Dimensionless Technologies Private Limited

Visit us at: www.dimensionless.in

✉ - info@dimensionless.in

☎ - 9923170071, 8108094992

DOCUMENT CLUSTERING WITH DAILY KOS

Document clustering, or text clustering, is a very popular application of clustering algorithms. A web search engine, like Google, often returns thousands of results for a simple query. For example, if you type the search term "jaguar" into Google, around 200 million results are returned. This makes it very difficult to browse or find relevant information, especially if the search term has multiple meanings. If we search for "jaguar", we might be looking for information about the animal, the car, or the Jacksonville Jaguars football team.

Clustering methods can be used to automatically group search results into categories, making it easier to find relavent results. This method is used in the search engines PolyMeta and Helioid, as well as on FirstGov.gov, the official Web portal for the U.S. government. The two most common algorithms used for document clustering are Hierarchical and k-means.

In this problem, we'll be clustering articles published on Daily Kos, an American political blog that publishes news and opinion articles written from a progressive point of view. Daily Kos was founded by Markos Moulitsas in 2002, and as of September 2014, the site had average weekday traffic of hundreds of thousands of visits.

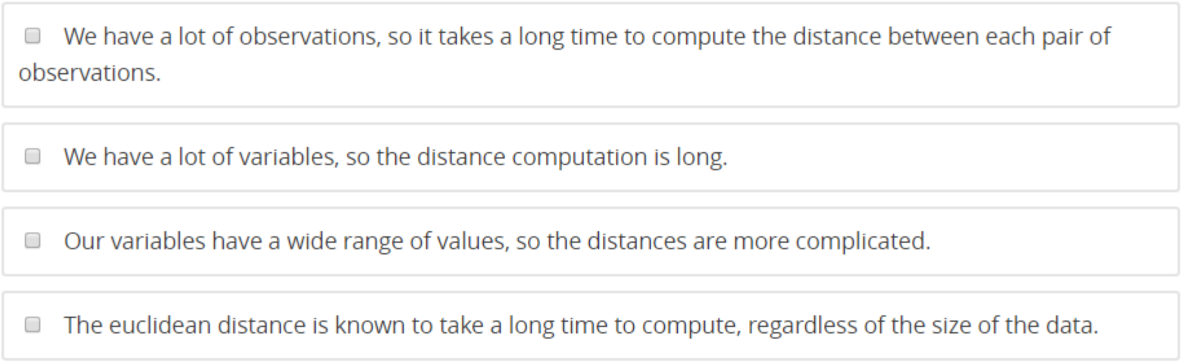
The file dailykos.csv contains data on 3,430 news articles or blogs that have been posted on Daily Kos. These articles were posted in 2004, leading up to the United States Presidential Election. The leading candidates were incumbent President George W. Bush (republican) and John Kerry (democratic). Foreign policy was a dominant topic of the election, specifically, the 2003 invasion of Iraq.

Each of the variables in the dataset is a word that has appeared in at least 50 different articles (1,545 words in total). For each document, the variable values are the number of times that word appeared in the document.

Problem 1.1 - Hierarchical Clustering

Let's start by building a hierarchical clustering model. First, read the data set into R. Then, compute the distances (using method="euclidean"), and use hclust to build the model (using method="ward.D"). You should cluster on all of the variables.

Running the dist function will probably take you a while. Why? Select all that apply.



□ We have a lot of observations, so it takes a long time to compute the distance between each pair of observations.

