University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2015

5:00pm-5:15pm, Wednesday, February 17th

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. With linear interpolation smoothing, the order between words from maximum likelihood estimation will be preserved.
   ***False**, and Explain*: after smoothing, a previously unseen word might have higher probability than a previously seen word.

2. Stopword removal is not necessary if we use IDF term weighting.
   True

2 Multi-choice Questions (4pts×2)

1. Cosine similarity is usually preferred over Euclidian distance because of: (c)
   (a) computational complexity; (b) interpretability;
   (c) invariant to document length; (d) in the range of [-1,1].

2. Additive smoothing is inferior to absolute discount smoothing because: (a)
   (a) not all words are equally important;
   (b) it is tricky to decide the constant δ in additive smoothing;
   (c) too much probability mass is reallocated in it;
   (d) empirically bad performance.

3 Short Answer Question (6 pts)

1. Given a unigram language model θ, we generate N documents by performing independent sampling from θ. Then, based on those generated documents, we estimate a new unigram language model θ' by maximum likelihood estimation. What can you say about the relationship between θ and θ'.

   $\theta = \lim_{N\to\infty} \theta'$