NDB048 Data Science – Structure of Report

1. Introduction

- What problem do we solve? What should be the result of our work?
- What questions are asked?
- What is included in this report? What parts does it have?
- What data do we have (generally)? What sources and technologies can we use for processing?

2. Planned steps

- What methods are available for problem solution? What do we choose and why?
- What individual steps do we take? Can we plan them to the end or do they depend on current results? What points of resolution do we expect?
- Are there any conditions for using some methods? Are they met (and what if not)?

3. Data

- What is the data source? Are there any marks that we cannot trust the data?
- Data description (size, format, fields, data types, …)
- Data quality and sanity check (missing, errors, weird values) and resolution for data preprocessing and cleaning

4. Exploratory analysis

- Basic statistics + visualizations
- Conclusions about relationships, typical and rare cases
- Can one expect it? If not, what is surprising and what is expectable?
- Recommendations for data transformation and (possibly) excluding some data

5. Data transformation and Modeling

- Feature engineering, data transformation
- Fitting predictive or classification model (if not stated differently; possibly simple) for a proper target
- Model performance evaluation

6. Results and Discussion

- Description and assessment of the results
- Description of usage of the approaches for Big Data processing (MapReduce / Spark / multi-model DB) for a selected part of the analysis
- Possibly description of the iterations made

7. Summary

- Summarization of the findings, answers to questions in the Introduction
- Possible recommendations for next steps

Typical problems:

1. The problem not introduced
2. Missing or too brief description of steps
3. Too complex aim of analysis
4. Useless graphs/tables (Does it carry any useful information? Why is it there? Is it worth it?)
5. Too many similar graphs, inappropriate type of graphs
6. No comments on the graphs/tables (What insights a graph provides should be summarised in text, too.)
7. Graphs without appropriate description (axis labels, titles, population constraints, ...)
8. Graph inconsistencies across a whole document (color, scales, graph types, …)
9. Too little of explaining text, bad arrangement of text, chart and tables
10. No summary/conclusion
11. Useless information (copying information from slides, list of used technologies, …)
12. Reproducibility problems, code behind the report issues