your pre-processing function goes here
def compute_tf_with_normalization(sentence, vocab, smoothing=False):
# Write code for term frequency with normalization based on the total number of words
        if smoothing:
        # Write code for smoothing to handle unseen words
        else:
        # Code for unsmoothed IDF score computation
# Write your code when
def compute_idf(sentence, sentences, vocab, smoothing=False):
        if smoothing:
        # Write code for smoothing to handle unseen words
        else:
        # Code for unsmoothed IDF score computation
def compute_tf_idf_scores(sentences):
# Write code for TF-IDF computation
def main():
# Write the main code for preprocessing, TF, IDF, and TF-IDF computation for all the sentences. Use log scores for both TF and IDF.
```

2. Given the following dataset (First tokenize based on punctuations and lower case all the sentences): **[CO3][7+3=10]**

The boy hugs the cat.
The boys are hugging the dogs.
The dogs are chasing the cats.
The dog and the cat sit quietly.
The boy is sitting on the dog.

Apply the wordpiece algorithm to find the vocabulary of tokens. Apply the merge step for 20 iterations. How will the following sentence be tokenized according to the wordpiece based on the vocabulary you got from the dataset.
The cat is chasing the dog quietly.

3. The following table represents the sentences and corresponding labels in a sentence classification task constituting the training data for the task: **[CO4][3+4+3=10]**

Sentence	**Label**
Check out https://example.com for more info!	Inform
Order 3 items, get 1 free! Limited offer!!!	Promo
Your package #12345 will arrive tomorrow.	Inform
Win $1000 now, visit http://winbig.com!!!	Promo
Meeting at 3pm, don't forget to bring the files.	Reminder
Exclusive deal for you: buy 2, get 1 free!!!	Promo
Download the report from https://reports.com.	Inform
The meeting is starting in 10 minutes.	Reminder
Reminder: submit your timesheet by 5pm today.	Reminder

Design specific features that define a class/label very well. [Hint: Use binary/frequency based features for presence of URLs, Numbers, and Punctuations] Use bag-of-words representation for bigrams appearing in a sentence and use probability of bigrams with add-K smoothing (where K=0.3) as features. Apply the same preprocessing techniques that you used in Question 1 and write all sentences after preprocessing. Find all features and their corresponding probabilities for a Naive Bayes' classifier. Predict the label for the following sentence using the Naive Bayes' classification algorithm:
You will get an exclusive offer in the meeting!