

## Review of Bachelor's Thesis

**Student:** Trampeška Václav  
**Title:** Exploratory Analysis of Big Data in "jmenomesto" (id 23348)  
**Reviewer:** Plchot Oldřich, Ing., Ph.D., DCGM FIT BUT

1. **Assignment complexity** **less demanding assignment**  
This work was focused mainly on analyzing the fixed dataset, building a database out of it and an application that enables queries into the dataset. The direction was straightforward and did not require the student to get acquainted with any complicated theory.
2. **Completeness of assignment requirements** **student reasonably departed from the assignment with minor reservations**  
The main goals of the work were achieved as the student managed to build a working database and application allowing queries into the dataset that would give statistical insights into the public knowledge of people from different countries or speakers of various languages. Student has also performed several analyses that were allowed by the data itself and demonstrated a potential of data mining methods.  
As also stated by the supervisor in his assessment, the point 7 of the assignment was not fully achieved. The reason for this is understandable as the application was allowed to work only autonomously without any interaction to the actual game. Student have, however, created a working application that can help the game creators to have easier insight into the data and trends that are appearing in the game.  
  
I am also missing more attention to the point 2 in the text of the work. Given the fact that a point 2 of the assignment says to get acquainted with statistical tests, I would expect an attempt to provide some information about the statistical significance of the results mined by the application. An ideal place would be trying to perform some statistical tests in the data presented in section 5.2, where several graphs show that some answers are significantly more prevalent for people from different regions.
3. **Length of technical report** **in usual extent**  
The report is of length. Report also contains number of graphs, schemes and tables that complement the text. More details related to the implementation are left for an appendix and do not clutter the core of the report.
4. **Presentation level of technical report** **75 p. (C)**  
All chapters are logically following each other and the text is very accessible to the reader. I would expect more attention to the theory of data mining and to questions of statistical significance of results. In my opinion the Chapter 1 (Introduction) should contain a deeper review of the problematic and not be concentrated on this single example related to the given dataset.
5. **Formal aspects of technical report** **95 p. (A)**  
I do not have any issues with typographical or stylistic side of the document. It is written in a good English without any obvious errors. All of the text is easily readable and understandable.
6. **Literature usage** **75 p. (C)**  
The problem itself did not require studying much of the literature and therefore the references are mainly online sources for the toolkit and libraries being used throughout the project. I do have, however, a comment about not diving more deeply into the statistical theory as it is also part of the assignment and I also find the use of ref [17] (Statistical Analysis Handbook) on page 12 unfortunate and little bit forced. It would be proper to reference it in some text dealing with a concrete statistical method and not generally at the word "statistics" as here.
7. **Implementation results** **85 p. (B)**  
The created application is easily usable and can provide quick insight into the game's data. It also allows easy export of the results in the form of graphs or tables. The design and source code is well documented.
8. **Utilizability of results**  
In my opinion, this is an engineering assignment, which required author to deal with a large dataset by building a SQL database and an application around it. Moreover, the resulting tools were used to provide some interesting insights into the data, which can be perhaps, in this stage, used by the authors of the game.
9. **Questions for defence**  
-
10. **Total assessment** **75 p. good (C)**  
I found the presented work to be nicely written and the application technically well executed. Student has also

managed to conduct several analyses with the dataset and pointed some challenges given the nature of the data, especially cross-lingual analysis, which is hard to perform here. Overall, I was satisfied, but I would appreciate more focus on the underlying theory of data mining both in theoretical review and in the conducted experiments. Given the relatively easier assignment, I suggest the mark C.

In Brno 3 June 2021

Píchoč Oldřich, Ing., Ph.D.  
reviewer