Cpts 315

Project Proposal

Team member: Qixuan Jiang(11677932), Hanzhi Zheng(11648592)

3/21/2022

Proposal of SMS Spam Classification

1) Data mining task:

   With the advancement of technology, the types of fraudulent information have become diverse and difficult to distinguish. According to Stats, nearly 122.23 billion spam emails are sent every day. At the same time, cross-checking with the total number of emails sent in a day, the count of spam emails is huge. Spam messages accompany our daily lives, although most of them are identifiable, a small part of fraudulent information can lead to poisoning of the user's device, resulting in the leakage of user information. Our task is purpose to help users determine which text message or email are spam risk by using data mining.

2) Dataset:

   We will use dataset from https://www.kaggle.com/uciml/sms-spam-collection-dataset. In this dataset, it includes spam and non-spam messages.

3) Methodology:

   We will use A-priori algorithm to run through every message, Look for text with potentially fraudulent sequence of words in each message, and then distinguish potentially dangerous messages from normal messages.

4) Final product:

We will learn how to use data mining and machine learning skills to determine spam messages and non-spam messages. In the future, this project can help many users classify their messages easily, decrease rate of being scammed.