hw3_NB

2023 年 7 月 9 日

张峪齐 3200105176

0.0.1 第一题

使用朴素贝叶斯过滤垃圾邮件

题目 (a). 收集数据：提供文本文件。

(b). 准备数据：将文本文件解析成词条向量。

(c). 分析数据：检查词条确保解析的正确性。

(d). 训练算法：使用我们之前建立的 trainNB0() 函数。

(e). 测试算法：使用 classifyNB()，并且构建一个新的测试函数来计算文档集的错误率。

(f). 使用算法：构建一个完整的程序对一组文档进行分类，将错分的文档输出到屏幕上。

解答

收集数据 读取所有文本，构建单词表

[ ]:
```
import re# 这个库中包含了正则表达式的函数
import numpy as np
# 定义一个函数，用于将文本转换为词列表表
def text_to_word_list(text):
  # 使用正则表达式将文本中的所有非字母数字字符替换为空格
  text = re.sub(r'[^a-zA-Z0-9]', ' ', text)
  # 将文本转换为小写
```

```
  text = text.lower()
  # 使用 split() 函数将文本分割成单词列表
  word_list = text.split()
  # 返回单词列表
  return word_list

def createMailList(path):
    mail_list=[]
    for i in range(1,26):
        mail_list.append(text_to_word_list(open(path+"/ham/"+str(i)+".
 ↪txt",'r',encoding='gbk').read()))
        #print(open(route,'r',encoding='gbk').read())
    for i in range(1,26):
        mail_list.append(text_to_word_list(open(path+"/spam/"+str(i)+".
 ↪txt",'r',encoding='gbk').read()))
        #print(open(route,'r',encoding='gbk').read())
    return mail_list

def createVocabList(mail_list):
    vocabSet = set([]) # 创建一个空的不重复列表
    for words_list in mail_list:
        vocabSet = vocabSet | set(words_list) # 取并集
    return list(vocabSet)

path="./hw3_NB/email"
mail_list=createMailList(path)
vocab_list=createVocabList(mail_list)
class_vec=np.concatenate((np.zeros(25,dtype=int),np.ones(25,dtype=int))).
 ↪tolist()

#print(vocab_list)
#print(class_vec)
```

准备数据 将文本文件解析成词条向量

[ ]:
```
def setOfWords2Vec(vocabList, inputSet):
    returnVec = [0] * len(vocabList) # 创建一个其中所含元素都为 0 的向量
    for word in inputSet: # 遍历每个词条
        if word in vocabList: # 如果词条存在于词汇表中，则置 1
            returnVec[vocabList.index(word)] = 1
        else: print("the word: %s is not in my Vocabulary!" % word)
    return returnVec # 返回文档向量
```

分析数据　检查词条确保解析的正确性。

[ ]:
```
print("字典列表: ")
print(vocab_list)
print("\n\n")
temp=setOfWords2Vec(vocab_list,["this","is","a","test","text"])
```

字典列表:

['www', 'll', 'these', 'such', 'order', 'others', 'watson', '562', 'right', 'explosive', 'moderate', 'thing', 'reputable', 'party', 'far', 'cheap', 'permanantly', 'download', 'good', 'amex', 'changing', 'accept', 'professional', 'don', 'hope', 'winter', 'hermes', 'life', 'couple', 'help', 'notification', 'attaching', 'holiday', 'articles', 'x', 'jquery', 'father', 'butt', 'also', 'here', '513', 'want', 'message', 'o', 'should', 'china', 'earn', 'ideas', 'website', 'mandarin', 'xp', 'questions', 'coast', 'rock', 'comment', 'courier', 'business', 'suggest', 'being', '180', 'and', 'could', 'mom', 'safest', 'moneyback', 'success', 'en', 'docs', 'cca', '7', 'answer', 'stuff', 'while', 'below', 'connection', 'tickets', 'past', 'programming', 'gpu', 'have', 'high', '15mg', 'via', 'sky', '15', 'longer', 'cost', 'fermi', 'certified', 'price', 'ferguson', 'book', 'hold', 'bargains', 'buyviagra', 'mailing', '588', 'from', '750', 'prices', 'penisen1argement', 'page', 'expertise', 'fine', 'mail', '100m', 'noprescription', 'reliever', 'quality', 'another', 'at', 'narcotic', 'not', 'level', '50mg', 'where', 'extended', 'thought', 'out', 'creation', 'sent', 'lunch', '100', 'important', 'wasn', 'used', 'moderately', 'window', 'sounds', 'mandatory', '10mg', 'guy', 'members', 'how', 'ready', 'who', 'features', 'sure', 'wilson', 'edit', 'those', 'care', 'approach', 'two', '86152', 'brands', 'discussions', 'we', 'spaying', 'focusing', 'style', 'http', 'femaleviagra', '5', 'giants', 'hotels', 'competitive', 't', 'use', 'opportunity', 'check', '8', 'discount', 'since', 'died', 'louis', 'went', 'is',

