Министерство образования Республики Беларусь

Учреждение образования

"Белорусский Государственный университет информатики

и радиоэлектроники"

Лабораторная работа №5

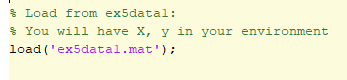
по учебной дисциплине “Машинное обучение”

|  |  |
| --- | --- |
| Выполнил: | Студент гр. 956241 Дубовик Н.О. |
|  |  |
| Проверил: | Стержанов М.В. |

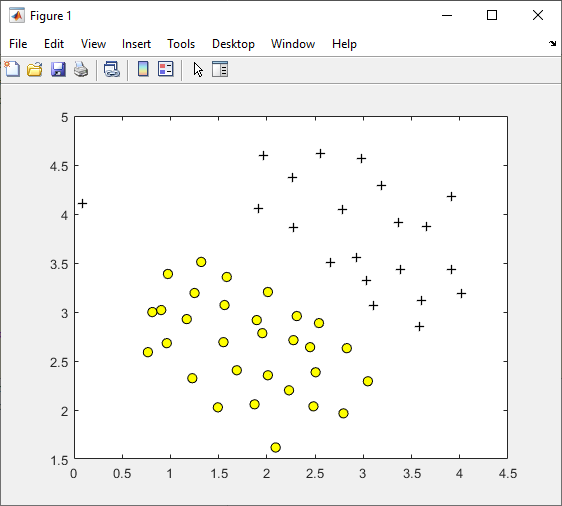
Минск 2019

**Задание.**

1. Загрузите данные ex5data1.mat из файла.



1. Постройте график для загруженного набора данных: по осям - переменные X1, X2, а точки, принадлежащие различным классам должны быть обозначены различными маркерами.

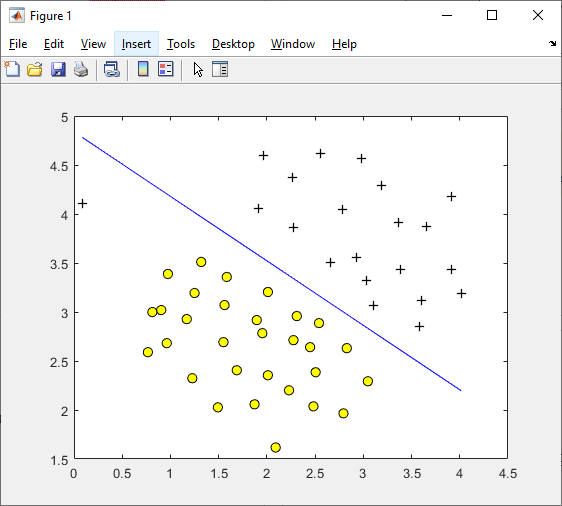


1. Обучите классификатор с помощью библиотечной реализации SVM с линейным ядром на данном наборе.

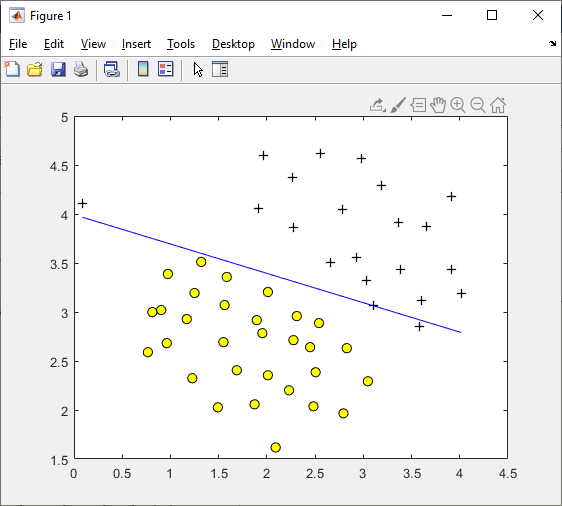


1. Постройте разделяющую прямую для классификаторов с различными параметрами C = 1, C = 100 (совместно с графиком из пункта 2). Объясните различия в полученных прямых?

