1. **比较词集模型与词袋模型的分类效果。**

我选用了KNN分类器分别使用词集模式和词袋模式对影评进行分类

主要代码：

词集模式：

for n in range(len(documents)):

document\_words = set(documents[n][0])

for m in range(len(word\_features)):

if word\_features[m] in document\_words:

features[n,m] = 1

#只关注这条影评里有没有这个特征词

词袋模式：

for n in range(len(documents)):

document\_words1 = set(documents[n][0])

for m in range(len(word\_features)):

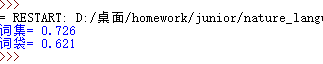
if word\_features[m] in document\_words1:

features1[n,m] = documents[n][0].count(word\_features[m])

#如果这条影评里有这个词，计算它有几个这个词

结果：

结果并不符合我们的预想，词袋模式的准确率比词集模式还要低



1. 在一个语料库上，提取每个文本中TFIDF最大的前n（5）个词，并观察与文本主题的关系。

我选用了搜狗语料库08及财经类的语料库，使用了前100个文本计算了TFIDF

代码：

import os

import codecs

import jieba

import re

from sklearn import feature\_extraction

from sklearn.feature\_extraction.text import TfidfTransformer

from sklearn.feature\_extraction.text import CountVectorizer

import numpy as np

from scipy import spatial

documents = []

txt = []

dirname = r'D:\桌面\homework\junior\nature\_language\_processing\data\Sogou\c000008\\'

files = os.listdir(dirname)

fm = 0

for fn in files:

try:

f = codecs.open(dirname+fn,mode='r')

text = f.readlines()

f.close()

voca = []

for t in text:

words = re.sub(r'[\t\r\n\u3000]','',t)

words = list(jieba.cut(words))

voca = voca + words

documents.append(' '.join(voca))

txt.append(fn)

fm += 1

except:

cotinue

if fm>100:

break

stopwords = ['(',')','，','。','？','“','”','‘','’','：','；','【','】','……']

v = CountVectorizer(min\_df=5,stop\_words=stopwords)

transformer=TfidfTransformer()#该类会统计每个词语的tf-idf权值

tfidf=transformer.fit\_transform(v.fit\_transform(documents))#fit\_transform计算tf-idf，fit\_transform将文本转为词频矩阵

word=v.get\_feature\_names()#获取词袋模型中的所有词语

tfidfmat=tfidf.toarray()#将tf-idf矩阵抽取出来，元素a[i][j]表示j词在i类文本中的tf-idf权重

for n in range(len(documents)):

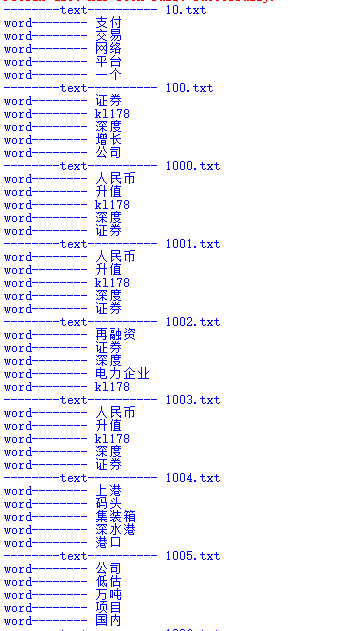
a=tfidfmat[n,:].argsort()

print('--------text----------',txt[n])

for m in range(-1,-6,-1):

print('word--------',word[a[m]])

结果：



可以看出，财经类的文本中，‘公交’、‘人民币’、‘证券’、‘深度’等词在一些文本里都属于TFIDF高的词语，符合财经的特征

1. 使用词袋模型作为垃圾邮件过滤器效果如何？会比词集模型好么

我使用聚类了k-means方法分别用词袋和词集对垃圾邮件进行了分类，结果发现词袋并没有比词集模式好

主要代码：

#----------------------模型建立-------------------------------------------------------

features = np.zeros([len(alltext),len(wordfeatures)],dtype=float)

features1 = np.zeros([len(alltext),len(wordfeatures)],dtype=float)

for n in range(len(alltext)):

for m in range(len(wordfeatures)):

if wordfeatures[m][0] in alltext[n][0]:

features[n,m] = 1 #词集

features1[n,m] = alltext[n][0].count(wordfeatures[m][0]) #词袋

#------------------------分类器训练与测试----------------------------------------

data1 = whiten(features)

centroids1,\_ = kmeans(data1,2)

idx1,\_ = vq(data1,centroids1)

data2 = whiten(features1)

centroids2,\_ = kmeans(data2,2)

idx2,\_ = vq(data2,centroids2)

a1 = sum(target1 == idx1)/len(target1)

a2 = sum(target1 == idx2)/len(target1)

结果：