'behind', 'lists', '90563', 'cannot', 'hours', 'guaranteeed', 'name', 'monte',
'cs5', '20', 'encourage', 'microsoft', 'just', 'may', 'peter', 'hi',
'thickness', 'u', 'severepain', 'withoutprescription', 'leaves', 'google',
'day', 'benoit', 'intenseorgasns', 'take', 'zach', '1924', 'eugene',
'financial', 'need', 'storedetailview', 'both', 'about', 'runs', 'bike',
'survive', 'yeah', 'working', 'enjoy', 'will', 'inches', 'arvind', 'advocate',
'call', 'class', 'link', 'shape', 'net', 'top', 'freeviagra', 'easily', 'up',
'year', 'color', 'amazing', '100mg', 've', 'grow', 'office', 'pills', 'blue',
'brand', 'dozen', 'differ', '9', 'yay', 'per', 'ok', '70', 'watchesstore',
'plus', 'often', 'magazine', 'over', 'if', 'bad', 'methylmorphine', 'using',
'2', 'announcement', 'sophisticated', 'now', 'fast', 'ofejacu1ate', 'pill',
'pls', 'inspired', 'retirement', 'must', 'tabs', 'superb', 'strategic', 'hamm',
'groups', 'on', 'drugs', 'definitely', 'percocet', 'phone', 'going', 'kerry',
'his', 'ones', '129', '119', 'because', 'inconvenience', 'yo', 'or',
'customized', 'to', 'thank', 'plane', 'site', 'same', 'worldwide', '203',
'follow', 'functionalities', 'your', 'methods', 'the', 'source', 'items', 'mba',
'pro', 'codeine', 'serial', 'as', 'canadian', 'location', 'pricing', 'linkedin',
'save', 'ordercializviagra', 'had', 'improving', 'yourpenis', 'experts', 'you',
'interesting', 'naturalpenisenhancement', 'doors', 'express', '39', '2010',
'assistance', 'much', 'works', '174623', 'school', 'discreet', 'mg', 'latest',
'138', 'es', 'i', 'reservation', 'support', 'turd', 'with', 'assigning', '30',
'prototype', '25mg', 'wholesale', 'once', 'writing', 'gas', 'gucci', 'away',
'thread', 'tent', 'most', 'uses', '570', 'doing', 'watches', 'glimpse', 'place',
'hangzhou', 'john', 'huge', 'adobe', 'share', 'dusty', 'pages', 'access',
'food', 'be', 'go', 'meet', 'hello', 'night', 'then', 'issues', 'zolpidem', '4',
'156', 'owner', 'm', 'private', '200', 'trip', 'softwares', '14th', 'can',
'view', 'one', 'mathematics', '37', 'me', 'placed', 'let', 'hl', 'genuine',
'supplement', 'com', 'grounds', 'pavilion', 'please', 'endorsed', '00',
'design', 'note', 'than', 'generation', '199', 'bin', 'files', 'drunk', 'might',
'got', '385', 'things', 'rent', 'focus', 'listed', 'photoshop', 'ap',
'knocking', 'nvidia', 'signed', 'either', 'store', 'aged', 'web', 'cartier',
'transformed', 'team', '50', 'stepp', 'inform', 'art', 'get', 'perhaps',
'supporting', 'approved', 'co', 'expo', '322', 'chinese', 'reply', 'tv', 'came',
'income', '38', 'management', 'fedex', 'tour', 'in', 'featured', 'designed',
'commented', 'derivatives', 'february', 'c', '2011', 'held', 'when', 'acrobat',
'launch', 'full', 'ma1eenhancement', 'experience', 'jose', 'haloney', 'opioid',
'selected', 'new', 'rain', 'know', 'storage', 'recieve', 'a', 'of', 'enough',