C = 1

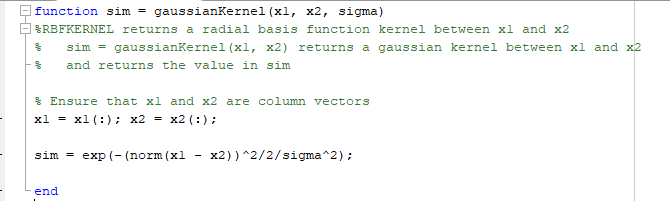


C = 100

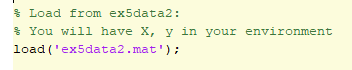


Параметр C является положительным значением, которое контролирует штраф за неверно классифицированные примеры обучения. Большой параметр C говорит SVM о необходимости правильно классифицировать все примеры.

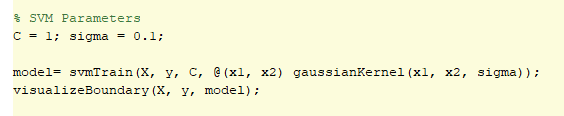
1. Реализуйте функцию вычисления Гауссового ядра для алгоритма SVM.



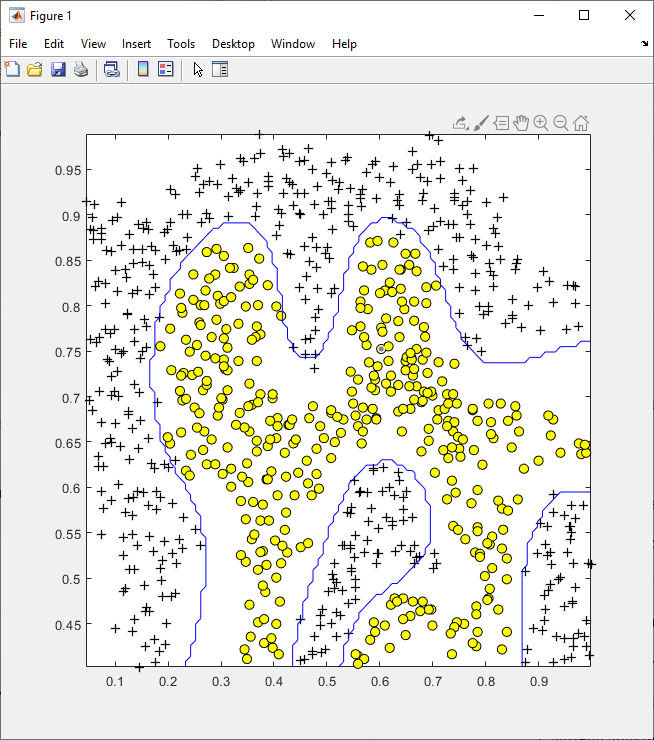
1. Загрузите данные ex5data2.mat из файла.



1. Обработайте данные с помощью функции Гауссового ядра.
2. Обучите классификатор SVM.



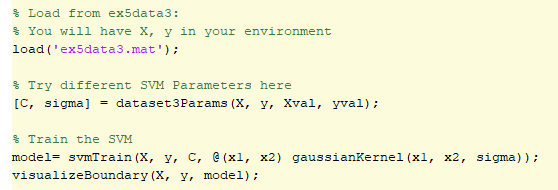
1. Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).

