CASE 5

Group 3： Huanhan Liu, Fangzheng Sun, Neel Gehlot, Yun Yue

PART 1

What do we think

Definition:

- One of the largest frustrations for Internet users.
- For businesses, this frustration adds up to dollars lost and spent trying to prevent it

Spam
/spam/ noun

Spam is an unsolicited email message, instant message, or text message – usually sent to the recipient for commercial purposes

Dangerous email

WPI Mail Admin

AA

Aihaitijiang, Abudula

今天, 10:15

全部答复

WPI

Final Notice, upgrade your WPI.EDU email to office 2017 server for better performance and more storage space, CLICK HERE and update. Failure to follow this instruction will lead to permanent deactivation of your mail box in the next 24 hours.

Worcester Polytechnic Institute | 100 Institute Road | Worcester, MA

The Effects

1. Spam contributes to a loss of productivity and profit.

2. Spam poses legal risks

3. Spam contains various malware threats

4. Spam can also hurt the reputation of your business

Business Problem

To Filter Junk/Spam Emails more efficiently using Machine Learning.

We are going to use sklearn package to do email classification:
ham (non-junk mail)and spam(junk mail)

Potential Clients

Gmail by Google

Outlook

YAHOO! MAIL

Hotmail

PART 2

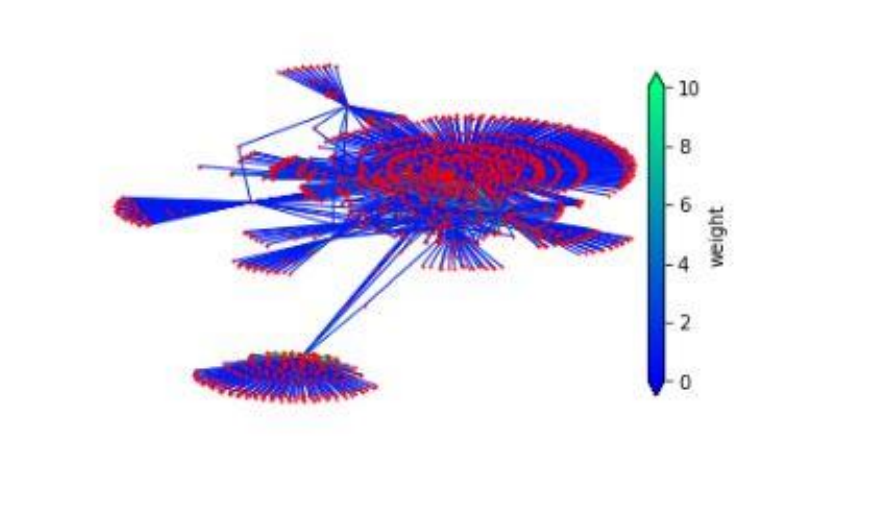
Data Collection

Data Exploration

Exploring the Email Dataset

Plot Email Communication Graph/Network

- Each node is an email account
- Take totally 10812 emails
- Color represents the weight of each edge
- The weight of an edge between two accounts depends on how many emails have been sent between them.



Data Collection

- Loading raw email data into a workable format
- The Enron Email Dataset
- Used Inbox and Deleted folder of all the users

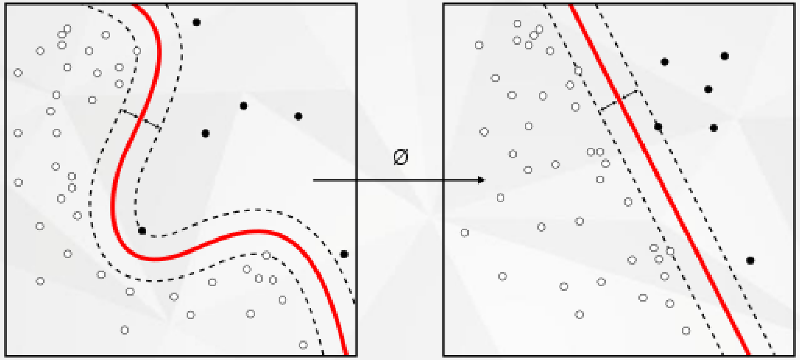
PART 3

Methods

CountVectorizer

- It learns the vocabulary of the corpus and extracts word count features.
- This method is an efficient way to do both steps, and for us it does the job.
- CountVectorizer provides fit and transform methods to do them separately.
- Additionally, you can provide a vocabulary in the constructor.