'vivek', 'door', 'no', 'cats', 'so', 'insights', 'heard', 'foaming',
'everything', 'ryan', 'them', 'volume', 'copy', 'windows', 'wednesday', 'money',
'292', 'brained', 'cheers', 'try', 'province', 'significantly', 'hotel', '3',
'mathematician', 'am', 'release', 'borders', 'pretty', 'tiffany',
'automatically', 'well', 'model', 'told', 'car', 'length', 'wallets', 'does',
'pharmacy', 'jar', 'development', 'instead', 'fans', 'forward', 'pain',
'automatic', 'chapter', 'october', 'ultimate', 'each', 'create', 'example',
'75', 'jewerly', 'update', 'whybrew', 'jay', '225', 'more', 'sites', 'decision',
'trusted', 'do', 'jocelyn', 'come', 'thousand', 'network', 'this', 'off', 'e',
'shipping', 'increase', 'd', '195', '1', 'talked', '366', 'brandviagra',
'capabilities', 'julius', 'only', 'bathroom', 'today', 'quantitative', 'fbi',
'computer', 'generates', 'like', 'address', 'done', 'status', 'online', 'sorry',
'vuitton', 's', 'girl', 'thanks', '430', '492', 'visa', 'file', '130', 'fda',
'learn', 'are', 'oris', 'vicodin', 'creative', 'job', 'products', 'find', 'ups',
'bettererections', 'york', 'herbal', '25', 'that', 'forum', 'credit', 'back',
'starting', 'regards', '98', 'low', 'ambiem', 'finder', 'proven', 'close',
'would', 'bags', 'dior', 'add', 'plugin', '120', 'treat', 'yesterday',
'specifically', 'wrote', 'think', 'computing', '85', 'knew', 'py', 'some',
'gain', 'doctor', 'thirumalai', 'germany', 'home', 'time', 'service', 'town',
'hommies', 'risk', '80', '5mg', 'running', 'enabled', 'phentermin', 'possible',
'rude', 'days', 'nature', 'k', 'said', 'parallel', 'troy', '625', 'effective',
'but', 'contact', 'invitation', 'horn', 'was', 'all', 'upload', 'favorite',
'required', '300x', 'located', 'by', 'viagranoprescription', '50092',
'specifications', 'code', '60', '325', 'safe', 'includes', 'needed', '291',
'item', 'sf', '10', 'thailand', 'jqplot', 'control', 'biggerpenis', 'series',
'incoming', 'having', 'logged', 'tool', 'program', 'tokyo', 'has', 'station',
'roofer', 'concise', 'mandelbrot', '66343', '30mg', 'email', 'dhl', 'famous',
'welcome', 'information', 'jpgs', 'work', '396', 'incredib1e', '0nline',
'millions', 'requested', 'through', 'my', 'game', 'warranty', 'delivery',
'harderecetions', 'received', 'prepared', 'python', 'buy', 'natural', 'least',
'individual', 'looking', 'way', 'ems', 'fractal', 'hydrocodone', 'wilmott',
'sliding', 'inside', '2007', 'chance', 'pick', 'fundamental', 'cold', 'keep',
'it', 'great', 'any', 'museum', 'speedpost', 'been', '11', 'too', 'for',
'accepted', 'modelling', 'riding', 'analgesic', 'pictures', 'cat', 'major',
'group', 'there', 'titles', 'see', 'oem', 'faster', 'gains', '90', 'scenic',
'saw', 'arolexbvlgari', 'train', 'free', 'finance', '219', 'doggy', 'number',
'they', 'exhibit', 'shipment', 'cards', 'carlo', 'tesla', 'strategy',

```
'scifinance', 'what', 'cuda', 'based', 'betterejacu1ation', 'an', 'made',
'core', 'changes', 'lined']
```

```
the word: test is not in my Vocabulary!
the word: text is not in my Vocabulary!
```

