6.Problem: Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task. Built-in Libraries can be used to write the program. Calculate the accuracy, precision, and recall for your data set.

Example1

Loading the 20 newsgroups dataset : The dataset is called “Twenty Newsgroups”. Here is the official description, quoted from the website:http://qwone.com/~jason/20Newsgroups/

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of our knowledge, it was originally collected by Ken Lang, probably for his paper “Newsweeder: Learning to filter netnews,” though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

In [2]:
```
from sklearn.datasets import fetch_20newsgroups
twenty_train = fetch_20newsgroups(subset='train', shuffle=True)
x = len(twenty_train.target_names)
print("\n The number of categories:",x)
print("\n The %d Different Categories of 20Newsgroups\n" %x)
i=1
for cat in twenty_train.target_names:
    print("Category[%d]:" %i,cat)
    i=i+1
```

```
print("\n Length of training data is",len(twenty_train.data))
print("\n Length of file names is ",len(twenty_train.filenames))

print("\n The Content/Data of First File is :\n")

print(twenty_train.data[0])
```

```
 The number of categories: 20

 The 20 Different Categories of 20Newsgroups

Category[1]: alt.atheism
Category[2]: comp.graphics
Category[3]: comp.os.ms-windows.misc
Category[4]: comp.sys.ibm.pc.hardware
Category[5]: comp.sys.mac.hardware
Category[6]: comp.windows.x
Category[7]: misc.forsale
Category[8]: rec.autos
Category[9]: rec.motorcycles
Category[10]: rec.sport.baseball
Category[11]: rec.sport.hockey
Category[12]: sci.crypt
Category[13]: sci.electronics
Category[14]: sci.med
Category[15]: sci.space
Category[16]: soc.religion.christian
Category[17]: talk.politics.guns
Category[18]: talk.politics.mideast
Category[19]: talk.politics.misc
Category[20]: talk.religion.misc

 Length of training data is 11314

 Length of file names is  11314

 The Content/Data of First File is :

From: lerxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
```

```
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
Lines: 15

 I was wondering if anyone out there could enlighten me on this car I s
aw
the other day. It was a 2-door sports car, looked to be from the late 6
0s/
early 70s. It was called a Bricklin. The doors were really small. In ad
dition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,
- IL
   ---- brought to you by your neighborhood Lerxst ----
```

