Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Рубежный контроль №2

по дисциплине

«Методы машинного обучения»

Выполнил:
студент группы ИУ5-23М
Дин Но

Москва — 2021 г.

Краткое описание набора данных

Набор данных 20 групп новостей представляет собой набор из примерно 20 000 документов групп новостей, распределенных (почти) равномерно по 20 различным группам новостей. Насколько мне известно, первоначально она была собрана Кеном Лэнгом, вероятно, для его Newsweeder: Learning to filter netnews paper, хотя он явно не упоминает об этой коллекции. Коллекция 20 групп новостей стала популярным набором данных для экспериментов в текстовых приложениях методов машинного обучения, таких как классификация текста и кластеризация текста.

Импорт Наборов данных

```
import numpy as np
from sklearn.datasets import fetch_20newsgroups
from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
```

Поскольку количество наборов данных 20 классов слишком велико, а стоимость обучения слишком высока, я выбрал 7 из 20 классов, чтобы поэкспериментировать с более типичными категориями.

```
select = ['alt.atheism','comp.graphics','misc.forsale','rec.autos',
          'sci.crypt','soc.religion.christian','talk.politics.guns']
train=fetch_20newsgroups(subset='train',categories=select)
test=fetch_20newsgroups(subset='test',categories=select)
```

Примеры в наборе данных

Train:


```
 "From: C445585@mizzou1.missouri.edu (John Kelsey)\nSubject: Re: How large are commercial keys?\nNntp-Posting-Host: mizzou1.missouri.edu\nOrganization: Universi
 'From: mpaul@unl.edu (marxhausen paul)\nSubject: Re: "National repentance"\nOrganization: University of Nebraska--Lincoln\nLines: 37\n\nmcovingt@aisun3.ai.uga.
 'From: mike@pyrdc.UUCP (Mike Whitman)\nSubject: 49cm Womens bike for sale\nOrganization: Pyramid Technology, Government Systems\nLines: 29\n\nI have the follow
 "From: loki@acca.nmsu.edu (Entropic Destroyer)\nSubject: Re: Need info on 43:1 and suicide for refutation\nOrganization: New Mexico State University\nLines: 26
 'From: halle@rebecca.its.rpi.edu (Ezra D.B. Hall)\nSubject: Re: Receiver and C-101 equilizer for sale\nKeywords: receiver, equilizer ,sterio,amp\nArticle-I.D.:
 'From: James Edward Burns <ddujeb@arco.com>\nSubject: Re: SUPER MEGA AUTOMOBILE SIGHTING(s)!!!!! Exotics together!\nX-Xxdate: Tue, 20 Apr 93 00:07:01 GMT\nOrga
 'From: xx155@yfn.ysu.edu (Family Magazine Sysops)\nSubject: WITNESS & PROOF OF CHRIST\'S RESURRECTION\nReply-To: xx155@yfn.ysu.edu (Family Magazine Sysops)\nOr
 'Subject: Re: islamic authority over women\nFrom: bobbe@vice.ICO.TEK.COM (Robert Beauchaine)\nOrganization: Tektronix, Inc., Beaverton,  OR.\nLines: 46\n\nIn a
 'From: ed@cwis.unomaha.edu (Ed Stastny)\nSubject: The OTIS Project (FTP sites for original art and images)\nKeywords: Mr.Owl, how many licks...\nOrganization:
 'From: ricky@watson.ibm.com (Rick Turner)\nSubject: Re: images of earth\nDisclaimer: This posting represents the poster\'s views, not necessarily those of IBM.
 ...],
'filenames': array(['/root/scikit_learn_data/20news_home/20news-bydate-train/alt.atheism/53070',
       '/root/scikit_learn_data/20news_home/20news-bydate-train/soc.religion.christian/20838',
       '/root/scikit_learn_data/20news_home/20news-bydate-train/comp.graphics/38244',
       ...,
       '/root/scikit_learn_data/20news_home/20news-bydate-train/alt.atheism/53228',
       '/root/scikit_learn_data/20news_home/20news-bydate-train/soc.religion.christian/20734',
       '/root/scikit_learn_data/20news_home/20news-bydate-train/soc.religion.christian/20556'],
      dtype='<U86'),
'target': array([0, 5, 1, ..., 0, 5, 5]),
'target_names': ['alt.atheism',
 'comp.graphics',
 'misc.forsale',
 'rec.autos',
 'sci.crypt',
 'soc.religion.christian',
 'talk.politics.guns']}
```


