University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2018

2:30pm-2:45pm, Thursday, May 3rd

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. k-means cannot provide any detailed internal structure within a cluster.
   True

2. The purity metric is intrinsically biased when fewer clusters are returned.
   False, and Explain: It biases towards algorithms that return more clusters; and it fails when every instance is assigned to a different cluster.

2 Multi-choice Questions (4pts×2)

1. Which of the following metric(s) can be used for clustering evaluation: (a), (d)
   (a) Normalized mutual information; (b) Accuracy; (c) Perplexity; (d) Dunn index.

2. What is true about k-means algorithm: (a), (b)
   (a) it is a greedy algorithm; (b) it converges in final steps;
   (c) it is not sensitive to initialization;
   (d) it provides more accurate clustering structure than hierarchical clustering algorithms.

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.

   \[
   TP + FP = \binom{5}{2} + \binom{5}{2} = 20, \quad TP = \binom{3}{2} + \binom{2}{2} + \binom{3}{2} + \binom{3}{2} = 8, \quad TP + FN = \binom{5}{2} + \binom{5}{2} = 20 \\
   TN = \binom{10}{2} - TP - FP - FN = 13, \quad \text{RandIndex} = \frac{8 + 13}{8 + 12 + 12 + 13} = \frac{7}{15}
   \]
   
   \[\begin{array}{|c|c|c|} 
   \hline 
   w(i) = w(j) & w(i) \neq w(j) \\
   \hline 
   c(i) = c(j) & 8 & 12 \\
   c(i) \neq c(j) & 12 & 13 \\
   \hline 
   \end{array}\]