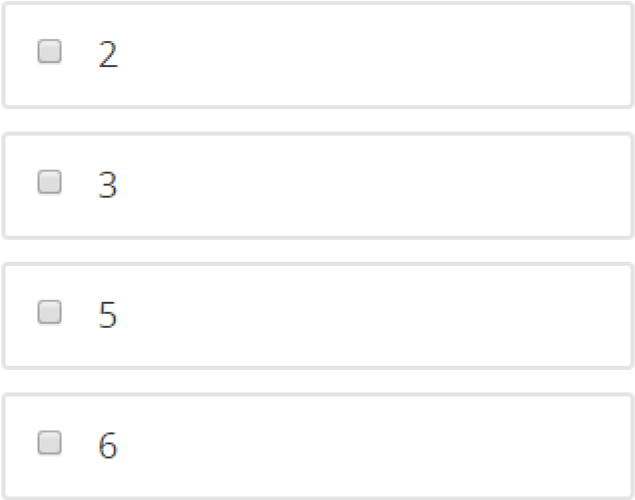
□ We have a lot of variables, so the distance computation is long.

□ Our variables have a wide range of values, so the distances are more complicated.

□ The euclidean distance is known to take a long time to compute, regardless of the size of the data.

Problem 1.2 - Hierarchical Clustering

Plot the dendrogram of your hierarchical clustering model. Just looking at the dendrogram, which of the following seem like good choices for the number of clusters? Select all that apply.



□ 2

□ 3

□ 5

□ 6

Problem 1.3 - Hierarchical Clustering

In this problem, we are trying to cluster news articles or blog posts into groups. This can be used to show readers categories to choose from when trying to decide what to read. Just thinking about this application, what are good choices for the number of clusters? Select all that apply.

☐ 2

☐ 3

☐ 7

☐ 8

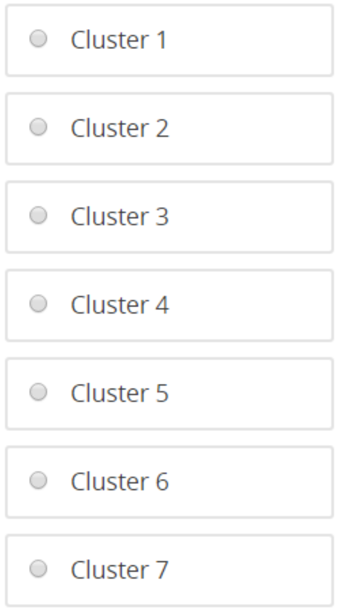
Problem 1.4 - Hierarchical Clustering

Let's pick 7 clusters. This number is reasonable according to the dendrogram, and also seems reasonable for the application. Use the cutree function to split your data into 7 clusters.

Now, we don't really want to run tapply on every single variable when we have over 1,000 different variables. Let's instead use the subset function to subset our data by cluster. Create 7 new datasets, each containing the observations from one of the clusters.

How many observations are in cluster 3?

Which cluster has the most observations?



○ Cluster 1

○ Cluster 2

○ Cluster 3

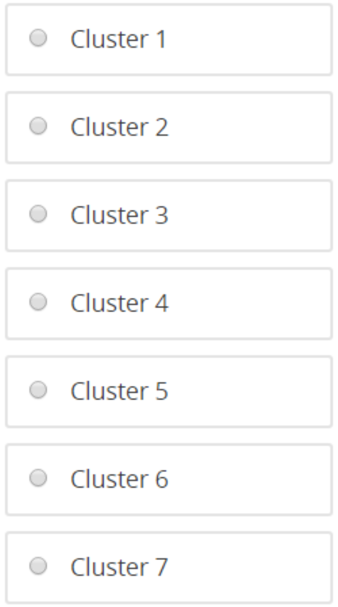
○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

Which cluster has the fewest observations?



