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.

1. It does not matter for an HMM based POS tagger whether it labels from left to right or from right to left. Prove or disprove with rigour. Prove for general K-order HMM. **10**

2. (a). Define "derivational" and "inflectional" morphology with appropriate examples. Distinguish and draw the relationship between morphology analysis and morphology synthesis. Sketch out how can the CYK parser be implemented?

   2+3+5

3. Why is left-recursion a problem in top-down parsing? How can this be eliminated? Distinguish between top-down and bottom-up parsing. Does bottom up filtering improves the efficiency of top-down parsing?-justify your claim. Write down the various steps for implementing a top-down early parser. **3+3+4+4+6**

4. (a). Consider the weighted term vectors of two documents as:

   $D_1 = 2T_1 + 3T_2 + 5T_3$ $D_2 = 3T_1 + 7T_2 + 1T_3$

   For a query, $Q = 5T_1 + 5T_2 + 2T_3$, compute the similarities using *inner product* and *cosine similarity* metrics. With respect to this problem, which one is the better measurement?

   (b). Describe the working principles of K Nearest Neighbor algorithm with respect to document classification. Show its training and testing time complexities. Mention about the shortcomings of this approach and their possible remedies.

   (c). Given a document, containing terms with given frequencies: A (3), B(2), C(1). Assume that collection contains 10,000 documents and document frequencies of these terms are: A (50), B (1300), C (250). Compute the term *frequency-inverse document frequency* of the document collection.

   5+(4+3+2+3)+3