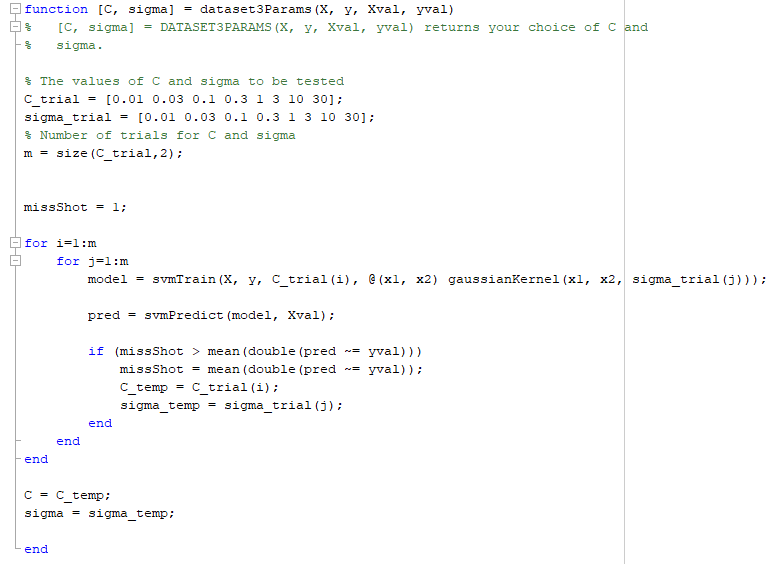


1. Загрузите данные ex5data3.mat из файла.

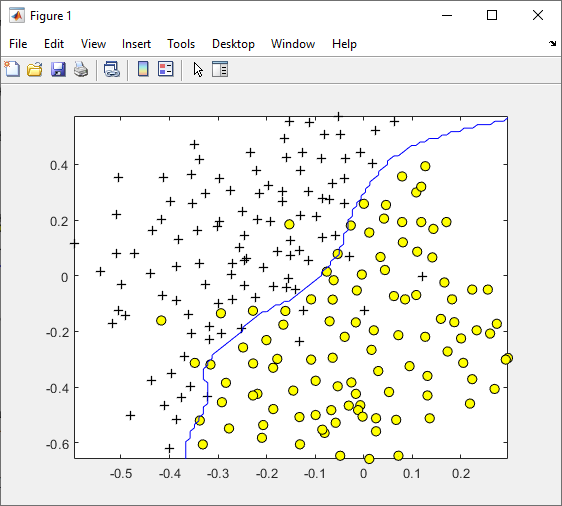


1. Вычислите параметры классификатора SVM на обучающей выборке, а также подберите параметры C и σ2 на валидационной выборке.

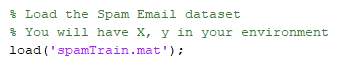




1. Визуализируйте данные вместе с разделяющей кривой (аналогично пункту 4).



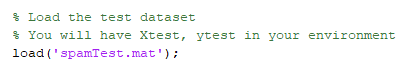
1. Загрузите данные spamTrain.mat из файла.



1. Обучите классификатор SVM.



1. Загрузите данные spamTest.mat из файла.



1. Реализуйте функцию предобработки текста письма, включающую в себя:

a. перевод в нижний регистр;

b. удаление HTML тэгов;

c. замена URL на одно слово (например, “httpaddr”);

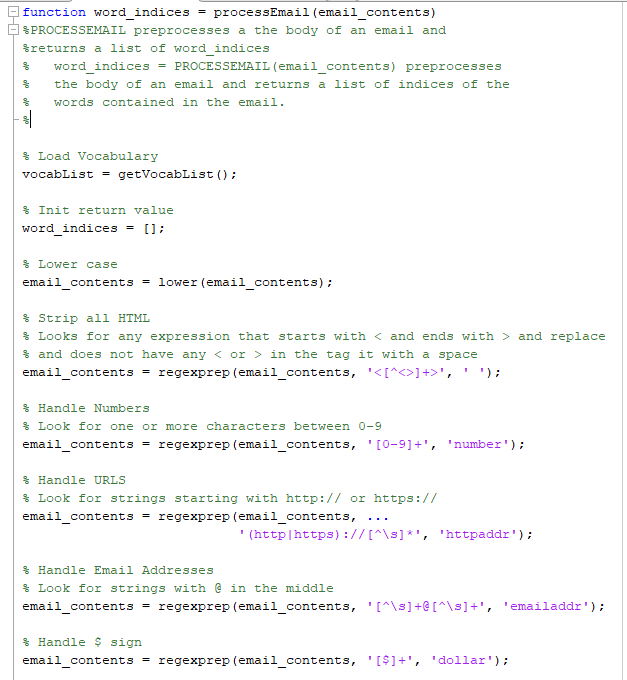
d. замена email-адресов на одно слово (например, “emailaddr”);

