University of Virginia
Department of Computer Science

CS 6501: Text Mining
Spring 2019

3:30pm-3:50pm, Tuesday, April 30

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 clustering is an NP-hard problem.
   False*, and Explain*: k-means is linear to the number of instances and clusters; instead, the original partitional clustering problem is NP-hard.

2. In all clustering algorithms, users need to specify the number of clusters ahead of time.

   False*, and Explain*: In hierarchical clustering, users do not need to specify the number of clusters ahead of time.

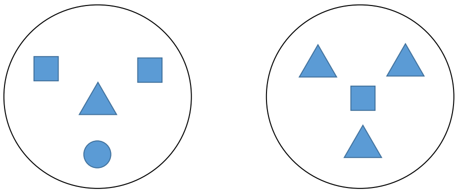
2 Multi-choice Questions (4pts×2)

1. Which of the following metrics can be used to evaluate clustering results when we do not have class labels on instances: (a) (c)
   (a) Davies–Bouldin index; (b) Purity; (c) Dunn index; (d) Entropy.

2. What is/are the input(s) to k-means: (a) (b) (c)
   (a) number of clusters; (b) distance metric; (c) feature vectors of instances;
   (d) class labels on a subset of instances.

3 Short Questions (6 pts)

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



TP+FP $= \binom{3}{2} + \binom{4}{2} = 9$, TP $= 1 + \binom{3}{2} = 4$, TP+FN $= \binom{4}{2} + \binom{4}{2} = 12$
TN $= \binom{8}{2} - TP - FP - FN = 11$, RandIndex $= \frac{4+11}{4+8+5+11} = \frac{15}{28}$

	$w(i) = w(j)$	$w(i) \neq w(j)$
$c(i) = c(j)$	4	8
$c(i) \neq c(j)$	5	11