In [4]:
```
print("\n The Contents/Data of First 10 Files is in Training Data :\n")

for i in range(0,10):
    print("\n FILE NO:%d \n"%(i+1))
    print(twenty_train.data[i])
```

```
 The Contents/Data of First 10 Files is in Training Data :


 FILE NO:1

From: lerxst@wam.umd.edu (where's my thing)
Subject: WHAT car is this!?
Nntp-Posting-Host: rac3.wam.umd.edu
Organization: University of Maryland, College Park
```

```
Lines: 15

 I was wondering if anyone out there could enlighten me on this car I s
aw
the other day. It was a 2-door sports car, looked to be from the late 6
0s/
early 70s. It was called a Bricklin. The doors were really small. In ad
dition,
the front bumper was separate from the rest of the body. This is
all I know. If anyone can tellme a model name, engine specs, years
of production, where this car is made, history, or whatever info you
have on this funky looking car, please e-mail.

Thanks,
- IL
   ---- brought to you by your neighborhood Lerxst ----




 FILE NO:2

From: guykuo@carson.u.washington.edu (Guy Kuo)
Subject: SI Clock Poll - Final Call
Summary: Final call for SI clock reports
Keywords: SI,acceleration,clock,upgrade
Article-I.D.: shelley.1qvfo9INNc3s
Organization: University of Washington
Lines: 11
NNTP-Posting-Host: carson.u.washington.edu

A fair number of brave souls who upgraded their SI clock oscillator hav
e
shared their experiences for this poll. Please send a brief message det
ailing
your experiences with the procedure. Top speed attained, CPU rated spee
d,
```

```
add on cards and adapters, heat sinks, hour of usage per day, floppy di
sk
functionality with 800 and 1.4 m floppies are especially requested.

I will be summarizing in the next two days, so please add to the networ
k
knowledge base if you have done the clock upgrade and haven't answered
this
poll. Thanks.

Guy Kuo <guykuo@u.washington.edu>


 FILE NO:3

From: twillis@ec.ecn.purdue.edu (Thomas E Willis)
Subject: PB questions...
Organization: Purdue University Engineering Computer Network
Distribution: usa
Lines: 36

well folks, my mac plus finally gave up the ghost this weekend after
starting life as a 512k way back in 1985.  sooo, i'm in the market for
a
new machine a bit sooner than i intended to be...

i'm looking into picking up a powerbook 160 or maybe 180 and have a bun
ch
of questions that (hopefully) somebody can answer:

* does anybody know any dirt on when the next round of powerbook
introductions are expected?  i'd heard the 185c was supposed to make an
appearence "this summer" but haven't heard anymore on it - and since i
don't have access to macleak, i was wondering if anybody out there had
more info...

* has anybody heard rumors about price drops to the powerbook line like
the
ones the duo's just went through recently?
```

```
* what's the impression of the display on the 180?  i could probably sw
ing
a 180 if i got the 80Mb disk rather than the 120, but i don't really ha
ve
a feel for how much "better" the display is (yea, it looks great in the
store, but is that all "wow" or is it really that good?).  could i soli
cit
some opinions of people who use the 160 and 180 day-to-day on if its wo
rth
taking the disk size and money hit to get the active display?  (i reali
ze
this is a real subjective question, but i've only played around with th
e
machines in a computer store breifly and figured the opinions of somebo
dy
who actually uses the machine daily might prove helpful).

* how well does hellcats perform?  ;)

thanks a bunch in advance for any info - if you could email, i'll post
a
summary (news reading time is at a premium with finals just around the
corner... :( )
--
Tom Willis  \  twillis@ecn.purdue.edu    \    Purdue Electrical Enginee
ring
-------------------------------------------------------------------------
----
"Convictions are more dangerous enemies of truth than lies."  - F. W.
Nietzsche


 FILE NO:4

From: jgreen@amber (Joe Green)
Subject: Re: Weitek P9000 ?