Naïve Bayes



- We're going to use a naïve Bayes classifier to learn from the features.
- A naïve Bayes classifier applies the Bayes theorem with naïve independence assumptions.
- Each feature is independent from every other one and each one contributes to the probability that an example belongs to a particular class.

PART 4

Data Processing

```
# of users we considered:  149
# of mails in ham set:  44542
# of mails in spam set:  50941
```

	class	text
C:/Users/sun_f_000/Documents/maildir\steffes-j\deleted_items\583	spam	I was thinking about what to get Dad for Chris...
C:/Users/sun_f_000/Documents/maildir\steffes-j\deleted_items\40	spam	\n\nDear Customer,\n\nThe electric utility i...
C:/Users/sun_f_000/Documents/maildir\shackleton-s\deleted_items\299	spam	\n\n[IMAGE] Forums Discuss these points in the...
C:/Users/sun_f_000/Documents/maildir\heard-m\inbox\64	ham	Sara,\n\nGSI Give up agreement. We would want...
C:/Users/sun_f_000/Documents/maildir\kaminski-v\deleted_items\2122	spam	\n\n From the Desk of George W. Pratt, III, Di...
C:/Users/sun_f_000/Documents/maildir\nemec-g\inbox\64	ham	Gerald,\n\n\n\nThis CA is to cover a proposed ...
C:/Users/sun_f_000/Documents/maildir\skilling-j\inbox\1386	ham	\n\n\n\n____...
C:/Users/sun_f_000/Documents/maildir\benson-r\inbox\292	ham	\n\n\n\n -----Original Message-----\n\nFrom: \...
C:/Users/sun_f_000/Documents/maildir\heard-m\inbox\master_netting\109	ham	Marie,\n\n\n\nThanks for your response. Pleas...
C:/Users/sun_f_000/Documents/maildir\meyers-a\deleted_items\914	spam	\n\n\n\nStart Date: 1/5/02; HourAhead hour: 18...

- Store the text data in a pandas data frame with class label “spam” or “ham”
- Shuffle the rows of the data frame so that the dataset become random
- 44542 ham emails and 50941 spam emails are selected under 149 users

- Reduce the mass of unstructured data into some uniform set of attributes that an algorithm can learn from by vectorizing all mail text to a sparse matrix with the same row number as the data frame and large column number

- (In numerical analysis, a sparse matrix is a matrix in which most of the elements are zero)

- In our sparse matrix, each element is an integer from 0 to 10, representing the feature of a word in the text

- The sparse matrix is our predictors matrix.

- Train the dataset and try the following example

```
# here's one example of classification test after the training
examples = ['Free Viagra call today!', "Tomorrow's meeting canceled."]
example_counts = count_vectorizer.transform(examples)
predictions = classifier.predict(example_counts)
predictions
```

```
array(['spam', 'ham'],
      dtype='<U4')
```

- In this example, with our vectorizer classifier, the two sentences are accurately classified by computer

Classification Results

- Apply 6-flod cross validation
- The overall accuracy regarding to all users' mails is 0.633
- If we turn to the dataset of some single users, accuracies are larger, this will be explained in our data limitation slide

All users' classification result

```
Total emails classified: 95483
Score: 0.55118575603 5
Confusion matrix:
[[38868  5674]
 [29399 21542]]
Accuracy: 0.632678068347
```

Single users' classification result

```
Total emails classified: 427
Score: 0.873316711025
Confusion matrix:
[[ 12  54]
 [ 39 322]]
Accuracy: 0.782201405152
```

```
Total emails classified: 1253
Score: 0.319930746275
Confusion matrix:
[[1114   29]
 [  84   26]]
Accuracy: 0.909816440543
```

PART 5

Limitation

Limitation

1. Highly personalized

- People have different habits, some people just delete the advertisement emails and people's attitude toward junk mails are different.
- People could delete some emails by mistake.
- Some people even don't delete junk mails.

2. Overfitting

- After transferring the text into sparse matrix, the number of predictors is very large, even larger than the training size which will lead to overfitting.

PART 6

Conclusion

Conclusion

- We analyzed the email data and use machine learning to filter junk mails.
- Our analysis can help email users to have a cleaner and safer using environment.
- We can also use our methods to do more specific classification of all emails in the future to increase working efficiency.

Thank you!