Contribution ID: **57** Type: **Oral Presentation**

Using Natural Language Processing to Extract Information from Unstructured code-change version control data: lessons learned

Friday, March 26, 2021 10:00 AM (30 minutes)

Natural Language Processing (NLP) is a branch of artificial intelligence that extracts information from language. It has been applied to a wide range of fields [1], such as voice recognition, email categorization, social network analysis and others. In the field of software engineering, NLP has been adopted to extract key information from free-form text, to generate models from the analysis of text or to categorize code changes according to their commit messages [2].

In the present study, NLP has been applied to software code and its documentation in order to detect any defects that may be omitted by other techniques. NLP has helped in the identification of patterns of software code modifications [3, 4], to determine how it has evolved over time and to understand the complexity faced by developers (who have implemented new features and fixed faults).

The wide adoption of version control systems, such as github, to manage software code opened the access to different unstructured information. Each modification entry, in fact, is characterized by specific elements such as the creation date of a file, the date of the modification, the name of the developer who performed the change, the lines of code that were added or removed, and a message that explains the reasons for the change. According to the content of the message, it is possible to identify key terms that can be used during the classification of the entries.

Willing to define useful instruments to support developers in their daily activities, in this study we present its key terms and the lessons learned from the assessment of the change history of WLCG software available on github to the HEP community [5]. Adopting NLP techniques, we have cleaned the messages and extracted some key terms to categorize software problems and other code changes performed by developers like the integration of a third-party dependency or the control sequence introduced for a given service. By combining the information gathered from the WLCG software and its change history present in the github repository [5] with the one in already existing literature [6] we have built a code change dictionary. Finally, we have applied some Machine Learning (ML) techniques to investigate possible, unconventional connections between code changes and software defects.

References

[1] S. Falkenstine, A. Thornton, B. Meiners, “Natural Language Processing for Autonomous Identification of Impactful Changes to Specification Documents”, in Proceedings of the 2020 AIAA/IEEE 39th Digital Avionics System Conference (DASC), doi: 10.1109/DASC50938.2020.9256611

[2] F. Gilson and D. Weyns, “When Natural Language Processing Jumps into Collaborative Software Engineering”, in Proceedings of the 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), 2019. doi: 10.1109/ICSA-C.2019.00049.

[3] C. Rosen, B. Grawi, and E. Shihab, “Commit guru: analytics and risk prediction of software commits”, in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, pp. 966-969, ACM, 2015. doi: 10.1145/2786805.2803183

[4] S. Majumder, J. Chakraborty, A. Agrawal, and T. Menzies, “Why Software Projects need Heroes (Lessons Learned from 1100+ Projects”, in cs.SE, 2020. arXiv:1904.09954v2

[5] E. Ronchieri, Y. Yang, M. Canaparo, A. Costantini, D. C. Duma, D. Salomoni, “A new code change prediction dataset: a case study based on HEP software”, under publication in Proceedings of IEEE NSS MIC 2020

[6] R. Ferenc, P. Gyimesi, G. Gyimesi, Z. Toth and T. Gyimothy, “An automatically created novel bug dataset and its validation in bug prediction”, in The Journal of Systems & Software, vol. 169, 2020. doi: 10.1016/j.jss.2020.110691

Primary authors: Dr COSTANTINI, Alessandro (INFN-CNAF); Dr SALOMONI, Davide (INFN); DUMA, Doina Cristina (INFN - CNAF); Dr RONCHIERI, Elisabetta (INFN CNAF); Mr CANAPARO, Marco (INFN); YANG,

Yue (Department of Statistical Sciences, University of Bologna)

Presenter: Dr RONCHIERI, Elisabetta (INFN CNAF)

Session Classification: Physics & Engineering Session

Track Classification: Physics (including HEP) and Engineering Applications