Title of your report

S. Tudent (12345678)

In these guidelines we provide a more detailed idea of what is expected of your experiments and of the report. For each section there is a rough indication of how much text is expected and which figures. Depending on your results certain parts may require more or less text or more or fewer figures. The main goal of your experiments is to study the performance, running time, and number of rounds of one of the MapReduce algorithms for the (k, ϵ)-coreset for the k-means problem that is presented in the class. We have learned a MPC algorithm that computes a (k, ϵ)-coreset for the k-means problem. The algorithm is for d-dimensional Euclidean spaces and needs $O(\log n)$-rounds of communication.

1 Introduction

[1-2 paragraphs] A brief introduction of what the goal of your experiments is.

2 Data sets and experiments

[3-4 paragraphs, with images of data types] A description of your data sets and experimental setup (which algorithms are you running and with which parameters). Also briefly explain how these help you answer your research question. For efficiency you are usually interested in larger point sets. A point set of 20 or even 100 points is not likely to really show a significant difference between the algorithms. Often, how the running time scales with the number of points is more interesting. You might want to create a sequence of data sets that have similar properties, but more increase in number of points. Generate or find data for you experiment.

Please choose datasets whose size are reasonably big, say having all these sizes

$$1000, 5000, 10000, 50000, 100000, 500000, \text{ and } 1000000 \ .$$

See how your algorithms behave when you increase the number of points. You can check how large could be the size of the point set that you can test your algorithm. Please test your algorithm on at least 5 datasets, two of them should be real datasets or image segmentation/compression/processing.

3 Experimental evaluation

[2-3 paragraphs with figures] A presentation of the results of your experiments as well as a brief discussion of them. The presentation would generally be in the form of a graph or table. Be sure to focus on the interesting parts. For example, if in most of your experiments the results are exactly as expected, then you can briefly show or state that and then focus on

less expected results. In the discussion you consider what the results imply and discuss any potential irregularities or unexpected results.

Run experiments on your data sets using your implementation of the MPC algorithm for the (k, ϵ)-coreset for the k-means problem so you can compare the resulting clustering with the true clustering of your data and the efficiency of the algorithms. You present and discuss the results in your report and explain what conclusions can (or possibly cannot) be drawn from your results. Try to specify the number of logical machines that you use, the space per machine, and the number of rounds that your algorithm use. Make a plot of them.

Example research questions:

- Compare the coreset sizes and see how it effects the running time and the quality of clustering. For example, if you increase or decrease ϵ and/or k, what will happen to the size of the coreset and the quality of clustering.
- Instead of placing a grid inside the inner balls of each cluster or the annuli, you can sample points inside each inner ball or annulus and take the union of the sample set as the coreset. The number of samples from each annuli can be a good research question. You can plot the number of samples against the quality of the clustering. You can also compare the coreset where you place grids inside annuli and the coreset where you sample points inside annuli and compare the running time and the quality of clustering for both coresets.
- You can test the coreset for image quantization, segmentation and compression. You can also look at the following papers:
  - Medical Image Segmentation Using k-Means Clustering and Improved Watershed Algorithm, Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation
  - There is a beautiful paper on k-means with lots of datasets and images segmentation. Here is the name of the paper: "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" [1]. Table 1 and Figures 9 and 10 explain lots of examples. You can run your experiments on these datasets.
  - Also, you can have a look at experiments in this paper: “Gereon Frahling, Piotr Indyk, Christian Sohler: Sampling in dynamic data streams and applications. SCG 2005: 142-149”.

Depending on your research question you can focus on one or several of the following measures depending on your research question.

- How long is the actual running time of the algorithm, how does it scale with the number of points.
- Which parts of the algorithm are the bottleneck? How long do different parts, such as initializing data structures, coreset construction and so on?

[1] The paper is available at https://www.cs.umd.edu/ mount/Projects/KMeans/pami02.pdf

4 Concluding remarks

[1-2 paragraph] A summary of the main conclusions of the experiments. Do your results help answer (part of) your initial research question?