Homework #4

CS 525/DS 595, Spring 2018

100 points total [6% of your final grade]

Due: April 5, 2018 by 11:59pm
[no submission will be accepted after April 8, 2018 at 11:59pm]

Delivery: Submit via Canvas

K-means clustering
In this assignment, you will implement a basic K-means clustering engine for finding clusters in a collection of text documents. Start by downloading hw4.zip and decompressing it; you should find two python files and two sample document collections.

a) cs525.py - Just like HW1, this helper class will be used to represent a student's identifying information. Any assignments without an instantiated student object of type Student will not be graded. You do NOT need to modify this file.

b) hw4.py: This is a file in which you will write the main code for k-means clustering.

Your goal is straightforward: consume a collection of files (tokenize, calculate tf-idf vectors on their contents) and find k clusters (In this assignment, **k is assigned as 5** – Do **NOT** change it). You should select r random restarts to ensure getting good clusters. To determine the distance of each document to centroids, use **Euclidean distance**. Use Residual Sum of Squares (RSS), an internal metric, to calculate best clusters.

We provide two set of test documents: test10 contains 10 documents, and test50 contains 50 documents. You need to find best 5 clusters for documents in each folder.

Since we have advanced so much in this course (this is the fourth homework, after all!), we leave the details up to you -- including how should the initial seeds be selected? How many restarts are necessary? Etc.

In order to make your implementation more readable, please **briefly explain how you implemented k-means clustering in a document (txt / doc / pdf)** with addressing following questions in the document: What is your r value (the number of random restarts)? Why do you select this number (r value)? What are your stopping criteria to terminate k-means clustering? Why did you select it?

What to turn in:

- Rename hw4.py to hw4_firstname_lastname.py (e.g., hw4_steve_jobs.py) and submit to Canvas your **hw4.py file** and a **document file** explaining how you implemented k-means clustering.
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy). At the top of hw1.py you will see a list of COLLABORATORS. Please fill this out with the names of classmates you consulted and the nature of your discussion.