The main goal of your experiments is to study the performance, running time, and number of rounds of one of the MapReduce algorithms that is presented in the class. We have learned two algorithms for the MST problem. The first algorithm is for dense graphs and needs $O(1/\epsilon)$ communication rounds. The second algorithm is for general graphs and needs $O(\log(n))$ rounds. You can choose either one that you think you can implement better and do your experiments.

You will need to do several things:

- Decide on a more precise research question that you want to answer. Some examples for inspiration are listed below.
- Generate or find data for you experiment.
- Run experiments on your data sets using your implementation of the MST algorithm 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.

This then goes into a report that we use to grade your experiments. On the next page you can find the rubric that we will use to grade the report so you have an idea of what to focus on.

Example research questions:

- In data with much noise interested, does the MST algorithm or the affinity clustering algorithm that you implemented creates clusterings that are close to the true clustering of your data?
- For data sets with clearly separated, but different sized clusters, does the MST algorithm that you implemented provides the best clustering? If not, what do you suggest to improve the quality of the solution reported by your algorithm?
- Does your MapReduce algorithm converge very fast on your data sets and for what datasets do you think your MapReduce algorithm would perform better? Is there a heuristic or any other method that could reduce the communication rounds of your algorithm?
- Does your algorithm works well when the number of clusters is large (over 100)?

Data generation:

- The python library sklearn (https://scikit-learn.org/stable/datasets/index.html) has several functions that allow easy data generation. You can use this to generate many different types of data. It also contains some data-sets, but they are mainly higher-dimensional.

Assessment of the reports for 2IMA35 is based on an evaluation of various aspects, as listed below. The aspects are evaluated on a scale consisting of five possible scores, which can be interpreted as follows:

poor	almost satisfactory	decent	good	excellent
O	O	O	O	O

For each aspect there is a short explanation about what “weak” means, and what you need to do get the score “excellent”. You can earn a maximum of 15 points for this reports based on scores for the listed aspects. Here’s a rough indication of how the scores are translated to the final number of points:

- Several scores on the experiments are poor, other scores are mostly almost satisfactory ➔ final grade ≤ 5
- Most scores, in particular on the contents, are almost satisfactory ➔ final grade 6-7
- Most scores are decent, some may be almost satisfactory ➔ final grade 8-9
- Most score are good, some may be decent ➔ final grade 10-12
- Almost all scores are at least good, and most scores on the contents are excellent ➔ final grade ≥ 13

General

Appearance

No use of sections or it is generally hard to find things in the report.	O	O	O	O	O	The report looks nice, figures fit with the text and the contents is divided well into the sections.

Writing

The report is hard to read due to the many grammatical and spelling errors or very chaotic sentence structures.	O	O	O	O	O	The report is a pleasure to read: sentences are grammatically correct, easy to parse and follow each other naturally; there are hardly any spelling errors.

Setup

Research question

Research question has an obvious answer or is impossible to answer.	O	O	O	O	O	Clear, non-trivial research question that fits the proposed area.

Data sets

Chosen data sets do not help answer research question.	O	O	O	O	O	Data sets are varied and highly relevant for the research question.

Experimental Set up

Experiments are too limited to provide any reasonable evidence towards answering research question.	O	O	O	O	O	Chosen experiments are varied and will be able to contribute strongly towards answering research question.

Experimental results

Presentation and discussion of results

Results are missing or unreadable	O	O	O	O	O	Results are presented clearly and concisely and most important features are highlighted.

Conclusion

Results are not used to make any conclusions about the research question	O	O	O	O	O	A conclusion is drawn from the results about the research question that is substantiated by the results.

-