训练算法 使用我们之前建立的 trainNB0() 函数

[ ]:
```
def trainNB0(trainMatrix,trainCategory):
    '''
    Parameters:

    trainMatrix - 训练文档矩阵，即 setOfWords2Vec 返回的 returnVec 构成的矩阵

    trainCategory - 训练类别标签向量，即 loadDataSet 返回的 classVec

    Returns:

    p0Vect - 侮辱类的条件概率数组

    p1Vect - 非侮辱类的条件概率数组

    pAbusive - 文档属于侮辱类的概率

    p0Vect

    = p(w| 文档属于非侮辱类)

    = [p(w0| 文档属于非侮辱类),p(w1| 文档属于非侮辱类),.....]

    = 文档属于非侮辱类的情况下：[第 0 个单词的出现频率，第 1 个单词的出现频率，.....
    ↪]
```

```
    = 文档属于非侮辱类的情况下：[第 0 个单词的出现次数，第 1 个单词的出现次数，.....
↪] / 非侮辱类文档单词总数

    = p0Num / p0Denom
    '''
    # 计算训练的文档数目
    numTrainDocs = len(trainMatrix)
    # 计算每篇文档的词条数
    numWords = len(trainMatrix[0])
    # 文档属于侮辱类的概率
    pAbusive = sum(trainCategory)/float(numTrainDocs)



    p0Num = np.ones(numWords)
    p1Num = np.ones(numWords)
    #p1Num 是分子列表，其长度为单词表的长度，[]
    p0Denom = 2.0
    p1Denom = 2.0 # 分母初始化为 2，拉普拉斯平滑

    for i in range(numTrainDocs):
    # 统计属于垃圾邮件的条件概率所需的数据
        if trainCategory[i] == 1: # 如果 trainMatrix[i] 是垃圾邮件
            p1Num += trainMatrix[i]
            p1Denom += sum(trainMatrix[i])
        else:# 统计属于正常邮件的条件概率所需的数据
            p0Num += trainMatrix[i]
            p0Denom += sum(trainMatrix[i])
    # 取对数，防止下溢出
    p1Vect = np.log(p1Num/p1Denom)
    p0Vect = np.log(p0Num/p0Denom)
    # 返回属于正常邮件的条件概率数组，属于垃圾邮件的条件概率数组，文档属于垃圾邮件的
概率
    return p0Vect,p1Vect,pAbusive


trainMat = []
```

```
for postinDoc in mail_list:
    trainMat.append(setOfWords2Vec(vocab_list, postinDoc))
p0V, p1V, pAb = trainNB0(trainMat, class_vec)
print('p0V:\n', p0V)
print('p1V:\n', p1V)
print('classVec:\n', class_vec)
print('pAb:\n', pAb)
```

```
p0V:
 [-5.72684775 -6.13231286 -5.43916568 -5.72684775 -6.82546004 -6.13231286
 -6.82546004 -6.82546004 -5.72684775 -6.82546004 -6.82546004 -6.13231286
 -6.82546004 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -5.21602212 -6.82546004 -6.13231286 -5.72684775 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286
 -6.13231286 -6.13231286 -5.72684775 -6.13231286 -6.82546004 -6.13231286
 -5.72684775 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -5.72684775
 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.82546004
 -5.72684775 -6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.82546004
 -4.05287131 -5.43916568 -6.13231286 -6.82546004 -6.82546004 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -5.72684775
 -6.13231286 -5.72684775 -5.72684775 -6.13231286 -6.13231286 -6.13231286
 -6.13231286 -5.03370057 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -6.82546004 -6.13231286 -6.82546004 -6.13231286 -6.82546004 -6.82546004
 -6.13231286 -5.43916568 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -6.82546004 -4.62823546 -6.82546004 -6.13231286 -6.82546004 -6.13231286
 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004
 -6.82546004 -6.13231286 -4.74601849 -6.82546004 -5.21602212 -6.13231286
 -6.82546004 -6.13231286 -6.13231286 -5.72684775 -5.72684775 -6.13231286
 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.13231286
 -6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286
 -6.13231286 -5.72684775 -6.13231286 -6.13231286 -6.13231286 -6.13231286
 -6.82546004 -6.13231286 -5.72684775 -6.13231286 -6.13231286 -6.13231286
 -6.13231286 -6.82546004 -6.13231286 -5.03370057 -6.13231286 -6.13231286
 -6.13231286 -5.72684775 -6.82546004 -6.82546004 -6.13231286 -6.13231286
 -6.82546004 -5.72684775 -6.13231286 -6.82546004 -6.13231286 -6.13231286
 -6.82546004 -6.13231286 -6.13231286 -6.82546004 -5.72684775 -5.03370057
```