Organization: Harris Computer Systems Division
Lines: 14
```

```
Distribution: world
NNTP-Posting-Host: amber.ssd.csd.harris.com
X-Newsreader: TIN [version 1.1 PL9]

Robert J.C. Kyanko (rob@rjck.UUCP) wrote:
> abraxis@iastate.edu writes in article <abraxis.734340159@class1.iasta
te.edu>:
> > Anyone know about the Weitek P9000 graphics chip?
> As far as the low-level stuff goes, it looks pretty nice.  It's got t
his
> quadrilateral fill command that requires just the four points.

Do you have Weitek's address/phone number?  I'd like to get some inform
ation
about this chip.

--
Joe Green                                Harris Corporation
jgreen@csd.harris.com                    Computer Systems Division
"The only thing that really scares me is a person with no sense of humo
r."
                                                  -- Jonathan Winters


 FILE NO:5

From: jcm@head-cfa.harvard.edu (Jonathan McDowell)
Subject: Re: Shuttle Launch Question
Organization: Smithsonian Astrophysical Observatory, Cambridge, MA,  US
A
Distribution: sci
Lines: 23

From article <C5owCB.n3p@world.std.com>, by tombaker@world.std.com (Tom
A Baker):
>>In article <C5JLwx.4H9.1@cs.cmu.edu>, ETRAT@ttacs1.ttu.edu (Pack Rat)
writes...
>>>"Clear caution & warning memory.  Verify no unexpected
>>>errors. ...".  I am wondering what an "expected error" might
```

```
>>>be.  Sorry if this is a really dumb question, but
>
> Parity errors in memory or previously known conditions that were waiv
ered.
>    "Yes that is an error, but we already knew about it"
> I'd be curious as to what the real meaning of the quote is.
>
> tom


My understanding is that the 'expected errors' are basically
known bugs in the warning system software - things are checked
that don't have the right values in yet because they aren't
set till after launch, and suchlike. Rather than fix the code
and possibly introduce new bugs, they just tell the crew
'ok, if you see a warning no. 213 before liftoff, ignore it'.

 - Jonathan



 FILE NO:6

From: dfo@vttoulu.tko.vtt.fi (Foxvog Douglas)
Subject: Re: Rewording the Second Amendment (ideas)
Organization: VTT
Lines: 58

In article <1r1eu1$4t@transfer.stratus.com> cdt@sw.stratus.com (C. D. T
avares) writes:
>In article <1993Apr20.083057.16899@ousrvr.oulu.fi>, dfo@vttoulu.tko.vt
t.fi (Foxvog Douglas) writes:
>> In article <1qv87v$4j3@transfer.stratus.com> cdt@sw.stratus.com (C.
D. Tavares) writes:
>> >In article <C5n3GI.F8F@ulowell.ulowell.edu>, jrutledg@cs.ulowell.ed
u (John Lawrence Rutledge) writes:
>
>> >> The massive destructive power of many modern weapons, makes the
```

```
>> >> cost of an accidental or crimial usage of these weapons to great.
>> >> The weapons of mass destruction need to be in the control of
>> >> the government only.  Individual access would result in the
>> >> needless deaths of millions.  This makes the right of the people
>> >> to keep and bear many modern weapons non-existant.

>> >Thanks for stating where you're coming from.  Needless to say, I
>> >disagree on every count.

>> You believe that individuals should have the right to own weapons of
>> mass destruction?  I find it hard to believe that you would support
a
>> neighbor's right to keep nuclear weapons, biological weapons, and ne
rve
>> gas on his/her property.

>> If we cannot even agree on keeping weapons of mass destruction out o
f
>> the hands of individuals, can there be any hope for us?

>I don't sign any blank checks.

Of course.  The term must be rigidly defined in any bill.

>When Doug Foxvog says "weapons of mass destruction," he means CBW and
>nukes.  When Sarah Brady says "weapons of mass destruction" she means
>Street Sweeper shotguns and semi-automatic SKS rifles.

I doubt she uses this term for that.  You are using a quote allegedly
from her, can you back it up?

>When John
>Lawrence Rutledge says "weapons of mass destruction," and then immedia
tely
>follows it with:

