Assignment Classification: Spam Prediction

Toon Calders

Deadline: May 1st

1 Context

You are working in the research and development department of a company. One day you are asked to research the possibility of developing a plugin for the company's mail system for detecting spam emails. The background of this request is the observation that, despite the use of a spam filter at the company level, the employees of your company still receive a lot of spam emails. Your first step is to discuss the issue with the administrator of the email server, who explains to you that for the company it is absolutely essential not to lose any legitimate email. Therefore, the spam filter's settings have been configured to be very conservative and only exclude emails that are spam emails beyond any doubt.

2 Project Description

Based on the input of the administrator, you come up with the plan of creating a company-specific spam filter. This spam filter will predict for each mail whether the mail is legitimate, suspicious, or spam. Spam emails will be removed, legitimate emails added into the inbox of the employee, and suspicious emails to a dedicated folder in the employee's mailbox. You present this plan to your manager, who agrees that you start a project to evaluate the potential of such an approach, researching the following questions:

- How can a company-specific spam filter be developed?
- What performance can be expected of the spam filter?
- Which features in an email are the most important ones for deciding if an email is spam?
- (Extra) Is it helpful to personalize the spam filter? That is, is it possible to improve the quality of the spam filter by personalizing it based on the contents of an employee's inbox?

For the final evaluation of the project and to make a go/no go decision, it is agreed with management to use the following cost estimations:

- Classifying a legitimate email as spam has an estimated cost of 0.5h of work to cover up for missed business opportunities and image damage. The average cost per working hour is 40 Euros.
- Classifying a legitimate email as suspicious is estimated to cost 5 Euros as this may imply a late answer or a missed business opportunity.
- On average every spam email delivered in the inbox or suspicious folder of an employee costs 40 cents.

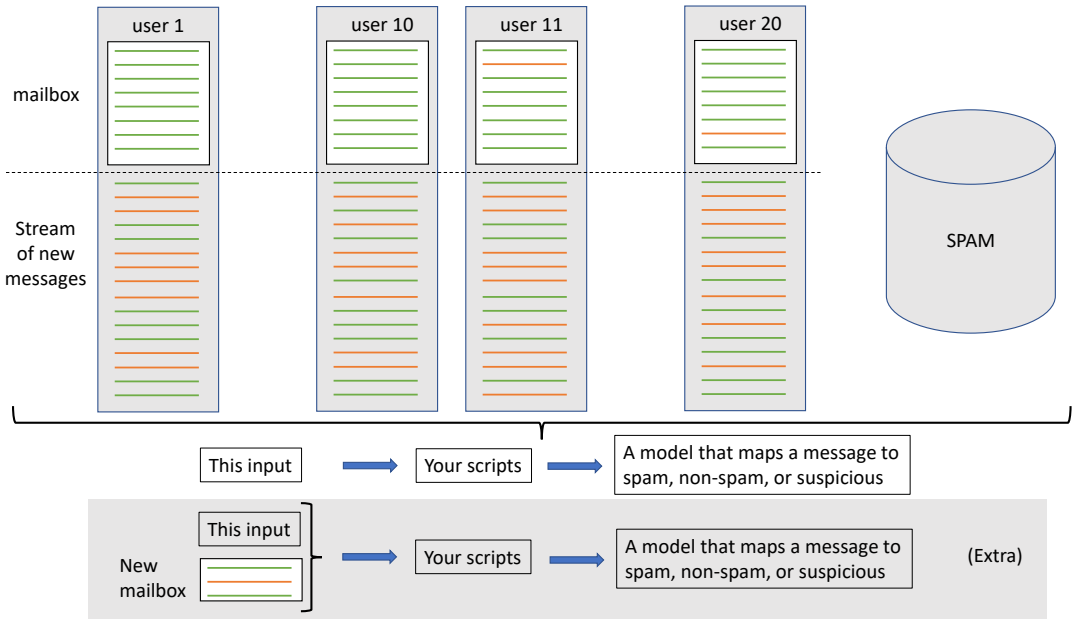

Figure 1: Schematic overview of the data available for the project

3 Project Execution

3.1 Project Start

To start your research you have to collect relevant data. After sending out a company-wide request, several employees agree that the contents of their inboxes will be used in this research. With the help of the system administrator you gather the following datasets:

- 10 employees are paid to meticulously go through their mailboxes and remove any spam message from it. In this way you get 10 mailboxes containing data which is guaranteed to contain only spam-free emails.
- Next to these 10 employees, 10 more employees donate the content of their inboxes, but due to budget restrictions it cannot be guaranteed that these mailboxes are completely spam-free. Nevertheless, it is reasonable to assume that less than 5% of these emails are spam.
- For all 20 employees, one week of mails is collected without any filtering. This set of mails is representative for the stream of emails your spam filter will get as input.
- Furthermore, a dataset was acquired externally consisting of spam emails only.

