Problem Description

“store_cluster_assignments” is a simple file with 245 "stores" (StoreId) along with a ClusterId that they are each assigned to. If you remember, we do some analytics to cluster client stores into cluster groups based on a host of variables (store sales, total population, total competitors, etc.) so each ClusterId represents some number of stores that are very similar to one another.

What I'd like you to do is to write a Python program (or R, or C++) that that uses this file to select an "optimal sample". The way we think about an optimal sample is for a given sample size (e.g., 20 stores, 21, 22, etc.), selecting an optimally proportional amount of sample stores to the amount of population stores. A simplified example:

If you had 100 total stores that fell into 4 clusters like this...

ClusterId	Count_of_Stores
1	25
2	25
3	10
4	40

(i.e., 25 stores are in cluster 1, etc.)

...than if you were picking a sample of 20 stores, you pick five in cluster 1, five in cluster 2, two in cluster 3, and eight in cluster 4 because that would be proportionally the sample as the count of sample stores across the cluster groups.

Also, the metric we have used to compare similar proportions for two vectors is cosine similarity, so if you took the cosine similarity of vector1=[25,25,10,40] and vector2=[5,5,2,8] you would get a value of 1.0, which is optimal (cosine similarity is a value between 0 and 1, with 1 being the most similar). In this link there is some documentation on how to do cosine similarity in Python (see response from charmoniumQ): http://stackoverflow.com/questions/18424228/cosine-similarity-between-2-number-lists.

Also note that once you get the right proportions of stores to select from each cluster we typically take a random sample of the stores in the cluster (i.e., in the example above, I would take a random sample of 5 stores from the 25 stores that are in cluster 1).

Lastly, for extra credit write your program so that the sample size is a parameter--i.e., I should be able to tell your program that I want a sample size of 20 or 30 or whatever, and it should spit out the StoreId's for that sample size.

Let me know if you have any questions. Spend no more than 2 hours with it and let me know where you get to. If you don't make it to writing any code by then send me some of your thoughts on how you'd write the code, and we can then evaluate that.

Hint: The trickiest part of this case study is determine an algorithm that will guarantee the optimal sample selection (ie, maximize cosine similarity). After you have read in the data and understand the prompt, be sure to give yourself time to think through various approaches to develop a program that ensures an optimal sample. (There are definitely more than one way to accomplish this.)

Have fun!