○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

Problem 1.5 - Hierarchical Clustering

Instead of looking at the average value in each variable individually, we'll just look at the top 6 words in each cluster. To do this for cluster 1, type the following in your R console (where "HierCluster1" should be replaced with the name of your first cluster subset):

```
tail(sort(colMeans(HierCluster1)))
```

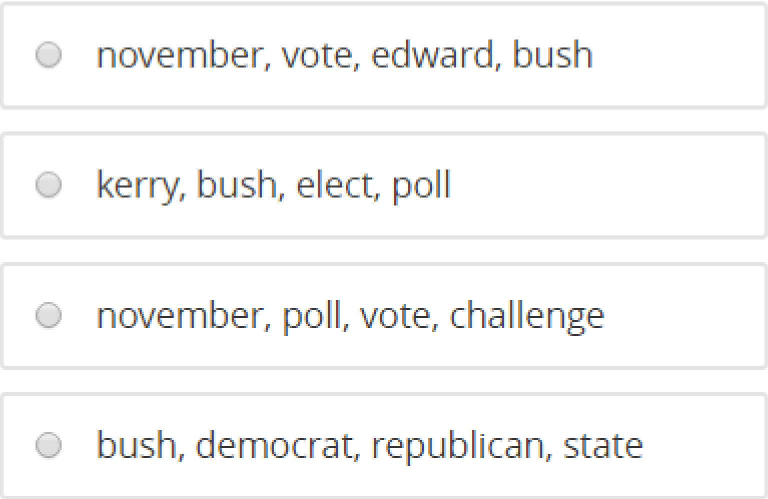
This computes the mean frequency values of each of the words in cluster 1, and then outputs the 6 words that occur the most frequently. The colMeans function computes the column (word) means, the sort function orders the words in increasing order of the mean values, and the tail function outputs the last 6 words listed, which are the ones with the largest column means.

What is the most frequent word in this cluster, in terms of average value? Enter the word exactly how you see it in the output:

Problem 1.6 - Hierarchical Clustering

Now repeat the command given in the previous problem for each of the other clusters, and answer the following questions.

Which words best describe cluster 2?



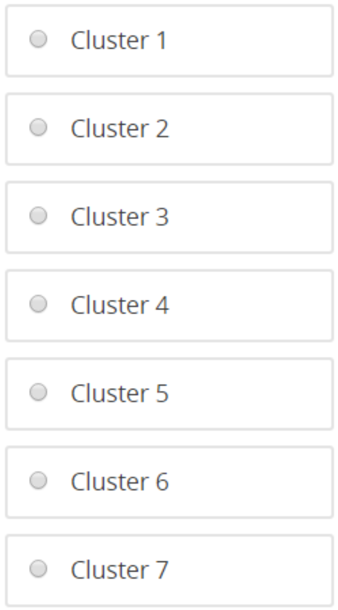
○ november, vote, edward, bush

○ kerry, bush, elect, poll

○ november, poll, vote, challenge

○ bush, democrat, republican, state

Which cluster could best be described as the cluster related to the Iraq war?



○ Cluster 1

○ Cluster 2

○ Cluster 3

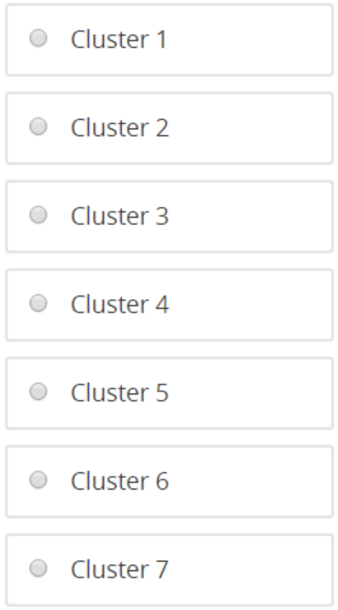
○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

In 2004, one of the candidates for the Democratic nomination for the President of the United States was Howard Dean, John Kerry was the candidate who won the democratic nomination, and John Edwards with the running mate of John Kerry (the Vice President nominee). Given this information, which cluster best corresponds to the democratic party?



○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

Problem 2.1 - K-Means Clustering

Now, run k-means clustering, setting the seed to 1000 right before you run the kmeans function. Again, pick the number of clusters equal to 7. You don't need to add the iters.max argument.

Subset your data into the 7 clusters (7 new datasets) by using the "cluster" variable of your kmeans output.

