Code Similarity Analysis

Yunze Hao

***ABSTRACT*—This project analyze the similarities among codes, which were developed for similar goals. Analyzing mainly relies on Universal sentence encoder method after deleting notations or comments of codes.**

I. INTRODUCTION

All codes were written by Python and a template was provided for all coders to implement different functions.

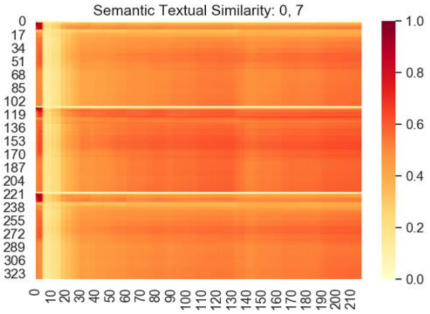
II. Methods

Preprocessing: Codes belong to one author were integrated into one file. Comments and notations were deleted, since they tends to be similar for the usage of the same template.

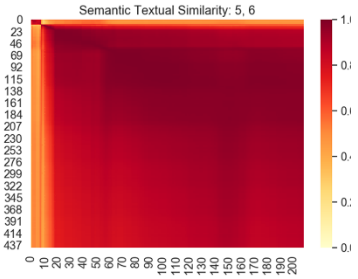
Analysis: Universal sentence encoder can detect different sentences representing similar meaning. Thus it was chosen as the method to analyzed similarities.

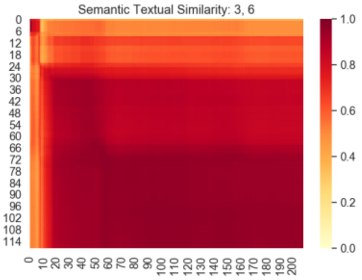
III. Results and discussion

There are eight codes writers. The similarities were compared among each writer's codes. The ID7 writer's codes are much different from others.



Comparing ID7's codes to other's codes, low similarities were found for each pairs. Looking more closely into the codes, the different approach of building protected health information dataset was found, supporting our analysis result. High similarities among ID3, 5, 6 were detected. After looking closely into codes, the similarities were probably due to the abuse of the template.





IV. Conclusion

The popularity of similar methods contribute to the similarities among codes. Besides, the abuse of the template also contribute to the similarities.

Reference:

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder. arXiv:1803.11175, 2018.