>>> The US has thousands of people killed each year by handguns,
>>> this number can easily be reduced by putting reasonable restriction
s
```

```
>>> on them.

>...what does Rutledge mean by the term?

I read the article as presenting first an argument about weapons of mas
s
destruction (as commonly understood) and then switching to other topic
s.
The first point evidently was to show that not all weapons should be
allowed, and then the later analysis was, given this understanding, to
consider another class.

>cdt@rocket.sw.stratus.com   --If you believe that I speak for my compa
ny,
>OR cdt@vos.stratus.com        write today for my special Investors' Pa
cket...



--
doug foxvog
douglas.foxvog@vtt.fi


 FILE NO:7

From: bmdelane@quads.uchicago.edu (brian manning delaney)
Subject: Brain Tumor Treatment (thanks)
Reply-To: bmdelane@midway.uchicago.edu
Organization: University of Chicago
Lines: 12

There were a few people who responded to my request for info on
treatment for astrocytomas through email, whom I couldn't thank
directly because of mail-bouncing probs (Sean, Debra, and Sharon).  So
I thought I'd publicly thank everyone.

Thanks!
```

```
(I'm sure glad I accidentally hit "rn" instead of "rm" when I was
trying to delete a file last September. "Hmmm... 'News?' What's
this?"....)

-Brian


 FILE NO:8

From: bgrubb@dante.nmsu.edu (GRUBB)
Subject: Re: IDE vs SCSI
Organization: New Mexico State University, Las Cruces, NM
Lines: 44
Distribution: world
NNTP-Posting-Host: dante.nmsu.edu

DXB132@psuvm.psu.edu writes:
>In article <1qlbrlINN7rk@dns1.NMSU.Edu>, bgrubb@dante.nmsu.edu (GRUBB)
says:
>>In PC Magazine April 27, 1993:29 "Although SCSI is twice as fasst as
ESDI,
>>20% faster than IDE, and support up to 7 devices its acceptance ...ha
s
>>long been stalled by incompatability problems and installation headac
hes."

>I love it when magazine writers make stupid statements like that re:

>performance. Where do they get those numbers? I'll list the actual
>performance ranges, which should convince anyone that such a

>statement is absurd:

>SCSI-I ranges from 0-5MB/s.

>SCSI-II ranges from 0-40MB/s.
>IDE ranges from 0-8.3MB/s.
>ESDI is always 1.25MB/s (although there are some non-standard version
s)
```

```
ALL this shows is that YOU don't know much about SCSI.

SCSI-1 {with a SCSI-1 controler chip} range is indeed 0-5MB/s
and that is ALL you have right about SCSI
SCSI-1 {With a SCSI-2 controller chip}: 4-6MB/s with 10MB/s burst {8-bi
t}
 Note the INCREASE in SPEED, the Mac Quadra uses this version of SCSI-1
 so it DOES exist. Some PC use this set up too.
SCSI-2 {8-bit/SCSI-1 mode}:          4-6MB/s with 10MB/s burst
SCSI-2 {16-bit/wide or fast mode}:  8-12MB/s with 20MB/s burst
SCSI-2 {32-bit/wide AND fast}:     15-20MB/s with 40MB/s burst

By your OWN data the "Although SCSI is twice as fast as ESDI" is correc
t
With a SCSI-2 controller chip SCSI-1 can reach 10MB/s which is indeed
"20% faster than IDE" {120% of 8.3 is 9.96}. ALL these SCSI facts have
been
posted to this newsgroup in my Mac & IBM info sheet {available by FTP o
n
sumex-aim.stanford.edu (36.44.0.6) in the info-mac/report as
mac-ibm-compare[version #].txt (It should be 173 but 161 may still be t
here)}

Part of this problem is both Mac and IBM PC are inconsiant about what S
CSI
is which.  Though it is WELL documented that the Quadra has a SCSI-2 ch
ip
an Apple salesperson said "it uses a fast SCSI-1 chip" {Not at a 6MB/s,
10MB/s burst it does not. SCSI-1 is 5MB/s maximum synchronous and Quadr
a
uses ANsynchronous SCSI which is SLOWER}  It seems that Mac and IBM see
SCSI-1 interface and think 'SCSI-1' when it maybe a SCSI-1 interface dr
iven
in the machine by a SCSi-2 controller chip in 8-bit mode {Which is MUCH
FASTER then true SCSI-1 can go}.

Don't slam an article because you don't understand what is going on.
One reference for the Quadra's SCSI-2 controller chip is
(Digital Review, Oct 21, 1991 v8 n33 p8(1)).
```

