University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2015

9:30am-9:45am, Thursday, March 5th

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. Tokenization can be modeled as a sequence labeling problem.
   True: we can use BIO label for each character in the input text sequence, i.e., Beginning of a token, Inside of a token, Outside of a token.

2. WordNet is organized as a fully connected graph, where synsets are the nodes on the graph.
   False*, and Explain*: WordNet is organized as a tree structure, rather than a fully connected graph.

2 Multi-choice Questions (4pts×2)

1. Which of the following is/are generative model(s): (a) (c)
   (a) Hidden Markov Models;
   (b) Conditional Random Field;
   (c) Language Models;
   (d) Maximum Entropy Markov Models.

2. In which of the following situation(s), similarity from distributional semantics is preferred over WordNet-based similarity: (b) (c)
   (a) we need efficient computation;
   (b) we have many out-of-vocabulary words;
   (c) we have a very large domain-specific corpus;
   (d) we need a more accurate similarity metric.

3 Short Questions (6 pts)

1. In POS tagging problems, we need to define the probability of $p(\mathbf{t}|\mathbf{w})$. Write down the specification of this probability in HMMs and MEMMs accordingly.

   Assumed to be both first-order models; answers with other assumptions about the order are also considered as correct.

   In HMM, $p(\mathbf{t}|\mathbf{w}) = \frac{\prod_i p(t_i|t_{i-1})p(w_i|t_i)}{p(\mathbf{w})}$

   In MEMM, $p(\mathbf{t}|\mathbf{w}) = \prod_i p(t_i|t_{i-1}, w_i)$