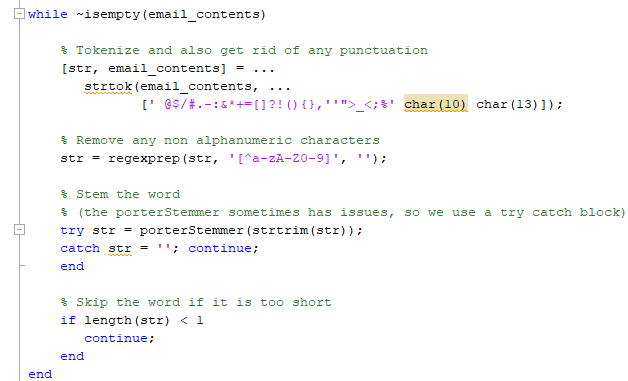
e. замена чисел на одно слово (например, “number”);

f. замена знаков доллара ($) на слово “dollar”;

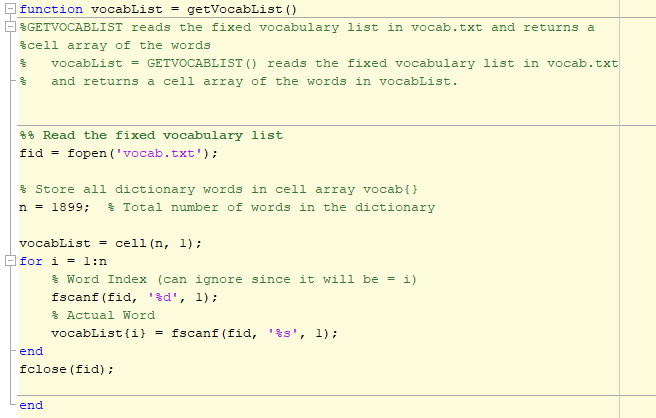
g. замена форм слов на исходное слово (например, слова “discount”, “discounts”, “discounted”, “discounting” должны быть заменены на слово “discount”). Такой подход называется stemming;

h. остальные символы должны быть удалены и заменены на пробелы, т.е. в результате получится текст, состоящий из слов, разделенных пробелами.



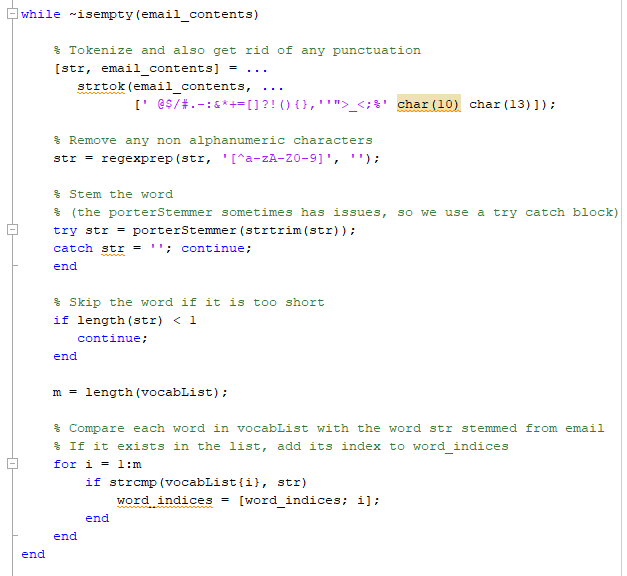


1. Загрузите коды слов из словаря vocab.txt.

