University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2015

9:30am-9:45am, Thursday, April 30th

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. Since EM algorithm is guaranteed to converge, initialization is not important for it.
   ***False**, and Explain*: EM only guarantees local maximum, so that initialization is important to find better local maximum.

2. Normalized mutual information is preferred over purity when evaluating clustering results, because it is normalized.
   ***False**, and Explain*: Purity does not penalize cluster size: in extreme case when one has equal number of clusters as instances, purity is maximized.

2 Multi-choice Questions (4pts×2)

1. What is true about k-means algorithm: (b),(c)
   (a) it is a variant of kNN;
   (b) convergency is guaranteed;
   (c) it is a greedy algorithm;
   (d) hard to be parallelized.

2. What is true about EM algorithm: (a),(c),(d)
   (a) it is a greedy algorithm;
   (b) it optimizes the upper bound of original objective function;
   (c) it maximize the expectation of the complete data likelihood;
   (d) it can deal with latent variable models.

3 Short Questions (6 pts)

1. Write down at least three different ways to compute distance between two clusters of instances.
   1. single link: minimum distance between any pair of instances from the two clusters;
   2. complete link: maximum distance between any pair of instances from the two clusters;
   3. average link: average distance between any pair of instances from the two clusters.