University of Victoria | Graduate Studies

Notice of the Final Oral Examination
for the Degree of Master of Science

of

KAMEL ALRASHEDY

BEd (University of Hail, 2013)

"Predicting the Programming Language of Questions and Snippets of Stack Overflow Using Natural Language Processing"

Department of Computer Science

Wednesday, August 29, 2018
11:00 A.M.
Engineering and Computer Science Building
Room 468

Supervisory Committee:
Dr. Venkatesh Srinivasan, Department of Computer Science, University of Victoria (Co-Supervisor)
Dr. T. Aaron Gulliver, Department of Electrical and Computer Engineering, UVic (Co-Supervisor)

External Examiner:
Dr. Alex Thomo, Department of Computer Science, UVic

Chair of Oral Examination:
Dr. Meyer Horowitz, School of Exercise, Science, Physical & Health Education, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

Stack Overflow is the most popular Q&A website among software developers. As a platform for knowledge sharing and acquisition, the questions posted in Stack Overflow usually contain a code snippet. Stack Overflow relies on users to properly tag the programming language of a question and it simply assumes that the programming language of the snippets inside a question is the same as the tag of the question itself. In this thesis, a classifier is proposed to predict the programming language of questions posted in Stack Overflow using Natural Language Processing (NLP) and Machine Learning (ML). The classifier achieves an accuracy of 91.1% in predicting the 24 most popular programming languages by combining features from the title, body and code snippets of the question. We also propose a classifier that only uses the title and body of the question and has an accuracy of 81.1%. Finally, we propose a classifier of code snippets only that achieves an accuracy of 77.7%. These results show that deploying ML techniques on the combination of text and code snippets of a question provides the best performance. These results demonstrate that it is possible to identify the programming language of a snippet of only a few lines of source code. We visualize the feature space of two programming languages Java and SQL in order to identify some properties of the information inside the questions corresponding to these languages.