All mails were preprocessed to remove headers, punctuation and some information from the receiver. In Appendix A, examples of a spam and non-spam mail are given.

A schematic view of the available data is given in Figure 1.

3.2 Project Continuation

At this point your assignment starts ... The data above is made available to you in a single zip file on the course website, containing the following folders:

- `user1-user10` : these folders contain subfolders `inbox` and `stream` containing respectively the cleaned mailboxes (of the 10 employees who were paid to clean out spam emails), and the stream of messages received by the user in one week. This stream is unlabeled and contains a mixture of spam and non-spam emails.
- `user11-user20` : these folders contain the data of 10 more users and has the same format at the for the clean mailbox users, with the difference that the inboxes of these 10 users may contain up to 5% spam.
- `spam` : this folder contains numerous spam messages gathered from an external source. They are nevertheless representative for the spam emails received by the employees.

In Appendix B, you find some suggestions for tools to use to proceed with your assignment.

4 Evaluation

The result of your project will consists of:

- A report regarding the research questions listed above. The report should contain a description of a solution and sufficient data to support your claims. Make sure to explicitly mention what performance you expect from your approach, and whether you recommend a go or no-go for the implementation of the project. The report should also succinctly describe weaknesses of your approach as well as potential threats to the validity of the results.
- The code to train and use a model to classify new emails into inbox, suspicious, or trash. Make sure your code is sufficiently documented and can be used easily to train the models you use on new data, and make predictions on new data.
- (Extra) A prototype to train a personalized spam filter for a new employee. Make sure the prototype is sufficiently documented to be used easily. Notice that it is perfectly allowed to use the datasets listed above to build the model in the prototype.

A Email examples

A.1 Spam Email

```
Subject: take the reins
become
your employer .
substantial profit processing money judgments .
from anywhere .
control when you want to work .
a substantial number of our members earn 5 , 000 us to 12 , 000 us per mo .
outstanding customer support and assistance .
here for more
info
while the couple were apparently examining the strange device , rob
started to his feet and walked toward them
the crowd fell back at his approach , but the man and the girl were so
interested that they did not notice himhe was still several paces away when
the girl put out her finger and touched the indicator on the dial
discontinue orange stad , and then mail stop 1 . 200 b , followed by a rub
a
to rob ' s horror and consternation the big turk began to rise slowly into
the air , while a howl of fear burst from the crowdbut the boy made a mighty
spring and caught the turk by his foot , clinging to it with desperate
tenacity , while they both mounted steadily upward until they were far above
the city of the desert
the big turk screamed pitifully at first , and then actually fainted away
from frightrob was much frightened , on his part , for he knew if his hands
slipped from their hold he would fall to his death
```

A.2 Non-spam Email

```
subject : february production estimate
tom ,
see the attached file listing by trade zone by meter producer services '
wellhead production estimate for the month of february , 2000 . please be
advised that this is a preliminary estimate ! ! ! as i am currently in the
process of collecting noms from a couple of producers . i will alert you of
any revisions as they arise . additionally , i have highlighted those deals
that we expect to come on in february , at such time will we request tickets
to be entered into sitara .
should you have any questions , please give me a call .
thanks ,
vlt
```

B Suggested Tools

My personal language of preference for data science tasks is Python, using the `pandas` and `sklearn` libraries. For text processing `sklearn.feature_extraction.text` contains many handy tools, including the `CountVectorizer`. Some basic information and scripts for spam mail prediction can be found at `https://www.kdnuggets.com/2017/03/email-spam-filtering-an-implementation-with-python-and-scikit-learn.html`. See also `http://zacstewart.com/2015/04/28/document-classification-with-scikit-learn.html`.

You are allowed to use other tools as well; last year many students were quite happy about the tool `knime`, which, according to them, was easy to use and worked out-of-the-box. You can find more information on knime at `https://www.knime.com/` and you may want to inspect `https://www.knime.com/knime-text-processing`.