-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.82546004
-5.43916568 -5.72684775 -4.74601849 -4.87954989 -6.82546004 -6.13231286
-6.82546004 -6.82546004 -6.13231286 -6.13231286 -5.72684775 -6.13231286
-6.82546004 -5.72684775 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.13231286 -6.13231286 -6.13231286 -5.21602212 -6.13231286 -6.13231286
-6.13231286 -6.13231286 -5.72684775 -6.13231286 -5.03370057 -6.82546004
-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -6.82546004 -6.13231286 -5.43916568 -6.13231286
-6.13231286 -6.82546004 -6.82546004 -5.72684775 -6.82546004 -6.82546004
-6.82546004 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.82546004
-6.13231286 -6.82546004 -6.13231286 -6.82546004 -6.82546004 -6.82546004
-6.13231286 -6.13231286 -6.82546004 -5.21602212 -6.13231286 -6.82546004
-5.72684775 -5.72684775 -6.13231286 -6.13231286 -5.43916568 -6.82546004
-6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.82546004 -6.13231286 -6.13231286 -5.72684775 -4.05287131
-6.82546004 -6.13231286 -6.82546004 -5.72684775 -5.21602212 -6.13231286
-6.13231286 -5.72684775 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-6.13231286 -5.21602212 -6.13231286 -3.88102106 -5.72684775 -6.13231286
-6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-4.74601849 -6.82546004 -3.99224669 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.82546004 -6.13231286 -5.21602212 -6.82546004 -6.13231286
-6.13231286 -5.72684775 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-6.82546004 -6.82546004 -4.05287131 -5.72684775 -6.82546004 -6.13231286
-6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.13231286 -6.13231286 -6.82546004 -6.82546004 -5.72684775 -6.82546004
-6.82546004 -4.26051068 -6.13231286 -6.13231286 -6.13231286 -4.74601849
-6.13231286 -6.82546004 -6.13231286 -6.82546004 -6.82546004 -6.13231286
-6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -6.82546004 -5.72684775 -6.82546004 -6.13231286
-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286
-6.13231286 -6.13231286 -5.72684775 -6.13231286 -4.74601849 -5.72684775
-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.82546004 -6.82546004 -6.13231286 -5.72684775 -6.82546004 -6.82546004
-6.13231286 -6.82546004 -6.13231286 -4.87954989 -5.72684775 -5.21602212
-6.13231286 -6.82546004 -5.03370057 -6.13231286 -5.72684775 -6.13231286
-6.82546004 -6.82546004 -5.72684775 -6.13231286 -6.13231286 -6.13231286

