PROJECT GUIDELINES

MANAGING BIG DATA (MBD)
2022-2023

DOINA BUCUR

Project goals:

- Write a *research question* which can be answered by data processing fairly large datasets, using the new distributed programming tool you have learned in this course. The choice of research question is up to you; it could be a question that can be answered using a dataset we already have, or it may require you to find a dataset. There is **no minimum dataset size**, but have (say) 1+ million data records or (say) 5+ GB of data. There is, however, a **maximum dataset size**, because our computing cluster is limited: stay under 100 GB if possible, so we don't run into capacity problems.

  Your project should be *research*, namely:

  (a) You need to show that your work is in some way *new*, and has not been done in exactly the same way before (use `Google Scholar` to find similar work). Doing new work is not very hard: you could take an existing research question (say, analyse a type of social media to understand how humans react to certain events) but focus on a new time frame, geographical location, social or cultural events, or even language.

  (b) You should be able to argue that your research question is **interesting** to someone, and thus it should be done. Does it have some *impact* on society? Is it perhaps useful in a branch of industry? You can also pick a topic you yourselves have some passion for; this may lead to further ideas for (or even your future MSc thesis).

- Answer your research question in the time and with the computing resources you have at your disposal. This means that you should pick a *feasible* research question. Per student, you only have 2 EC worth of time allotted for this project (which means 50 hours of work per student over 4 weeks); multiply this by the number of students in your team. Your project will be evaluated keeping in mind this number of working hours.

- Work in project *groups*. This is a skill that our study program wants to give you a lot of practice with, and a skill which is valuable in the IT industry. You are free to choose your project team. Aim for 4 students per team.[1]

- Delivering the results of your project in *written and presentation form*. A single report (or paper) of 8-10 pages is expected, in any report style. You will be able to choose your presentation date from among some options (see the sheet with the course schedule).

Doina can give you feedback on your research question (its impact and feasibility) at any point in time. You are, however, responsible for its design, implementation, and debugging; you should already be familiar with the class of algorithms you plan to design or apply, and should 'explore' your datasets early on, to make sure your plan is feasible to complete in time.

---

[1] (Exceptions are possible if needed, say plus/minus 1 in the team. I advise against having fewer than 3 members. We had small teams sometimes before, and they had a hard time.)

Evaluation criteria. For all research projects, we have fairly standard grading guidelines. The grade components below are universal (they count for any research-oriented project you do); only the weight of each in the final grade can vary.

The grade components are: (1) the *scientific content* (always the largest component of the grade), (2) the *communication* of results (written and oral), and finally (3) the *process organization* by the project team (how independent was the team? did they keep to deadlines? could they solve their own problems? were they polite to others on the cluster?).

Here's what these components mean, and how they'll be weighted for this project:

- **Scientific content** (60%). This grade will reflect the answers to the following questions:
  - What is the utility of this research question? (everything counts: societal impact, industrial utility, value for other researchers).
  - Are the results obtained trustworthy (statistically significant)? Is the research methodology sound? Are the conclusions justified by the results? Have the authors pointed out themselves any weaknesses in their research procedure?
  - What was the difficulty level of this solution? Is it at the level of BSc/MSc courses in Computer Science at the UT, or has additional fundamental knowledge been used from the scientific literature? Does it match the time available to work on this project?
  - *Grading*: A grade of 7 ("amply sufficient") means, as an example, that work has been done at the level of the UT courses, and this has had led to some new insights with regard to the research question. A grade of 9 ("very good") could be given for work that goes beyond the level of the UT courses, and/or cross-disciplinary work which could be published as a scientific paper in the near future. A grade of 10 ("excellent") goes to brilliant work, which may begin a new research theme at the UT. Generally, plus points go for: high-impact research, timely research on a hot topic of broad interest, high difficulty in the solution, or high originality.
- **Communication** (30%). This is about the writing of the project report, and the presentation skills. It measures the following:
  - Was the report (and presentation) a complete description of the work done? Were the text and visualizations well chosen, clear, unobfuscated, neither longer nor shorter than required by the scientific content described? Was there a good related-work section?
  - Were good visual aides used? This can mean data visualization, graphical descriptions of procedures (e.g., flowcharts for tool chains), tables and lists.
  - *Grading*: A grade of 7 ("amply sufficient") means that the report would need only minor corrections to "polish" into a good report, and is quite readable; the presentation should be a correct representation of the entire work done. For a grade of 10, (I quote from an assessment guideline for theses), "the report can serve as teaching material or a publication; the presentation was pure entertainment, while leaving everybody feeling that they learned a lot."
- **Organization** (10%). This grade measures skills highly valued by your future employers:
  - how much guidance was needed[2];
  - the independence and initiative of the project team;
  - the ability of the project team to deal with (adapt to) any unforeseen circumstances, and figure out practical issues.
  - *Grading*: A grade of 9 can be given if the teacher is (quote again from an evaluation template) "happy that they were allowed to be involved in this assignment". A grade of 10 is given if a teacher close to your topic would learn something new from your project. 7 goes to the most standard project: guidance was necessary, but has been sought by the students in good time.

NB: Grades of 10 overall for project work with a research component are very difficult to obtain. However, project grades of 9 or above do happen almost every year.

[2]Updates about intermediate results, project status, and new ideas doesn't count as guidance, but as... socializing.