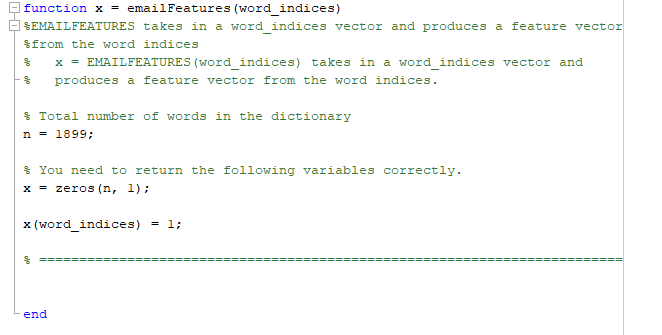


1. Реализуйте функцию замены слов в тексте письма после предобработки на их соответствующие коды.

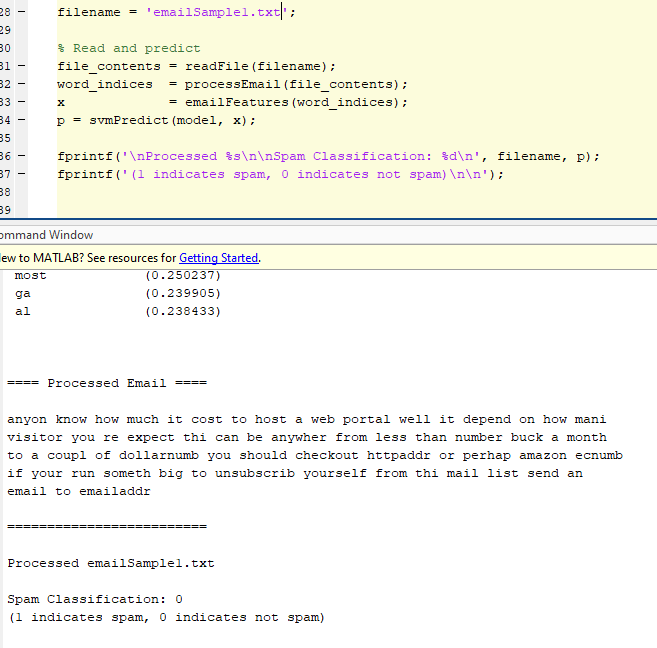


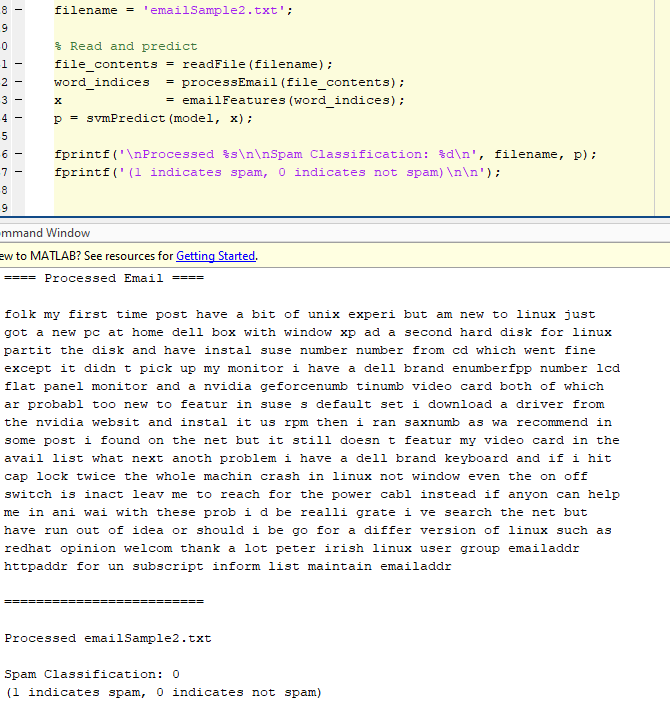


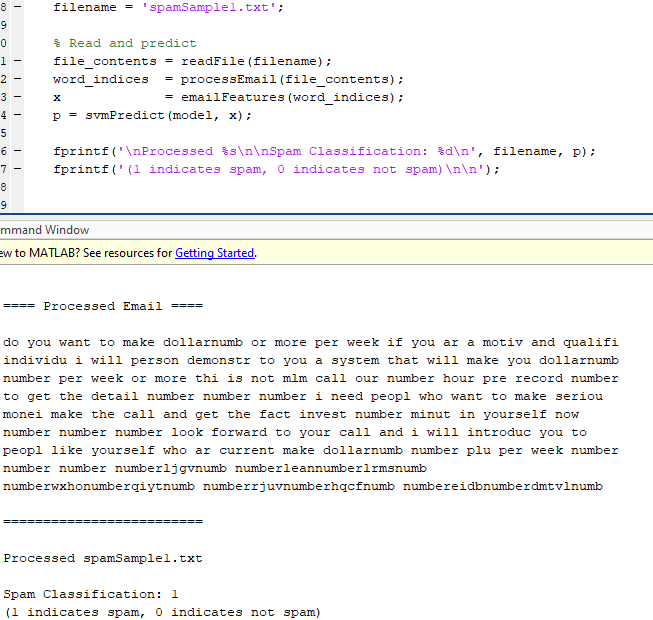