FILE NO:9

From: holmes7000@iscsvax.uni.edu
Subject: WIn 3.0 ICON HELP PLEASE!
Organization: University of Northern Iowa
Lines: 10

I have win 3.0 and downloaded several icons and BMP's but I can't figur
e out
how to change the "wallpaper" or use the icons. Any help would be appr
eciated.

Thanx,

-Brando

PS Please E-mail me

FILE NO:10

From: kerr@ux1.cso.uiuc.edu (Stan Kerr)
Subject: Re: Sigma Designs Double up??
Article-I.D.: ux1.C52u8x.B62
Organization: University of Illinois at Urbana
Lines: 29

jap10@po.CWRU.Edu (Joseph A. Pellettiere) writes:

>   I am looking for any information about the Sigma Designs
>   double up board. All I can figure out is that it is a
>   hardware compression board that works with AutoDoubler, but
>   I am not sure about this. Also how much would one cost?

```
I've had the board for over a year, and it does work with Diskdoubler,
but not with Autodoubler, due to a licensing problem with Stac Technolo
gies,
the owners of the board's compression technology. (I'm writing this
from memory; I've lost the reference. Please correct me if I'm wrong.)

Using the board, I've had problems with file icons being lost, but it's
hard to say whether it's the board's fault or something else; however,
if I decompress the troubled file and recompress it without the board,
the icon usually reappears. Because of the above mentioned licensing
problem, the freeware expansion utility DD Expand will not decompress
a board-compressed file unless you have the board installed.

Since Stac has its own product now, it seems unlikely that the holes
in Autodoubler/Diskdoubler related to the board will be fixed.
Which is sad, and makes me very reluctant to buy Stac's product since
they're being so stinky. (But hey, that's competition.)
--

Stan Kerr
Computing & Communications Services Office, U of Illinois/Urbana
Phone: 217-333-5217  Email: stankerr@uiuc.edu
```

Considering only four Categories

In order to get faster execution times for this first example we will work on a partial dataset with only 4 categories out of the 20 available in the dataset:

In [5]:
```
categories = ['alt.atheism', 'soc.religion.christian','comp.graphics',
'sci.med']
twenty_train = fetch_20newsgroups(subset='train',     categories=catego
ries, shuffle=True, random_state=42)
print("\n Reduced Target Names:\n",twenty_train.target_names)
print("\n Reduced Target Length:\n", len(twenty_train.data))
print("\nFirst Document : ",twenty_train.data[0])
```

```
 Reduced Target Names:
 ['alt.atheism', 'comp.graphics', 'sci.med', 'soc.religion.christian']

 Reduced Target Length:
 2257

First Document :  From: sd345@city.ac.uk (Michael Collier)
Subject: Converting images to HP LaserJet III?
Nntp-Posting-Host: hampton
Organization: The City University
Lines: 14

Does anyone know of a good way (standard PC application/PD utility) to
convert tif/img/tga files into LaserJet III format.  We would also like
to
do the same, converting to HPGL (HP plotter) files.

Please email any response.

Is this the correct group?

Thanks in advance.  Michael.
--
Michael Collier (Programmer)                The Computer Unit,
Email: M.P.Collier@uk.ac.city               The City University,
Tel: 071 477-8000 x3769                     London,
Fax: 071 477-8565                           EC1V 0HB.
```

Extracting features from text files

In order to perform machine learning on text documents, we first need to turn the text content into numerical feature vectors.The most intuitive way to do so is the bags of words representation: assign a fixed integer id to each word occurring in any document of the training set (for instance by building a dictionary from words to integer indices). for each document #i, count the number of occurrences of each word w and store it in X[i, j] as the value of feature #j where j is the index of word w in the dictionary The bags of words representation implies that n_features is the number of distinct words in the corpus: this number is typically larger than 100,000. Tokenizing text with scikit-learn: Text

preprocessing, tokenizing and filtering of stopwords are included in a high level component that is able to build a dictionary of features and transform documents to feature vectors:

Word Occurrences

In [11]:
```
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(twenty_train.data)
print("\n(Target Length , Distinct Words):",X_train_counts.shape)
print("\n Frequency of the word algorithm:", count_vect.vocabulary_.get
('algorithm'))
```

```
(Target Length , Distinct Words): (2257, 35788)

 Frequency of the word algorithm: 4690
```

From occurrences to frequencies

Term Frequencies : Divide the number of occurrences of each word in a document by the total number of words in the document: these new features are called tf for Term Frequencies. Another refinement on top of tf is to downscale weights for words that occur in many documents in the corpus and are therefore less informative than those that occur only in a smaller portion of the corpus.This downscaling is called tf–idf for “Term Frequency times Inverse Document Frequency”. Both tf and tf–idf can be computed as follows:

In [12]:
```
from sklearn.feature_extraction.text import TfidfTransformer
tf_transformer = TfidfTransformer(use_idf=False).fit(X_train_counts)
X_train_tf = tf_transformer.transform(X_train_counts)
X_train_tf.shape
```

Out[12]:
```
(2257, 35788)
```

In the above example-code, we firstly use the fit(..) method to fit our estimator to the data and secondly the transform(..) method to transform our count-matrix to a tf-idf representation. These two steps can be combined to

achieve the same end result faster by skipping redundant processing. This is done through using the fit_transform(..) method as shown below, and as mentioned in the note in the previous section:

In [14]:
```
tfidf_transformer = TfidfTransformer()
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape
```

Out[14]:
```
(2257, 35788)
```

Now that we have our features, we can train a classifier to try to predict the category of a post. Let's start with a naïve Bayes classifier, which provides a nice baseline for this task. scikit-learn includes several variants of this classifier; the one most suitable for word counts is the multinomial variant:

Training a classifier

In [15]:
```
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(X_train_tfidf, twenty_train.target)
```

Predicting the Outcome

To try to predict the outcome on a new document we need to extract the features using almost the same feature extracting chain as before. The difference is that we call transform instead of fit_transform on the transformers, since they have already been fit to the training set:

In [16]:
```
docs_new = ['God is love', 'OpenGL on the GPU is fast']
X_new_counts = count_vect.transform(docs_new)
X_new_tfidf = tfidf_transformer.transform(X_new_counts)

predicted = clf.predict(X_new_tfidf)

for doc, category in zip(docs_new, predicted):
    print('%r => %s' % (doc, twenty_train.target_names[category]))
```

```
'God is love' => soc.religion.christian
'OpenGL on the GPU is fast' => comp.graphics
```

Building a pipeline

In order to make the vectorizer => transformer => classifier easier to work with, scikit-learn provides a Pipeline class that behaves like a compound classifier:

In [19]:
```
from sklearn.pipeline import Pipeline
text_clf = Pipeline([('vect', CountVectorizer()),
                     ('tfidf', TfidfTransformer()),
                     ('clf', MultinomialNB()),
 ])
```

The names vect, tfidf and clf (classifier) are arbitrary. We shall see their use in the section on grid search, below. We can now train the model with a single command:

In [20]:
```
text_clf.fit(twenty_train.data, twenty_train.target)
```

Out[20]:
```
Pipeline(memory=None,
     steps=[('vect', CountVectorizer(analyzer='word', binary=False, dec
ode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip...inear_tf=False, use_idf=True)), ('clf', MultinomialNB(a
lpha=1.0, class_prior=None, fit_prior=True))])
```

Evaluation of the performance on the test set

In [22]:
```
#Evaluating the predictive accuracy of the model is equally easy:
import numpy as np
twenty_test = fetch_20newsgroups(subset='test',categories=categories, s
huffle=True, random_state=42)
docs_test = twenty_test.data
predicted = text_clf.predict(docs_test)
np.mean(predicted == twenty_test.target)
```

Out[22]:
```
0.83488681757656458
```

scikit-learn further provides utilities for more detailed performance analysis of the results:

In [23]:
```
from sklearn import metrics
print(metrics.classification_report(twenty_test.target, predicted,
     target_names=twenty_test.target_names))
```

```
                          precision    recall  f1-score   support

             alt.atheism       0.97      0.60      0.74       319
           comp.graphics       0.96      0.89      0.92       389
                 sci.med       0.97      0.81      0.88       396
soc.religion.christian       0.65      0.99      0.78       398

             avg / total       0.88      0.83      0.84      1502
```

In [24]:
```
metrics.confusion_matrix(twenty_test.target, predicted)
```

Out[24]:
```
array([[192,   2,   6, 119],
       [  2, 347,   4,  36],
       [  2,  11, 322,  61],
       [  2,   2,   1, 393]], dtype=int64)
```

As expected the confusion matrix shows that posts from the newsgroups on atheism and christian are more often confused for one another than with computer graphics.

In [ ]:
```
Reference : http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
```