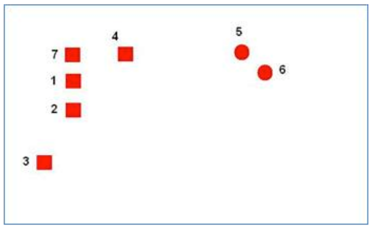
CS 422 Fall 2018

Homework 6

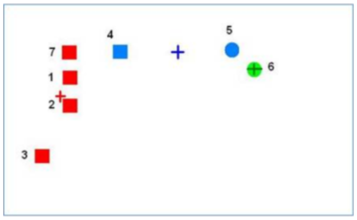
I Clustering

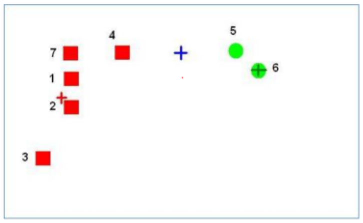
1) Analyze this example of a small set of points, with three initial centroids so that kMeans with k=3 converges to a clustering with an empty cluster. Explain in detail in your own words how does the empty cluster happen. 25 points



- Assume that points 3, 5, and 6 as our initial cluster centers

- explain in your own words why does that particular combination of points location and initial centroids gives an empty clusters. Here are the first 2 iterations of kMeans for this example:





- zero points without detailed explanation.

2) We use SSE(WSS) as the measure of cluster quality and kMeans minimizes it. If there is an empty cluster, can that clustering be the global minimum solution based on RSS? Show all details of you arguments. Use the equation for WSS and how we used it in our discussion in class to prove kMeans convergence. Use your own words. 10 points

II Recommender System

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Consider the following utility matrix, representing ratings on a 1-5 star scale of eight items a-h by three users A, B, and C. Treat ratings of 3, 4, and 5 as 1 and 1, 2 and blank as 0. Compute the Jaccard, and cosine distances between each pair of users. Which one is better? Explain your answer based on how the measures are computed. 15 points

III Page Rank

- Use the PageRank approach to find influential Twitter users
    - PageRank graph is constructed from web pages with hyperlinks. Pages are nodes, and hyperlinks are edges
    - Use the graph of Twitter users and their mentions of other Twitter users. Users are nodes, mention of another users are edges
    - Over this Twitter-user graph, apply the PageRank approach to rank the users. The main idea is that a user who is mentioned by other users is more influential
- Calculate the PageRank for a selection of four users based on the following four tweets:
    - user: Tim, tweet: "@Tom Howdy!"
    - user: Mike, tweet: "Welcome @Tom and @Anne!"
    - user: Tom, tweet: "Hi @Mike and @Anne!"
    - user: Anne, tweet: "Howdy!"
- There are four short tweets generated by four users. The @mentions between users form a directed graph with four nodes and five edges. E.g., the "Tim" node has a directed edge to the "Tom" node.

  Q: Compute manually the first 3 iterations of the PageRank iterations over this 4 node graph. You should use 0.1 as the probability of teleporting. Show all steps of your calculation, provide details and explanations for them (explain the matrices, the vectors you are using, and the equations). Write down the rank order of the 4 users after on you compute 3 iterations. 25 points