University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2016

2:00pm-2:15pm, Friday, May 6th

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 k-means is guaranteed to converge, initialization is not important for it.
   ***False**, and Explain*: k-means essentially is a greedy algorithm, initialization affects its convergence to local maximum.

2. Normalized mutual information is preferred over purity when evaluating clustering results, because it is normalized.
   ***False**, and Explain*: We prefer NMI because it penalizes results with many clusters, while purity cannot recognize such trivial results.

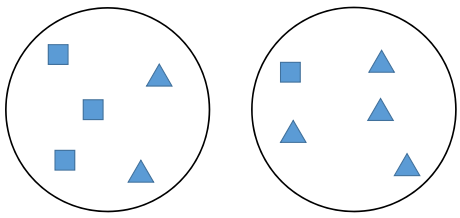
2 Multi-choice Questions (4pts×2)

1. Which of the follow will give us compact clustering results in agglomerative hierarchical clustering: (b)
   (a) single link; (b) complete link; (c) average link; (d) maximum link.

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

3 Short Questions (6 pts)

1. Compute Rand Index of the following clustering result.
   Hint: the two unshaded circles represent clustering results, and the triangles and squares stand for class labels.



TP+FP = $\binom{5}{2} + \binom{5}{2} = 20$, TP = $\binom{3}{2} + \binom{2}{2} + \binom{4}{2} = 10$, TP+FN = $\binom{4}{2} + \binom{6}{2} = 21$
TN = $\binom{10}{2} - TP - FP - FN = 14$, RandIndex = $\frac{10+14}{10+10+11+14} = \frac{8}{15}$

	$w(i) = w(j)$	$w(i) \neq w(j)$
$c(i) = c(j)$	10	10
$c(i) \neq c(j)$	11	14