Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220598

Jupyter Notebooks for Introducing Data Science to Novice Users

Mathilde FRUCHART[a,1] , Benjamin GUINHOUYA[a,b], Sylvia PELAYO[a,b], Christian VILHELM[a,b] and Antoine LAMER[a,b]
[a]*Univ. Lille, Faculté Ingénierie et Management de la Santé, F-59000, Lille, France.*
[b] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000 Lille, France.*

Abstract. Data science is a bridge discipline involving computer science, statistics, and knowledge of the health field. We developed a Jupyter Notebook to enable novice users to easily and autonomously analyze data from social networks. We conducted an experimentation with non-programmer students. They had to adapt a R Notebook and complete 14 questions and to perform descriptive analyses. The average rate of correct answers was 90.7. Jupyter Notebook enabled novice users to easily and autonomously analyze data from Twitter.

Keywords. Data Science; Education; Social media; Programming

1. Introduction

Data science may seem to be restricted to expert users because it requires the mastery of many technologies and knowledges [1]. Analyses are generally programmed from scripts consisting of a series of instructions, accompanied by comments, and developed with an integrated development environment (IDE) or in command line. Notebooks are another approach to programming and sharing results and were designed to improve the reproducibility of the studies [2]. The final notebook provides a full and comprehensive analysis report. We developed a R notebook to enable novice users to easily and autonomously analyze data from social networks. We assessed their ability to use it and captured their feedback on the value of such a tool for their daily practice.

2. Methods

Jupyter Notebook is an interactive computational laboratory notebook, which can work with code in many different programming languages such as Python, Java, R, or Julia [3]. Jupyter Notebook allows for the smooth integration of code and narrative text (in Markdown syntax) into a single document that can be executed and edited on the fly. We experimented the Jupyter Notebook with students from two master's degree formations not specialized in Data Science and without programming skills.

[1] Corresponding Author, Mathilde Fruchart, ULR 2694, Lille University, 2 place de Verdun, F-59000, Lille, France; E-mail: mathilde.fruchart@univ-lille.fr.

Wedistributed a R Notebook with datasets from Twitter. Students had to execute the Notebook and to adapt the script in order to obtain additional results.

We assessed the ability of the students to use the Jupyter Notebook in checking if they were able to perform the 14 tasks [4]: (i) instructions in relation to the notebook (modify and complete cells, execute cells, create new cells, export in pdf); (ii) R code 1st level (from an existing code, compute a new variable, change a color in a graphic, change a variable in a graphic); (iii) R code 2st level (find a instruction already implemented and modify it to produce a new result).

3. Results

We have delivered 20 health topics to 44 students. For each topic, we extracted 1000 tweets. Average (SD) rate of correct answers was 90.7 (13.3) for all the questions. Questions with the poorest scores were about (i) the modification of the required dataframe and the variable for the generation of a barplot, (ii) the updating of an instruction to display mentions instead of hastags and the adaptation of the title of the graphic, (iii) the modification of cells in Markdown format or (iv) the creation of two new cells for code or Mardown. The main difficulties that emerged were ensuring that users executed all the code cells, and that they had set up their cells in Markdown or code according to their needs. On the contrary, we did not report any major error on syntax problems which we usually encounter when learning a new language.

4. Discussion and conclusion

We developed a R notebook to enable novice users to easily and autonomously analyze data from Twitter. We tried it out with a class of 44 students without programming skills. Compared to Rmarkdown, which also integrate code and narrative text, Jupyter Notebook prevents the need to work locally and to have to install the necessary libraries himself [5]. IDE are rather designed for more advenced developers because they provide interesting features, such as code autocompletion, and displays the environment with all variables. However they could be more complex to handle for novice users than Jupyter notebook. Improvements would be to not display warnings and messages to provide training on manipulation of the notebook.

References

[1] Parker MS, Burgess AE, Bourne PE, Ten simple rules for starting (and sustaining) an academic data science initiative, PLoS Comput. Biol., vol. 17, no 2, p. e1008628, févr. 2021.
[2] Kluyver T et al. Jupyter Notebooks – a publishing format for reproducible computational workflows, Position. Power Acad. Publ. Play. Agents Agendas, p. 87-90, 2016.
[3] Project Jupyter. https://jupyter.org.
[4] Mathilde Fruchart / Twitter analysis for novice users, GitLab. https://gitlab.com/mathilde.frchrt/twitter-analysis-for-novice-users.
[5] Baumer B, Cetinkaya-Rundel M, Bray A, Loi L, Horton NJ. R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics, ArXiv14021894 Stat, févr. 2014.