-6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -5.43916568
-6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286 -5.72684775
-6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -5.72684775
-6.13231286 -6.13231286 -6.13231286 -5.21602212 -6.13231286 -6.13231286
-6.82546004 -6.82546004 -6.13231286 -6.82546004 -6.13231286 -5.43916568
-6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.82546004
-6.13231286 -4.62823546 -6.13231286 -6.13231286 -5.21602212 -6.13231286
-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004
-6.13231286 -6.82546004 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -5.43916568 -6.13231286 -5.21602212 -6.13231286
-6.82546004 -3.99224669 -4.74601849 -5.72684775 -6.13231286 -6.13231286
-5.72684775 -6.13231286 -5.72684775 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -5.43916568 -6.82546004 -6.13231286 -6.13231286
-6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286 -5.72684775
-6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286
-6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286
-6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286
-6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286
-6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-6.13231286 -6.82546004 -5.43916568 -6.13231286 -6.13231286 -6.82546004
-5.72684775 -6.13231286 -5.21602212 -6.82546004 -6.13231286 -4.62823546
-6.82546004 -5.72684775 -6.82546004 -6.82546004 -5.43916568 -6.82546004
-5.21602212 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
-5.43916568 -6.13231286 -5.72684775 -6.13231286 -6.13231286 -6.13231286
-6.13231286 -5.21602212 -6.13231286 -5.72684775 -5.21602212 -6.13231286
-6.13231286 -6.82546004 -5.03370057 -6.13231286 -6.13231286 -6.82546004
-6.82546004 -6.82546004 -6.13231286 -6.82546004 -6.82546004 -6.82546004
-4.74601849 -6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286
-6.82546004 -6.82546004 -6.82546004 -6.13231286 -6.82546004 -6.82546004
-4.74601849 -6.13231286 -6.82546004 -6.13231286 -6.13231286 -6.13231286
-6.13231286 -6.82546004 -6.82546004 -6.82546004 -6.82546004 -6.13231286
-5.72684775 -6.82546004 -6.82546004 -5.72684775 -6.13231286 -6.82546004
-6.82546004 -6.13231286 -6.13231286 -5.21602212 -5.72684775 -6.13231286
-6.13231286 -6.13231286 -6.13231286 -5.43916568 -6.82546004 -6.82546004

```
 -6.13231286 -6.13231286 -6.82546004 -5.43916568 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
 -6.82546004 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -6.13231286 -6.13231286 -6.13231286 -5.03370057 -5.21602212 -6.13231286
 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -5.43916568 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.82546004
 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -5.21602212
 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.13231286 -6.82546004
 -5.21602212 -6.82546004 -6.82546004 -6.13231286 -6.13231286 -6.13231286
 -5.72684775 -6.82546004 -6.82546004 -6.82546004 -6.82546004 -5.43916568
 -6.13231286 -5.21602212 -6.13231286 -6.82546004 -6.82546004 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -6.13231286 -5.72684775 -5.72684775 -6.82546004 -6.13231286 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.82546004 -6.13231286
 -6.13231286 -5.72684775 -6.13231286 -4.52287494 -6.82546004 -6.13231286
 -6.13231286 -6.82546004 -5.72684775 -6.13231286 -5.72684775 -4.87954989
 -6.82546004 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286
 -6.82546004 -6.13231286 -4.87954989 -6.82546004 -5.72684775 -6.82546004
 -6.13231286 -6.82546004 -6.82546004 -5.72684775 -6.13231286 -6.82546004
 -6.13231286 -6.13231286 -6.13231286 -6.82546004 -6.13231286 -6.13231286
 -5.43916568 -6.13231286 -6.82546004 -6.82546004 -6.13231286 -6.13231286
 -6.13231286 -6.13231286 -5.72684775 -6.13231286 -6.82546004 -6.82546004
 -5.43916568 -6.13231286 -6.13231286 -6.13231286 -6.13231286]
p1V:
 [-6.74641213 -6.74641213 -6.74641213 -6.74641213 -4.80050198 -6.74641213
 -6.05326495 -5.64779984 -6.74641213 -4.66697059 -5.64779984 -6.74641213
 -5.36011777 -6.74641213 -6.74641213 -5.64779984 -4.66697059 -6.05326495
 -6.74641213 -6.05326495 -6.74641213 -5.36011777 -6.05326495 -6.05326495
 -6.74641213 -6.74641213 -5.36011777 -6.05326495 -6.74641213 -6.74641213
 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -5.64779984 -6.74641213
 -6.74641213 -6.74641213 -6.74641213 -5.13697422 -5.64779984 -6.74641213
 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213
 -6.74641213 -6.74641213 -6.05326495 -6.74641213 -6.74641213 -4.66697059
 -6.74641213 -5.36011777 -6.05326495 -6.74641213 -6.74641213 -5.36011777
```

```
-3.97382341 -6.74641213 -6.74641213 -6.05326495 -5.64779984 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -5.64779984 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -4.44382704 -6.74641213 -5.64779984 -5.36011777 -6.74641213
-6.05326495 -6.74641213 -5.64779984 -6.74641213 -6.05326495 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -5.64779984 -6.05326495 -6.74641213
-6.05326495 -5.36011777 -6.05326495 -5.64779984 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -5.36011777 -5.64779984
-4.95465266 -6.74641213 -4.54918755 -5.13697422 -6.74641213 -6.74641213
-6.05326495 -6.74641213 -6.05326495 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -4.18146277 -6.74641213 -6.74641213 -5.64779984
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.05326495 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -5.36011777 -6.74641213 -5.36011777 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.05326495 -5.13697422 -6.74641213 -6.74641213
-6.05326495 -6.74641213 -6.74641213 -6.05326495 -5.64779984 -6.74641213
-5.36011777 -6.74641213 -6.74641213 -5.36011777 -6.74641213 -4.95465266
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -5.64779984
-6.74641213 -6.74641213 -6.05326495 -6.05326495 -6.74641213 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -4.66697059 -6.74641213
-5.64779984 -5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-4.66697059 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -5.36011777 -5.36011777 -4.54918755
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.05326495 -5.64779984 -6.05326495 -6.74641213 -4.34851686 -5.36011777
-6.74641213 -4.66697059 -6.05326495 -6.74641213 -6.05326495 -6.05326495
-5.13697422 -6.74641213 -6.05326495 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -6.05326495 -6.74641213 -5.64779984 -5.36011777 -6.05326495
-6.74641213 -6.74641213 -6.05326495 -6.74641213 -6.74641213 -5.64779984
-6.74641213 -6.05326495 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-4.66697059 -5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-5.64779984 -5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-5.64779984 -6.74641213 -5.64779984 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -5.64779984 -6.05326495 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -4.1073548  -6.74641213 -6.74641213
```