How many observations are in Cluster 3?

Which cluster has the most observations?

Which cluster has the fewest number of observations?

Problem 2.2 - K-Means Clustering

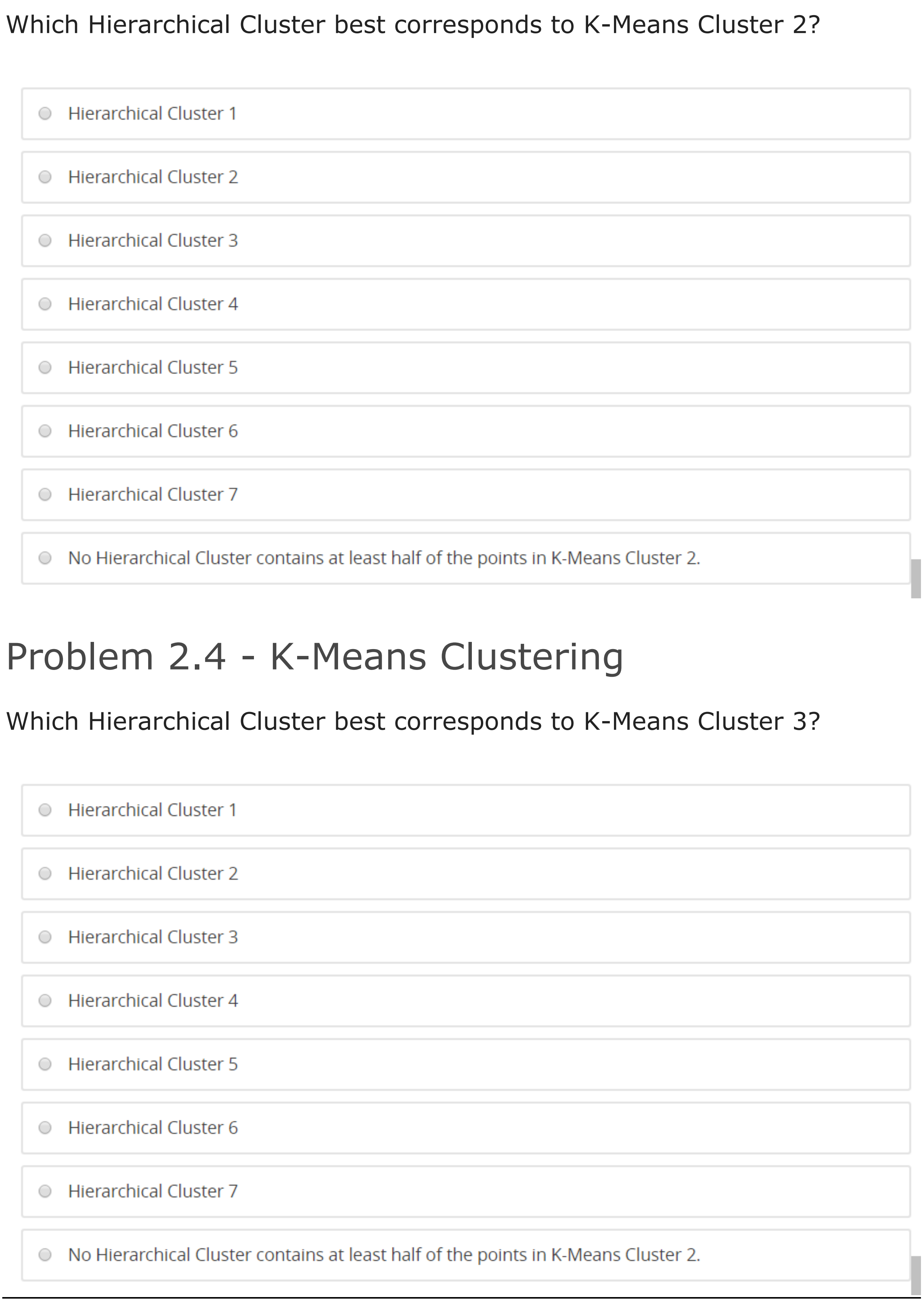
Now, output the six most frequent words in each cluster, like we did in the previous problem, for each of the k-means clusters.