1. Реализуйте функцию преобразования текста письма в вектор признаков (в таком же формате как в файлах spamTrain.mat и spamTest.mat).

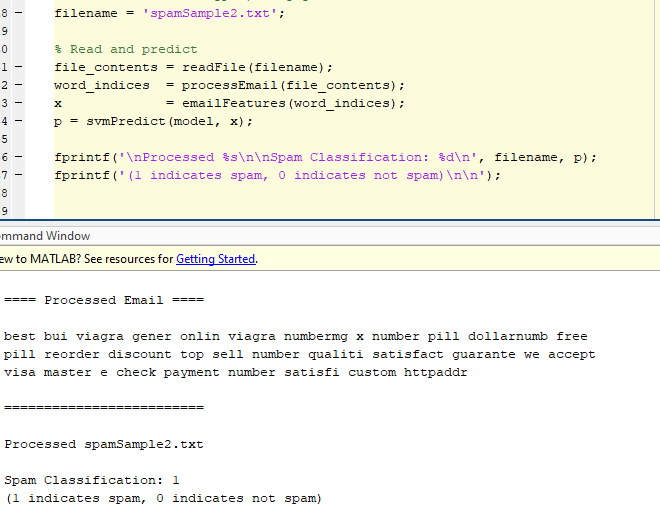


1. Проверьте работу классификатора на письмах из файлов emailSample1.txt, emailSample2.txt, spamSample1.txt и spamSample2.txt.







