University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2016

5:00pm-5:15pm, Wednesday, March 23rd

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*: : In HMM words are independent only when the tag sequence is given, i,e., $p(\boldsymbol{w}, \boldsymbol{t}) = \prod_i p(w_i|t_i)p(t_i|t_{i-1})$.

2. In machine translation, a parallel corpus is required to estimate the language model.
   ***False**, and Explain*: : a parallel corpus is used to estimate the translation probability; language model is trained on a monolingual corpus.

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 scenario.

2. How to use WordNet to measure semantic relatedness between words: (a) (b)
   (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 Questions (6 pts)

1. Briefly describe how does Lesk algorithm solve the word sense disambiguation problem.
   Simplified Lesk algorithm works as follows,

   (a) construct word signature vector for each word sense defined in a dictionary (based on the glosses and examples);

   (b) for the word in a given text input, e.g., a sentence, use the surrounding words (in a fixed length window) to construct its context vector;

   (c) compare this context vector with the word signature vectors for different word senses, and retrieve the most similar one as the result word sense.

   The original Lesk algorithm expands the context vector of a word by its surrounding words' signature vectors.