Which k-means cluster best corresponds to the Iraq War?

Which k-means cluster best corresponds to the democratic party? (Remember that we are looking for the names of the key democratic party leaders.)

Problem 2.3 - K-Means Clustering

For the rest of this problem, we'll ask you to compare how observations were assigned to clusters in the two different methods. Use the table function to compare the cluster assignment of hierarchical clustering to the cluster assignment of k-means clustering.



Which Hierarchical Cluster best corresponds to K-Means Cluster 2?

○ Hierarchical Cluster 1

○ Hierarchical Cluster 2

○ Hierarchical Cluster 3

○ Hierarchical Cluster 4

○ Hierarchical Cluster 5

○ Hierarchical Cluster 6

○ Hierarchical Cluster 7

○ No Hierarchical Cluster contains at least half of the points in K-Means Cluster 2.

Problem 2.4 - K-Means Clustering

Which Hierarchical Cluster best corresponds to K-Means Cluster 3?

○ Hierarchical Cluster 1

○ Hierarchical Cluster 2

○ Hierarchical Cluster 3

○ Hierarchical Cluster 4

○ Hierarchical Cluster 5

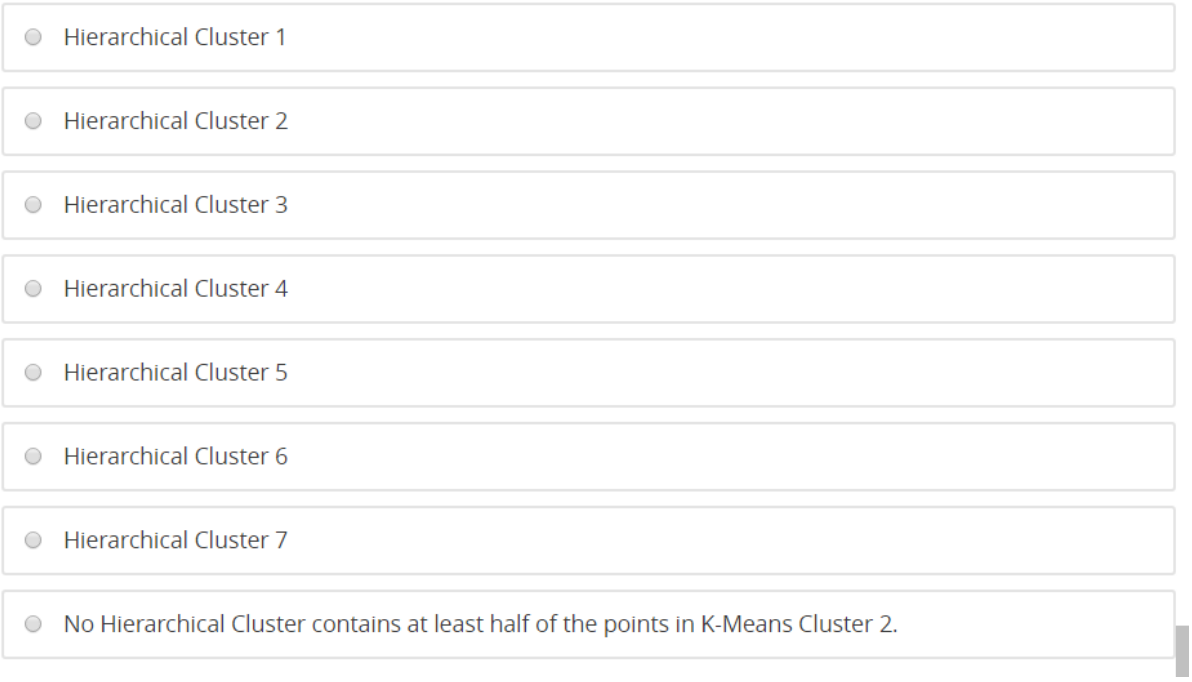
○ Hierarchical Cluster 6

○ Hierarchical Cluster 7

○ No Hierarchical Cluster contains at least half of the points in K-Means Cluster 2.