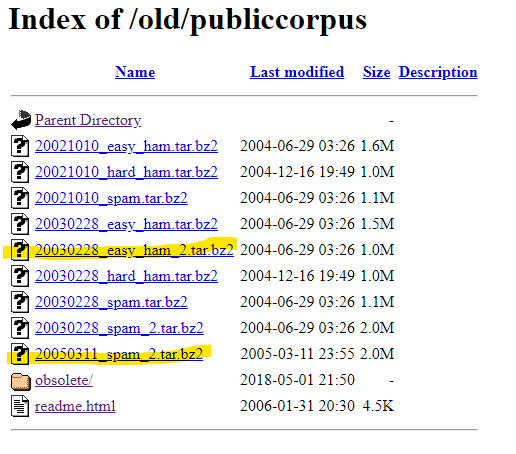


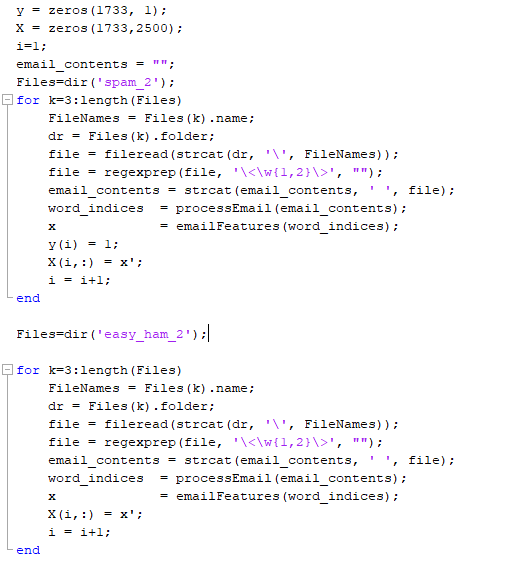
1. Создайте свой набор данных из оригинального корпуса текстов - <http://spamassassin.apache.org/old/publiccorpus/>.

Были взяты два датасета: один содержит спам([20050311\_spam\_2.tar.bz2](http://spamassassin.apache.org/old/publiccorpus/20050311_spam_2.tar.bz2)), второй обычные сообщения([20030228\_easy\_ham\_2.tar.bz2](http://spamassassin.apache.org/old/publiccorpus/20030228_easy_ham_2.tar.bz2)).

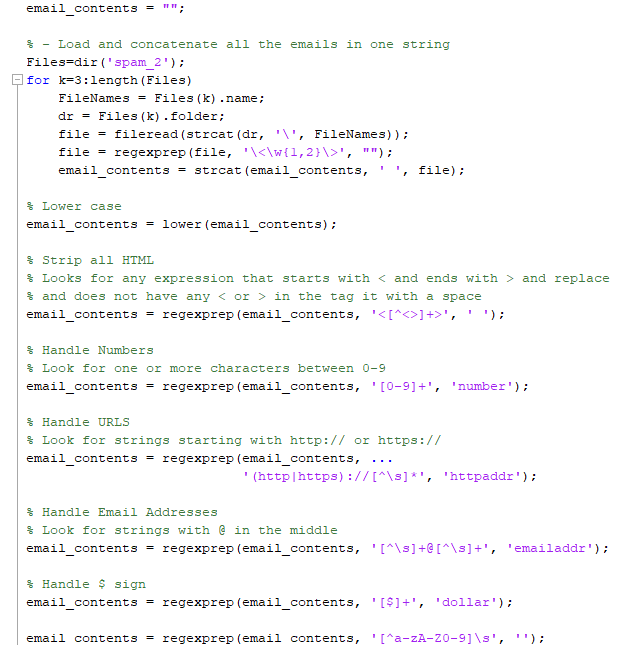
На их основе были заполнены X и y, а также вокабуляр, содержащий 2500 слов.

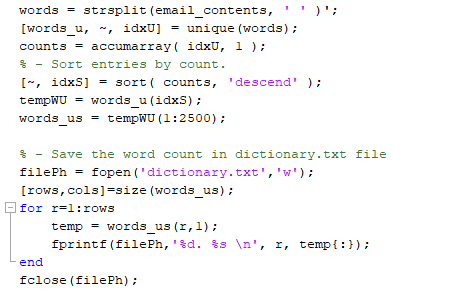
Получились массив Х размером 1733х2500 и вектор Y размером 1733х1.

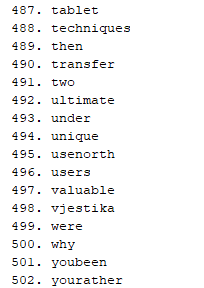




1. Постройте собственный словарь.







1. Как изменилось качество классификации? Почему?

Точность классификаций упала, так как в собранном датасете меньше данных чем из заранее заготовленных примеров.