-6.74641213 -6.74641213 -6.05326495 -5.64779984 -6.74641213 -6.74641213
-4.95465266 -6.05326495 -5.13697422 -6.74641213 -6.74641213 -6.74641213
-6.05326495 -5.36011777 -6.74641213 -6.74641213 -5.36011777 -6.74641213
-6.74641213 -6.74641213 -5.13697422 -5.64779984 -6.74641213 -6.74641213
-4.66697059 -6.05326495 -4.26150548 -6.74641213 -5.64779984 -6.74641213
-5.36011777 -6.05326495 -6.05326495 -6.74641213 -6.74641213 -5.64779984
-6.74641213 -6.74641213 -5.64779984 -4.80050198 -6.74641213 -6.05326495
-6.05326495 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -3.91319878 -6.74641213 -6.05326495 -5.64779984 -6.74641213
-6.74641213 -6.74641213 -5.36011777 -6.74641213 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.05326495 -6.74641213 -5.36011777 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-4.54918755 -6.05326495 -6.74641213 -6.74641213 -5.64779984 -6.05326495
-6.74641213 -6.05326495 -6.74641213 -6.05326495 -6.74641213 -6.74641213
-6.74641213 -5.64779984 -6.74641213 -6.74641213 -6.05326495 -6.74641213
-5.64779984 -6.05326495 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-4.66697059 -5.13697422 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-5.36011777 -6.05326495 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -5.36011777 -6.05326495 -6.74641213 -5.13697422
-6.74641213 -6.74641213 -6.74641213 -5.36011777 -6.74641213 -6.74641213
-5.64779984 -5.36011777 -6.74641213 -6.05326495 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.05326495 -6.05326495 -6.74641213 -5.13697422
-6.74641213 -4.66697059 -6.74641213 -4.66697059 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -5.36011777 -6.05326495 -4.54918755 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-5.36011777 -4.54918755 -4.54918755 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-4.66697059 -6.74641213 -6.74641213 -4.66697059 -6.74641213 -6.05326495
-6.74641213 -6.05326495 -5.64779984 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -4.44382704 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -5.36011777 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -4.66697059 -5.36011777 -6.74641213

-5.36011777 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -4.95465266 -5.36011777 -6.74641213 -6.74641213
-6.74641213 -5.64779984 -4.80050198 -6.74641213 -6.74641213 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213 -6.05326495
-5.36011777 -5.64779984 -5.64779984 -4.66697059 -6.74641213 -6.05326495
-4.80050198 -6.74641213 -5.64779984 -6.05326495 -6.74641213 -6.74641213
-5.13697422 -6.74641213 -5.64779984 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -4.66697059
-6.74641213 -5.36011777 -6.74641213 -6.74641213 -6.74641213 -6.05326495
-6.05326495 -4.95465266 -6.74641213 -5.64779984 -5.64779984 -6.05326495
-6.74641213 -5.36011777 -6.05326495 -6.74641213 -6.74641213 -6.05326495
-6.05326495 -5.36011777 -6.05326495 -6.74641213 -4.66697059 -5.64779984
-5.64779984 -6.74641213 -5.36011777 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -5.64779984 -6.05326495 -6.05326495 -5.64779984 -6.74641213
-6.74641213 -5.36011777 -5.36011777 -6.74641213 -6.74641213 -4.95465266
-5.64779984 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -4.66697059 -4.66697059
-6.74641213 -6.74641213 -6.05326495 -6.05326495 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -5.36011777 -6.05326495 -6.74641213 -6.74641213
-5.64779984 -6.74641213 -6.05326495 -6.05326495 -6.74641213 -5.64779984
-6.74641213 -6.74641213 -6.74641213 -5.64779984 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -4.95465266 -6.74641213
-6.74641213 -5.36011777 -6.74641213 -6.74641213 -5.64779984 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -4.80050198 -6.05326495 -4.44382704
-6.74641213 -6.05326495 -6.05326495 -6.74641213 -6.74641213 -4.95465266
-6.74641213 -6.74641213 -4.66697059 -6.05326495 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -5.36011777
-6.74641213 -5.36011777 -5.36011777 -6.74641213 -6.74641213 -6.74641213
-6.05326495 -6.05326495 -4.66697059 -5.64779984 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -5.36011777 -5.64779984 -4.66697059
-6.74641213 -6.74641213 -6.74641213 -4.44382704 -4.66697059 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -5.36011777 -6.74641213 -6.05326495
-6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.05326495 -6.74641213
-6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -6.74641213
-6.74641213 -5.36011777 -6.74641213 -6.74641213 -6.74641213 -4.80050198

```
 -5.64779984 -6.74641213 -6.74641213 -5.64779984 -6.74641213 -6.74641213
 -5.36011777 -6.74641213 -6.74641213 -6.05326495 -6.74641213 -6.05326495
 -6.74641213 -4.66697059 -4.66697059 -6.74641213 -6.74641213 -5.36011777
 -6.74641213 -6.05326495 -6.74641213 -5.64779984 -6.74641213 -6.74641213
 -6.74641213 -6.74641213 -5.36011777 -5.36011777 -6.74641213 -6.74641213
 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.05326495 -4.66697059
 -6.74641213 -6.74641213 -6.74641213 -6.74641213 -6.74641213]
classVec:
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
pAb:
 0.5
```

