University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2019

3:30pm-3:50pm, Tuesday, March 19th

Name:
ComputingID:

- This is a **closed book** and **closed notes** quiz. No electronic aids or cheat sheets are allowed.
- There are 2 pages, 3 parts of questions, and 20 total points in this quiz.
- The questions are printed on the **back** of this paper!
- Please carefully read the instructions and questions before you answer them.
- Please pay special attention on your handwriting; if the answers are not recognizable by the instructor, the grading might be inaccurate (*NO* argument about this after the grading is done).
- Try to keep your answers as concise as possible; grading is *not* by keyword matching.

Total	/20

1 True/False Questions (3pts×2)

For the statement you believe it is *False*, please give your brief explanation of it (you do not need to explain anything when you believe it is *True*). *Note the credit can only be granted if your explanation is correct.*

1. HMM for POS tagging problem assumes words are independent from each other.
   ***False**, and Explain*: HMM assumes the words are independent from each other given the tag sequence.

2. In machine translation, a parallel corpus is required to estimate the language model.
   ***False**, and Explain*: The parallel corpus is used to estimate the translation probabilities.

2 Multi-choice Questions (4pts×2)

1. The advantages of MEMM versus HMM in sequential labeling problems include: (a)
   (a) Incorporate arbitrary features;
   (b) More reasonable independence assumptions;
   (c) Guarantee better performance;
   (d) Wider application scenarios.

2. How to use WordNet to measure semantic relatedness between words: (a)
   (a) measure the shortest path between two words on WordNet;
   (b) count the number of shared parent nodes;
   (c) measure the difference between their depths in WordNet;
   (d) measure the difference between the size of child nodes they have.

3 Short Answer Questions (6 pts)

1. Briefly describe how Lesk algorithm solves the word sense disambiguation problem.
   1. Construct the signature of the target word in context by the signatures of the context words defined in the dictionary;
   2. Assign the dictionary sense of the target word whose gloss and examples are the most similar to the signature of context in which the word occurs.