CLASSIFICATION, CONTEXT EXTRACTION AND OTHER TECHNIQUES FOR EMAIL MANAGEMENT

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2012

By
Romain Francez
School of Computer Science

Abstract

Emails are time consuming, mainly because an average user receive and send a lot of messages every day. This master's dissertation tries to solve some of the problems encountered with email management by using text mining techniques. Information retrieval, clustering, classification and named entity recognition are put together to improve email usage. This document shows results for these techniques and explains why it is more complicated to have good performances for unsupervised techniques rather than for supervised ones. Spam detection is incredibly accurate (with more than 99% of accuracy), named entity recognition did well finding basic entities like persons, locations and dates. Automatic labelling was found to be more challenging.

Keywords: emails, spam, information retrieval, information extraction, clustering, classification, named entity recognition