ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

# SUPERVISOR‘S OPINION OF FINAL THESIS

## I. IDENTIFICATION DATA

| | |
|---|---|
| **Thesis name:** | **Application of machine learning methods to solve the problem of user identification on various digital devices** |
| **Author’s name:** | **Elena Ivanova** |
| **Type of thesis:** | master |
| **Faculty/Institute:** | Faculty of Electrical Engineering (FEE) |
| **Department:** | Department of Computer Science |
| **Thesis supervisor:** | Karel Frajták |
| **Supervisor’s department:** | Department of Computer Science |

## II. EVALUATION OF INDIVIDUAL CRITERIA

**Assignment** **challenging**

*Evaluation of thesis difficulty of assignment.*

The task of recognizing one user across multiple devices became extremely important nowadays as everyone owns different devices to perform tasks and user's identity becomes fragmented which isn't good for advertising. This thesis is focusing on various machine learning techniques that could be applied to link computers and mobile devices that belong to one person. The problem was presented on Kaggle website providing an anonymized data and the task was to „identify individual users across their digital devices“ (quoting Kaggle).

The diploma's main goal is to develop an algorithm which could link computers and mobile devices that belong to the same person. We are provided with anonymous data that include user's behavior on sites and mobile apps along with visited IP addresses. The thesis is focused on various machine learning techniques that could be applied to solve the problem.

**Satisfaction of assignment** **fulfilled**

*Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.*

The introduction to problem on Kaggle is described better than the author has described in the thesis and gave me better understanding of what the author is trying to solve. The introduction is quite short, and I would welcome it to be longer including an example of the input data, example of data coming from multiple devices but belonging to a single user and the basic way of how to stitch it together to identify the individual. What I am also missing is not the decision why the author decided to use machine learning techniques (high volume of data), but a more detailed description of how the data is classified.

The background chapter describes the formal definitions and methods that the author will use for experiment. There are few related works cited without any description about what the authors did and how it can help the author of the thesis.

Next chapter describes the experimental part. At the beginning the author divides the obtained set into two parts (85% and 15%) without any explanation of why those numbers were used or why the smaller part was chosen as validation set or how the validation set was created. The preprocessing part is well written, describing how the author dealt with data being split into multiple data sources and how it was joined together. The results of the experiment show that the final score was able to make top 5 results in Kaggle (but the author have not submitted it).

**Technical level** **A - excellent.**

*Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.*

The author proved her knowledge of machine learning methods and applied it to solve such difficult task.

**Formal and language level, scope of thesis** **B - very good.**

*Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.*

Although the thesis is 75 pages long, the actual text ends on page 40 followed by 2 pages of references, and the rest is the content of the Python source files. I don’t think it’s necessary to put these into the printed version of the thesis (think of the paper required to print it). The text is well written, well styled and quite well organized (although the chapter numbering skipped Chapter 2 and 4). Language level is above normal, I haven’t notices and typos or serious grammar error.

| **Selection of sources, citation correctness** | **B - very good.** |
|---|---|
| *Present your opinion to student’s activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.* | |
| All resources were properly cited. Author researched quite recent related work (although 2015 is a far past in this field). | |

## III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

*Questions for defense:*

- *Why didn’t you post your results to Kaggle?*
- *You are filtering and joining the data based on cookies and their IP addresses, the Kaggle problem description (see below) indicates that the user is visiting multiple websites with different IP addresses, but what is unifying the user is the interest (trip to Iceland). How would you approach the problem from this point of view (this might have been addressed in the text, the cookies have some anonymous features to describe them, but not explained thoroughly)?*

*Kaggle problem description:*

*Imagine you're planning a summer holiday to Iceland: you read a travel blog on your smartphone on the subway to work, search for hotels on your laptop during lunch, browse Reykjavik restaurants on a tablet while half-watching TV after dinner, and then download a travel book to your e-reader to skim before bed.*

I evaluate handed thesis with classification grade **B - very good.**

Date: **1.6.2018**

Signature: Frejlák