Problem 2.5 - K-Means Clustering

Which Hierarchical Cluster best corresponds to K-Means Cluster 7?



○ Hierarchical Cluster 1

○ Hierarchical Cluster 2

○ Hierarchical Cluster 3

○ Hierarchical Cluster 4

○ Hierarchical Cluster 5

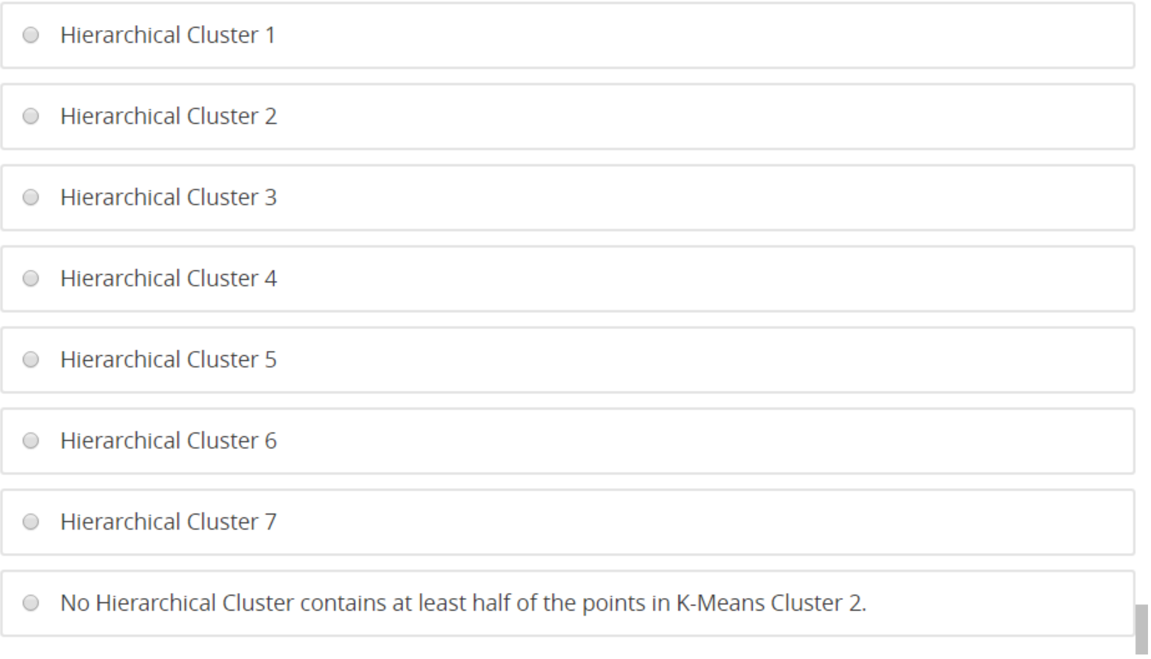
○ Hierarchical Cluster 6

○ Hierarchical Cluster 7

○ No Hierarchical Cluster contains at least half of the points in K-Means Cluster 2.

Problem 2.6 - K-Means Clustering

Which Hierarchical Cluster best corresponds to K-Means Cluster 6?



○ Hierarchical Cluster 1

○ Hierarchical Cluster 2

○ Hierarchical Cluster 3

○ Hierarchical Cluster 4

○ Hierarchical Cluster 5

○ Hierarchical Cluster 6

○ Hierarchical Cluster 7

○ No Hierarchical Cluster contains at least half of the points in K-Means Cluster 2.