Data Mining Hw1

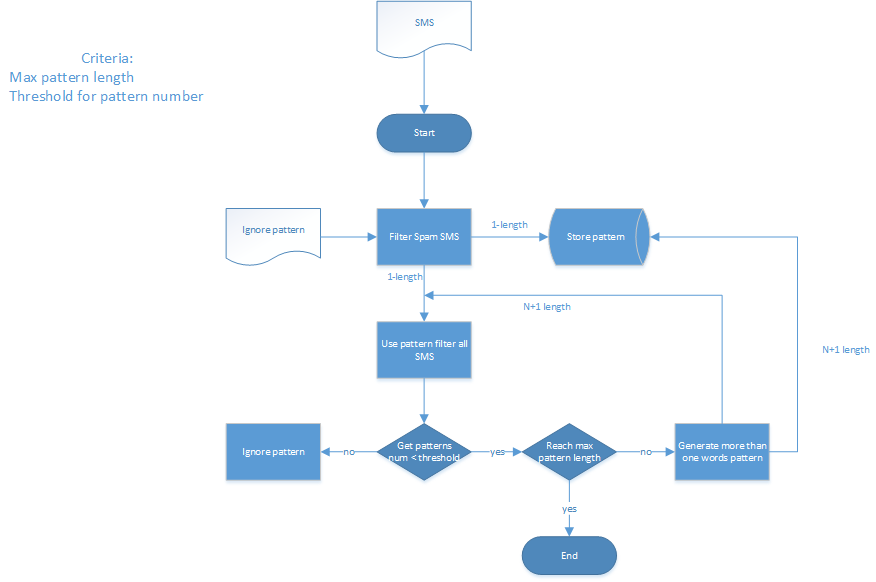
資工所 r02922103 張家瑋

資工所 r02922082 林士淳

1. Introduction：

We live in an information explosion generation, which everyone receive thousands of messages for a day. Through many different technologies, for instance, e-mail, SMS, MMS, instant message like “Line, Skype, Hangouts, etc.” It is convenient to communicate or trace with friends. Nonetheless, there are a large part of receiving messages belong useless messages. How to filter these useless messages and take others for users is significant to every service providers.

In this homework, we try to design a program and use some SMS messages to train them, hoping our program can differentiate spam from all of the messages by searching some patterns of useless messages.

 Below is the flow chart of our program:

II. preprocessing tactics

In the beginning of this program, we input some normal SMS messages and some spam SMS messages. According to the first words of each messages(spam or ham), we can branch them to two types of messages. By filtering all spam SMS messages, the program records some spam patterns and use it to filter all messages including normal and spam. After this step, the program chooses the most useful patterns and uses it to generate longer patterns, repeating the steps until finding enough patterns.

III. method & settings

(i). Variable Setting:

1. V is the value of the pattern.

2. ,: The number of the pattern contains in spam SMS and normal SMS message, respectively.

3. , : The number of all spam SMS and normal SMS.

4. : The factor to adjust how serious to is taking normal SMS to spam class

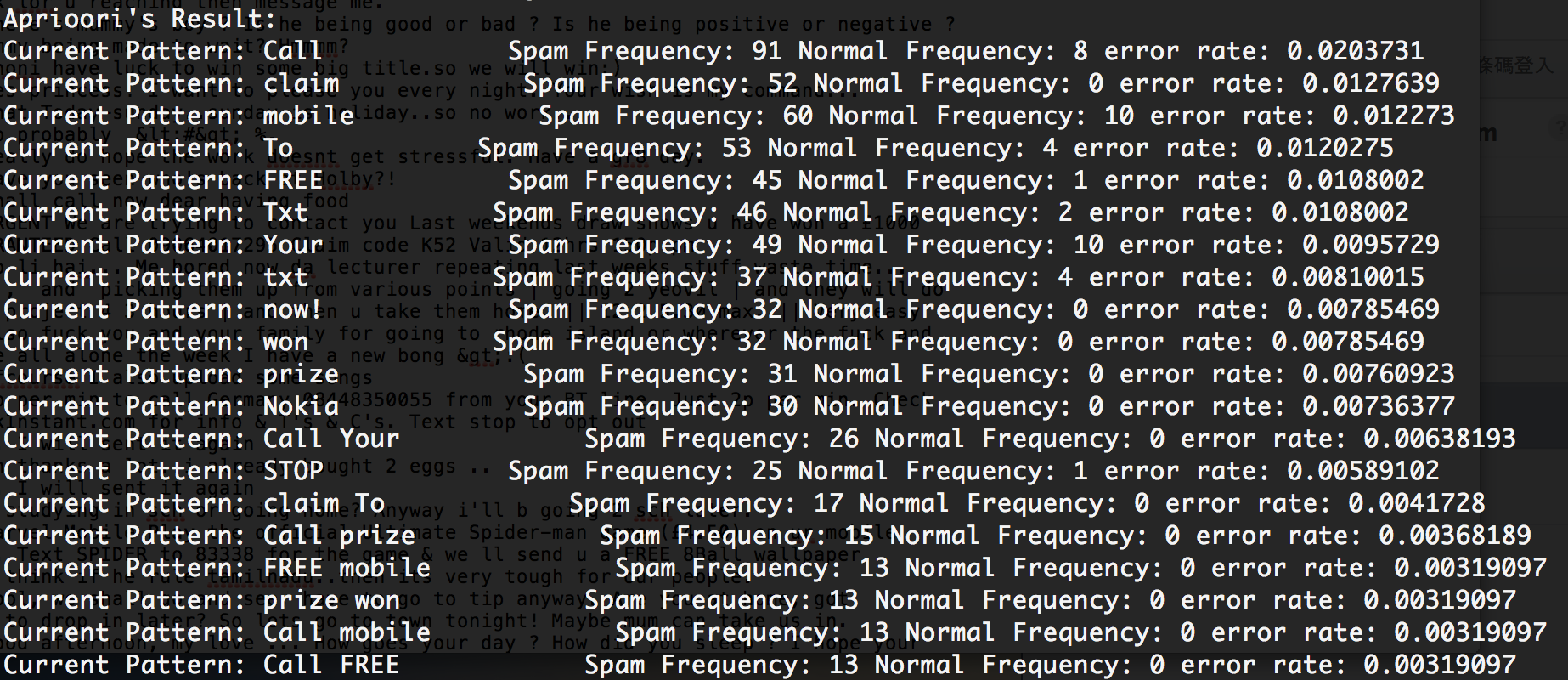
The criteria to calculate the value about a pattern:

V = ,

(ii). Method:

We use Apriori Algorithm and Sampling approach. The above mentioned method(preprocessing tactics) is Apriori’s preprocessing.

The result is like this:



We can see that there are some patterns having high normal frequency, ex: “Your”, “mobile”, both of them have normal frequency 10. As a result, we tried to use sample method, choosing only 100 spam messages to generate 20 patterns.

IV. results

After using the sample method, we get the better result:

