DATA MINING PROJECT

CSE4412 FALL 2012

[Pick the date]

Siwei Wang

DEREK XU

Contents

[Introduction 3](#_Toc341360794)

# Introduction

Internet and e mail become very popular, and also the spam or junk email become so annoying and are affecting almost everyone’s daily life.

Many people are working on to find an efficient way to filter out the spam. The major task of filtering the spam is to classify the emails into spam and ham. Numbers of data mining (text mining) techniques are employed to help to solve this problem.

The project is to use set of training data, which is set of emails text contains both spam and ham, to generate a model and then use that model to classify the test data set. The most important steps are preprocessing the training data and deciding which classification algorithm to use.