测试算法 使用 classifyNB()，并且构建一个新的测试函数来计算文档集的错误率。

[ ]:
```python
def classifyNB(vec2Classify, p0Vec, p1Vec, pClass1):
    '''
    比较 P(class=0 | w=vec2Classify ) 与 P(class=1 | w=vec2Classify ) 的大小

    因为对于 class=1 和 class=0 的 P(ci|w)=p(w|ci)*p(ci)/p(w) 中，p(w) 不变，因此只比较分子

    因为 p(w|ci) 是经过取对数的，因此对 p(ci) 也取对数
    '''

    p1 = sum(vec2Classify * p1Vec) + np.log(pClass1)
    # 对应元素相乘。logA * B = logA + logB,
    # 所以这里加上 log(pClass1)
    p0 = sum(vec2Classify * p0Vec) + np.log(1.0 - pClass1)
    if p1 > p0:
        return 1
    else:
        return 0
```

使用算法 构建一个完整的程序对一组文档进行分类，将错分的文档输出到屏幕上。

将本文档的代码块合起来并去除多余的输出语句即为完整的程序

[ ]:
```
trainingSet = list(range(50)) #trainingSet 是记录用作训练数据的单词列表在
mail_list 中的下标
testSet=[]#testSet 是记录用作测试数据的单词列表的在 mail_list 中的下标
for i in range(16):# 随机抽取 16 个邮件作为训练集
    randIndex = int(np.random.uniform(0,len(trainingSet)))
    testSet.append(trainingSet[randIndex])
    del trainingSet[randIndex]
trainMat=[]
trainClasses = []
for docIndex in trainingSet:
    trainMat.append(setOfWords2Vec(vocab_list, mail_list[docIndex]))
    trainClasses.append(class_vec[docIndex])
p0V,p1V,pSpam = trainNB0(np.array(trainMat),np.array(trainClasses))
# 错误数
errorCount = 0
for docIndex in testSet:
    wordVector = setOfWords2Vec(vocab_list, mail_list[docIndex])
    if classifyNB(np.array(wordVector),p0V,p1V,pSpam) != class_vec[docIndex]:
        errorCount += 1
        print ("分类错误的邮件: \n", mail_list[docIndex])
        print(f"应为{'垃圾邮件' if class_vec[docIndex]==1 else '非垃圾邮件' }, 被
错误分类为{'垃圾邮件' if class_vec[docIndex]==0 else '非垃圾邮件' }")

print ('the error rate is: ', float(errorCount)/len(testSet))
```

分类错误的邮件:
['oem', 'adobe', 'microsoft', 'softwares', 'fast', 'order', 'and', 'download',
'microsoft', 'office', 'professional', 'plus', '2007', '2010', '129',
'microsoft', 'windows', '7', 'ultimate', '119', 'adobe', 'photoshop', 'cs5',
'extended', 'adobe', 'acrobat', '9', 'pro', 'extended', 'windows', 'xp',
'professional', 'thousand', 'more', 'titles']
应为垃圾邮件, 被错误分类为非垃圾邮件
the error rate is: 0.0625