Mid Semester Examination
Course Name: Natural Language Processing
Code: CS 563
Full Marks-60 **Time: 2 hours**
Answer ALL the questions

Make reasonable assumptions as and whenever necessary. You can answer the questions in any sequence. However, answers of all the parts to any particular question should appear together.

(Q1). a). Assume PER (Person), LOC (Location), ORG (Organization), MISC (Date, Time, Monetary Expression etc.) and O (Other than named entities) as the possible tags for Named Entity Recognition (NER). Manually tag the following text with NEs. Consider BIO encoding scheme where B, I and O denote the beginning, intermediate and outside the NE.

The Asian giants last won in Australia under Anil Kumble at Perth during their trip Down Under in 2007-08. They were blanked 4-0 under MS Dhoni in 2011-12 and suffered a 2-0 defeat during the 2014-15 tour despite Kohli scoring a truckload of runs. Kohli has also become the first Asian captain to win Tests in Australia, South Africa, and England. BCCI congratulated the Indian Cricket Team for this victory.

b). Design a 3-layered feed-forward neural network (input-hidden-output) to tag the named entities in the above text using a two-step process, i.e. first step deals with discriminating NE (Named Entity) vs. O (Others) and the second step performs classification of NEs into PER, LOC, ORG and MISC. For both the cases, describe your approach and clearly mention

1. the number of neurons at input and output layers and explain why?
2. the dimensions of all the weight matrices that the network needs to learn (including bias terms), assuming hidden layer has 100 neurons.
3. the dimensions of the input and output matrices.

Also, explain one-hot encoding scheme and use it in the above network.

c). Explain with examples how is global information important for Named Entity Recognition (NER)?

(4+8+3)

(Q2). a) Discuss the various dimensions of sentiment analysis with appropriate examples. "Lexicon (containing positive, negative words etc.) of sentiment is not sufficient for sentiment analysis"- explain this with appropriate examples. **(5)**

b). Aspect based Sentiment Analysis (ABSA) provides more fine-grained information that helps in better decision making- explain this with proper examples (show sentence/document level vs. aspect level). **(3)**

c). ABSA comprises of two steps, *viz.* Aspect Term Extraction and Polarity Classification. Suppose feature-based supervised models are to be developed for solving these two problems. Mention the possible set of features that can be used for solving each of these two tasks. **(6)**

d). Consider the following ABSA problem:

No. of classes=3 (Positive, Negative and Neutral); **Domain**: Laptop reviews; **Classifier**: Conditional Random Field and HMM; **Features**: Lexical, syntactic and semantics; **No. of features for aspect term extraction**: 10; **No. of features for sentiment classification: 6**

Build a particle swarm optimization (PSO)/Genetic algorithm-based Feature Selection and Ensemble Construction algorithms for Aspect Term Extraction (i.e. aspect terms have to be identified) and Sentiment Classification (i.e. classification into positive, negative and neutral with respect to each aspect term). Consider (i). "Accuracy" and "F-score" as the objective functions for sentiment classification and aspect term extraction, respectively; and (ii). Majority and Weighted voting techniques are used for combining the outputs of classifiers. For aspect term extraction and sentiment classification, consider F-score and Accuracy as the weights. Show all the steps including encoding, fitness computation, velocity updates/crossover etc with proper examples and explanations. **(10)**

(Q3). *An Argentine international, Messi is his country's all-time leading goal scorer. At youth level, he won the 2005 FIFA World Youth Championship, finishing the tournament with both the Golden Ball and Golden Shoe, and an Olympic gold medal at the 2008 Summer Olympics. His style of play as a diminutive, left-footed dribbler drew comparisons with compatriot Diego Maradona, who declared the teenager his successor.*

For the example above, assume you have only 5 PoS tags N*N (noun), VB (verb), JJ (adjective), RB (adverb) and FF (other, i.e., function words).* Manually PoS tag the above text. Use the convention '***word_POS***'. Explain with mathematical intuitions which amongst the tags given by you will be possible for an HMM based POS tagger to tag and which not. Consider a second order HMM (***Hint***: *You may discuss with reasonable assumptions how lexical, transition probabilities would be computed and decoding would be performed.*) **(5+10)**

Q4). Distinguish between anaphora and cataphora with proper examples. Show the various steps for developing a machine learning based coreference resolution system (*show how positive examples, negative examples would be created, mention the basic set of features and how a algorithm like decision tree would operate to solve the problem*). **(6)**

Best of Luck!