University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2015

9:30am-9:45am, Thursday, February 12th

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. Given a well-tuned unigram language model $p(w|\theta)$ estimated based on all the text books about the topic of "text mining", we can safely conclude that p("text mining"$|\theta$) > p("mining text"$|\theta$).
   ***False**, and Explain*: they should be equal, since in a unigram language model we do not model the order of words.

2. Given a unigram language model and a bigram language model estimated on the same text collection **without smoothing**, perplexity of the unigram language model will be much larger than that of the bigram language model on this **same training corpus**.
   True

2 Multi-choice Questions (4pts×2)

1. Good "basic concepts" in a vector space model should be: (a) (c)
   (a) orthogonal to each other; (b) based on linguistic study;
   (c) able to automatically compute the weights in each document;
   (d) understandable by human.

2. Zipf's law tells us: (b) (d)
   (a) head words take major portion in English vocabulary;
   (b) in a given French corpus, if the most frequent word's frequency is 1, then the second frequent word's frequency is around 0.5;
   (c) comparing to tail words, removing head words helps more to reduce the storage of documents represented by a vector space model when using a dense matrix data structure;
   (d) smoothing is necessary.

3 Short Questions (6 pts)

1. Let D be a document in a text collection. Suppose we add a copy of D to the collection. How would this affect the IDF values of all the words in the collection? Why?

   The new IDF is,

   $$IDF(w) = \begin{cases} 1 + log(\frac{N+1}{DF(w)+1}) & \text{if } w \in D \\ 1 + log(\frac{N+1}{DF(w)}) & \text{otherwise} \end{cases}$$

   where N the original collection size, $DF(w)$ is the original document frequency of word w.
   Therefore, when w occurs in D, its new IDF decreases (since $N \geq DF(w)$); otherwise, its new IDF increases.