Homework Assignment 3

To learn the programming and analytics skills, master students are required to complete Option 1: hands-on exercises. PhD students have two options: (1) hands-on exercises and (2) paper reading.

Option 1: Hands-on exercises

1. Use the same data matrix (after the feature selection step) in the home work assignment 2
2. Implement the K-means algorithm by yourself or find K-means source code. While you are encouraged to implement K-mean, I have uploaded a python K-means code in Canvas. You can revise as you need.
3. K-means algorithm computes the distance of a given data point pair. Replace the computation function with Euclidean distance, 1- Cosine similarity, and 1 – Generalized Jarcard similarity. For instance, in the uploaded kmeans.py, there is a function called "distance" defined next:
   - def distance(instance1, instance2):
     - if instance1 == None or instance2 == None:
     - return float("inf")
     - sumOfSquares = 0
     - for i in range(1, len(instance1)):
     - sumOfSquares += (instance1[i] - instance2[i])**2
     - return sumOfSquares

   You can replace this function with different similarity/distance metrics.

4. Run K-means clustering with Euclidean, Cosine and Jarcard similarity. (Specify K as the number of categories of your news articles)
5. Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means. Which method is better and why?
6. Compare the accuracies of Euclidean-K-means Cosine-K-means, Jarcard-K-means. First, label each cluster with the article category of the highest votes. Later, compute the accuracy of the K-means with respect to the three similarity metrics. Which metric is better and why?
7. Compare the accuracies (not SSE) of K-means before feature selection and after feature selection. Can feature selection help improve the accuracies of classifiers and why?
8. Which of Euclidean-K-means, Cosine-K-means, Jarcard-K-means requires more iterations and times and why?
9. Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means with respect to the following three terminating conditions:
   - when there is no change in centroid position
   - when the SSE value increases in the next iteration
   - when the maximum preset value (100) of iteration is complete

   Which method requires more time or more iterations and why?

Deadline: please submit your results by 02/28. You may compress your data matrix, codes, and report into a zip file to submit.

Alternative option for PhD students: Paper reading

Select a clustering paper that has more than 200 citations, read, and present in the class on 02/28. You are encouraged to include motivation, problem formulation, methodology overview, technical details, and experiment interpretations in your presentation. The presentation should be no more than 10 minutes.