Heidelberg University
Database Systems Research Group
Prof. Dr. Michael Gertz
Satya Almasian, Dennis Aumiller

November 02, 2020
Text Analytics
Winter Semester 2020/21

Project Proposal Guidelines
Proposal due on Friday, November 27, 2020, 18:00, via Moodle

General Notes:

- The project proposal determines a significant part of the remaining work in this course. Think about your idea carefully. The proposal itself will count for 15% of your final grade for the group project.
- As mentioned before, the project is a **team effort of three to four students**. The students must be in the same groups as for the assignments. Groups of four students are expected to propose larger projects.
- You are not allowed to share projects between groups.
- Proposals need to be **uploaded as a single PDF of up to 5 pages**. More specific instructions on the contents can be found in this document.
- Source code needs to be made available via GitHub. If you do not want to publicly share the project, you must at the very least add instructors to the project as collaborators.
- We will assign project mentors based on the topics within the week of your proposal submission, and your group will be contacted by your respective supervisor shortly after submitting your proposal.
- For further questions, please reach out via the Moodle forums, or contact an instructor.

---

1 General Submission Guidelines

- **Title**: Title of your project (does not need to be final).
- **Team Members**: The name of all team members. The participating students must match a group that is handing in assignments together.
- **Mail Addresses**: The preferred email of each team member.
- **Project Source**: Indicate whether you are using an existing document collection (in this case, provide a link to the data), or alternatively how you intend to collect your dataset (include a link).
- **Existing Code Fragments**: In case you already know that you will be using a third-party or prior project, you must clearly indicate those parts of your project.
- **GitHub Repository Link**: This is the location where all changes need to be documented during your project.

---

2 Motivation (0.5 pages)

The first part of your proposal should focus on a *practical problem* that you are trying to solve. Try to give reasons for why your project is a relevant topic in text analytics, and how/what you are addressing in your specific project. It should give an idea of what to expect from the later parts of your proposal as well.

---

3 Research Topic Summary (1-2 pages)

The second part should focus on the theoretical parts of your work, and specifically related research in that area. To give you some idea of what to include, see the following list:

- **Paper/Article Sources**: Your proposal should include references of related work in the form of scientific papers or projects that are adjacent to your own idea. While you can also cite blog posts/articles (e.g., from medium.com / towardsdatascience.com, etc.), we strongly encourage you to find papers that have ideally been published in a peer-reviewed conference/journal. Papers are also relevant sources when you are using existing datasets and can help you with a description of how their data was obtained.
- **Citations**: Please include any articles that you are referencing in your work as a proper citation, including author names, publication venue, URL (if relevant) and title.
- **Theoretical Summary/Method**: Describe your main contributions from a theoretical point of view. What issues with current approaches are you trying to address? How can you optimize existing work to better fit your specific problem setting? What are the main areas that you plan to focus on?
- **Research Context (Advanced)**: How does your project fit into the wider context of the specific area of text analytics? Does it have potential application in other tasks as well? Can downstream tasks be expected to perform better with this method?

---

4 Project Description (1-2 pages)

Describing your project should take up the largest part in your project proposal. Here, it is essential to describe your approach in more detail, and indicate your text processing pipeline, as well as the planned distribution of tasks between members.

- **Main Project Goals**: In extension of the previous motivation, you should take some more time to describe the practicalities of your approach, this time with a clearer focus on the application. Detail the main goal, and - if possible - specific subgoals that you are trying to aim for along the way. This will help you to track your progress and make it easier to divide up single tasks between group members.
- **Text Analytics Tasks**: This should already be mentioned in the related work section of the theoretical section, but you can re-state the kind of Text Analytics task that is relevant for this project.
- **Pipeline**: To describe what sort of pipeline you are proposing, a schematic figure might help you in explaining your setup.

- **Used Dataset**: Similarly, please re-state the dataset (existing or self-collected) you are planning to use. Address potential problems (specifically when you are collecting your own data, or are changing the original purpose of another dataset) that may arise with the dataset.
- **Evaluation**: Describe how you plan to analyze the results from your experiments. Are single numerical evaluation metrics possible? Is an automatic evaluation possible? Consider the (dis-)advantages of specific evaluation metrics that you might be using, and ideally analyze multiple different scoring methods.
- **Baselines**: For the evaluation, you should also think about sensible baselines (Random, "Oracle", simple approaches, such as TF-IDF, etc.). Are there existing scores available in related literature?

---

5 Grading Criteria

For grading, we will be looking at several criteria in your submitted reports, among others:

- **Thematic Fit**: Does your project show a clear fit within the Text Analytics space? Did you incorporate methods and concepts from the lecture?
- **Creativity**: Additional bonus is given for creative approaches. Are you utilizing existing datasets, or did you find a smart way to acquire new samples for your project? Is there domain knowledge that you can incorporate?
- **Research Context**: Did you take the time to find relevant related approaches? Are there existing projects that have tried something similar in a practical context?
- **Well-roundedness**: Your report should convey a coherent project idea that gives a clear understanding of your planned approach, and do so across the different chapters.
- **Team Organization**: Did you consider sensible subtasks that can be solved in parallel by the various team members? Are the ideas concise and achievable enough to ensure a positive project outcome, or at least achieve partial completion?