Test:


```
 'From: brad@clarinet.com (Brad Templeton)\nSubject: Re: Dorothy Denning opposes Clipper, Capstone wiretap chips\nOrganization: ClariNet Communications Corp.\nL
 'From: pmetzger@snark.shearson.com (Perry E. Metzger)\nSubject: Re: Screw the people, crypto is for hard-core hackers & spooks  only\nOrganization: Partnership
 'From: PETCH@gvg47.gvg.tek.com (Chuck)\nSubject: Daily Verse\nLines: 4\n\nBut you will receive power when the Holy Spirit comes on you; and you will be\nmy wit
 'From: jburgin@ralph.cs.haverford.edu (Joshua Marc Burgin)\nSubject: FOR SALE - GRAPHIC EQUALIZER\nOrganization: Haverford College Computer Science Department\
 'From: tscwhitehead@vx9000.weber.edu (Clarke Whitehead)\nSubject: Re: Borland C++ 3.1 w/App Frmwrks ** FORSALE **\nNews-Software: VAX/VMS VNEWS 1.41    \nOrgan
 'From: mserv@mozart.cc.iup.edu (Mail Server)\nSubject: Re: Eternal Marriage\nLines: 31\n\n\n\n hall@vice.ico.tek.com (Hal F Lillywhite) writes:\n>In article <May
 'From: strnlght@netcom.com (David Sternlight)\nSubject: Back doors in Clipper?\nOrganization: DSI/USCRPAC\nLines: 20\n\n\nI think it very unlikely there are ba
 'From: tedwards@eng.umd.edu (Thomas Grant Edwards)\nSubject: Re: Clipper considered harmful\nOrganization: Project GLUE, University of Maryland, College Park\n
 'From: random@cbnewse.cb.att.com (David L. Pope)\nSubject: Re: CLINTON JOINS LIST OF GENOCIDAL SOCIALIST LEADERS\nOrganization: AT&T\nLines: 34\n\nFrom article
 'From: Alan.Olsen@p17.f40.n105.z1.fidonet.org (Alan Olsen)\nSubject: some thoughts on Christian books...\nLines: 32\n\nDN> I think I took on this \'liar, lunat
 "From: rvkl60@email.sps.mot.com (Tom Mathes)\nSubject: Re: locking lugnuts / tire rebalance??\nNntp-Posting-Host: 223.7.248.44\nOrganization: Motorola Inc, Aus
 "From: jimb@hpindda.cup.hp.com (James Bruder)\nSubject: Re: Changing brake fluid..is it necessary..\nOrganization: HP Information Networks, Cupertino, CA\nLine
 "From: shellgate!llo@uu4.psi.com (Larry L. Overacker)\nSubject: Re: If There Were No Hell\nOrganization: Shell Oil\nLines: 38\n\nOFM Comments:\n\n>[The only pr
 'From: alin@nyx.cs.du.edu (ailin lin)\nSubject: very cheap 386 motherboard\nOrganization: Nyx, Public Access Unix @ U. of Denver Math/CS dept.\nLines: 7\n\nNov
 ...],
'filenames': array(['/root/scikit_learn_data/20news_home/20news-bydate-test/alt.atheism/53583',
       '/root/scikit_learn_data/20news_home/20news-bydate-test/soc.religion.christian/21773',
       '/root/scikit_learn_data/20news_home/20news-bydate-test/talk.politics.guns/55485',
       ...,
       '/root/scikit_learn_data/20news_home/20news-bydate-test/soc.religion.christian/21511',
       '/root/scikit_learn_data/20news_home/20news-bydate-test/alt.atheism/54144',
       '/root/scikit_learn_data/20news_home/20news-bydate-test/rec.autos/103742'],
      dtype='<U85'),
'target': array([0, 5, 6, ..., 5, 0, 3]),
'target_names': ['alt.atheism',
 'comp.graphics',
 'misc.forsale',
 'rec.autos',
 'sci.crypt',
 'soc.religion.christian',
 'talk.politics.guns']}
```


Векторизация данных статьи

```
vectorizer1 = TfidfVectorizer(stop_words='english',lowercase=True)
vectorizer2 = CountVectorizer(stop_words='english',lowercase=True)
train1_v=vectorizer1.fit_transform(train.data)
train2_v=vectorizer2.fit_transform(train.data)
print(train1_v.shape)

test1_v=vectorizer1.transform(test.data)
test2_v=vectorizer2.transform(test.data)
print(test1_v.shape)
```

```
(3983, 40980)
(2652, 40980)
```

LinearSVC И Multinomial Naive Bayes (MNB)

```
from sklearn.metrics import accuracy_score,f1_score

clf_1=MultinomialNB(alpha=0.1,fit_prior = False)
clf_1.fit(train1_v,train.target)
pred_1=clf_1.predict(test1_v)

print(f1_score(test.target,pred_1,average='macro'))
print(accuracy_score(test.target,pred_1))
```

```
0.7991295102257887
0.807315233785822
```

```
clf_2=MultinomialNB(alpha=0.1,fit_prior = False)
clf_2.fit(train2_v,train.target)
pred_2=clf_2.predict(test2_v)

print(f1_score(test.target,pred_2,average='macro'))
print(accuracy_score(test.target,pred_2))
```

```
0.809256244555425
0.8122171945701357
```

```
clf_3=SVC()
clf_3.fit(train1_v,train.target)
pred_3=clf_3.predict(test1_v)

print(f1_score(test.target,pred_3,average='macro'))
print(accuracy_score(test.target,pred_3))
```

```
0.789824680407716
0.7967571644042232
```

```
clf_4=SVC()
clf_4.fit(train2_v,train.target)
pred_4=clf_4.predict(test2_v)

print(f1_score(test.target,pred_4,average='macro'))
print(accuracy_score(test.target,pred_4))
```

```
0.6205553116107276
0.610105580693816
```

Вывод

Мы видим, что точность составляет до 81,22%. При использовании классификатора MNB и использовании метода CountVectorizer для векторизации объектов.