Title: Generalized Language Identification

Speaker: Marco Lui
University of Melbourne

Date/Time: 11 April 2014, Friday, 02:00 PM to 03:00 PM

Venue: MR6, AS6-05-10

Chaired by: Dr Ng Hwee Tou, Provost's Chair Professor, School of Computing (nght@comp.nus.edu.sg)

ABSTRACT:

Language identification (LangID) is the task of determining the natural language a document is written in. When I visited NUS in 2012, I gave a talk introducing standard methods for LangID, and presented our work on building langid.py, a language identifier that is robust to variation in language across different sources of text. In this talk, I will briefly recap our previous work, and present our new work on language identification in multilingual documents, i.e. documents that contain text in more than one language. I will also present some results on research in discriminating similar languages, specifically on discriminating between Australian, British and Canadian varieties of English.

BIODATA:

Marco Lui is a PhD candidate in the Department of Computing and Information Systems at the University of Melbourne in Australia, and is in the final stages of completing his thesis. His main research area is language identification, and his research interests include text classification, natural language processing, machine learning, web-scale data and user-generated content. He grew up in Singapore, and speaks English, Italian, Cantonese and Mandarin.