University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2018

12:30pm-12:45pm, Thursday, February 22

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. Zipf’s law tells us that tail words take the major portion of occurrences in a text corpus.
   
   False, and Explain: tail words take the major portion of vocabulary, head words take the major portion of occurrences in a text corpus.

2. Maximum likelihood estimation is preferred over Bayesian estimation when one only has a small training set.
   
   False, and Explain: maximum likelihood estimation is problematic when the training set is small; instead, Bayesian estimation is preferred in this situation.

2 Multi-choice Questions (4pts×2)

1. Language model smoothing helps us: (a)
   (a) avoid zero probability; (b) reduce perplexity on training set; (c) make the curve of distribution smooth; (d) increase the probability of seen words.

2. Which of the following procedures will reduce the vocabulary size: (a), (d)
   (a) stemming; (b) construct N-gram; (c) TF normalization; (d) Stopword removal.

3 Short Answer Question (6 pts)

1. \( x \) and \( y \) are two n-dimensional unit vectors, i.e., \( \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = 1 \). Figure out the relationship between the cosine similarity between \( x \) and \( y \) and Euclidean distance between \( x \) and \( y \). (Hint: can you compute cosine similarity from Euclidean distance, and vice versa?)

\[
\sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} x_i y_i}}
= \sqrt{2 - \frac{2 \sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}}
= \sqrt{2 - 2 \cos(x, y)}
\]