Indian Institute of Technology Jammu, India

COL010P2E - Natural Language Processing

Mid-Term
2 hours

October 26, 2024 **50 marks**

INSTRUCTIONS

- *Conditions of Examination: Closed book; No dictionary; Non-programmable calculator is allowed.*
- *This question paper contains total of 5 Questions. Make necessary assumptions with clear mentions if any.*
- *Do not Copy.*

Question 1: You are given a language with two possible words namely True and False. The following Training sentences with its sentiment label are mentioned below. *(24 marks)*

Training Data: Tr
S1: Mohan does not like reading. (Label: Negative)
S2: Reading is a good habit. (Label: Positive)
S3: Reading is easy but hard to understand. (Label: Positive)

Test Data: Ts
S_{Test}: Mohan reads well. (Label: ?)

a) Represent the Training data Tr in a vector space after pre-processing the dataset. The processing steps should include stop word removal, lemmatization, stemming etc. (4 marks)

b) Represent the training data in a simple Binary, Term Frequency and TF-IDF using unigram Bag of Words representation. (4 marks)

c) Calculate the word embeddings for unique tokens in the training data (Tr) using One-Hot, CBOW, SKIP-GRAM, and LSA representations. For CBOW, SKIP-GRAM, and LSA, a one-dimensional embedding is needed. Consider a context window size of 1. In CBOW and Skip-Gram models, assume that all weight values in the single hidden layer neural network after training are 1, without bias, and with linear activation. The embedding layer has one unit/neuron. Assume the sum of squared error as the loss function. (10 marks)

d) Compute the embeddings of the training and test data using CBOW, Skip-Gram and LSA. [Use aggregate method to generate text embeddings.] (4 marks)

e) Comment on the embeddings generated for the classification task above. (2 marks)

Question 2: Write the appropriate methods to augment the above training data (ref Question 1). *(4 marks)*

Question 3: The court summoned three persons for hearing. The first person always speaks truth, the second person speaks a lie 75% of the time and the third one speaks a lie for 50% of the time. The judge asked a question to one of them. [Note: Make necessary assumptions if any.] *(12 marks)*

a) What is the probability that the answer is given by a liar? (4 marks)

b) What is the probability that the third person has given the answer? (4 marks)

c) Compute the probability that truth has been spoken in the court. (4 marks)

Question 4: Develop a learning framework to generate the title from the following text.

Gemini is Google DeepMind's advanced large language model (LLM), designed to compete with leading AI models. Launched in 2023, Gemini integrates powerful reasoning capabilities, grounded knowledge, and real-time search features, enhancing accuracy and adaptability across tasks. It emphasizes multi-modal capabilities, allowing it to process and generate text, images, and other data types seamlessly. The model is also designed with safety and ethical AI considerations, aiming to reduce biases and prevent harmful outputs. Gemini represents a major step forward in natural language processing, leveraging Google's vast data resources and AI expertise. *(8 marks)*

Question 5: Explain negative sampling in the context of word embedding with an appropriate example. *(2 marks)*

——— End Of Exam ———