S23CS7.401:
Introduction to
Natural Language Processing

Mid Semester Exam

March 2, 2023

MM: 50 Time: 1.5 hrs

Note: Marks are mentioned next to the questions.

1. Describe and compare Kneser-Ney and Witten-Bell Smoothing methods. Also, succinctly list down the key ideas for both.

   [5+5 marks]

2. Derive HMM's objective function (what are you maximizing?). What are the roles of Viterbi and Forward Algorithms in the conext of HMMs?

   [5+5 marks]

3. In your own words, explain what is Distributional Semantics. Whats are the methods for creating word representations based on Distributional semantics?

   [5+5 marks]

4. Language modeling relies on large amount of raw corpus. Given multiple LMs how would you determine which model has learned better representation from the given data? How do you make statistical LMs robust against sparsity in data?

   [5+5 marks]

5. How would you cluster words such that words with similar POS tags fall in same clusters. No annotated training data is available but large amount of raw text is available.

   [10 marks]

S23CS7.401: Introduction to Natural Language Processing

Mid Semester Exam

March 2, 2023

MM: 50 Time: 1.5 hrs

Note: Marks are mentioned next to the questions.

1. Describe and compare Kneser-Ney and Witten-Bell Smoothing methods. Also, succinctly list down the key ideas for both.

   [5+5 marks]

2. Derive HMM's objective function (what are you maximizing?). What are the roles of Viterbi and Forward Algorithms in the conext of HMMs?

   [5+5 marks]

3. In your own words, explain what is Distributional Semantics. Whats are the methods for creating word representations based on Distributional semantics?

   [5+5 marks]

4. Language modeling relies on large amount of raw corpus. Given multiple LMs how would you determine which model has learned better representation from the given data? How do you make statistical LMs robust against sparsity in data?

   [5+5 marks]

5. How would you cluster words such that words with similar POS tags fall in same clusters. No annotated training data is available but large amount of raw text is available.

   [10 marks]