CS 532: Assignment 9

Dinesh Kumar Paladhi

Spring 2016

Contents

1 Problem 1

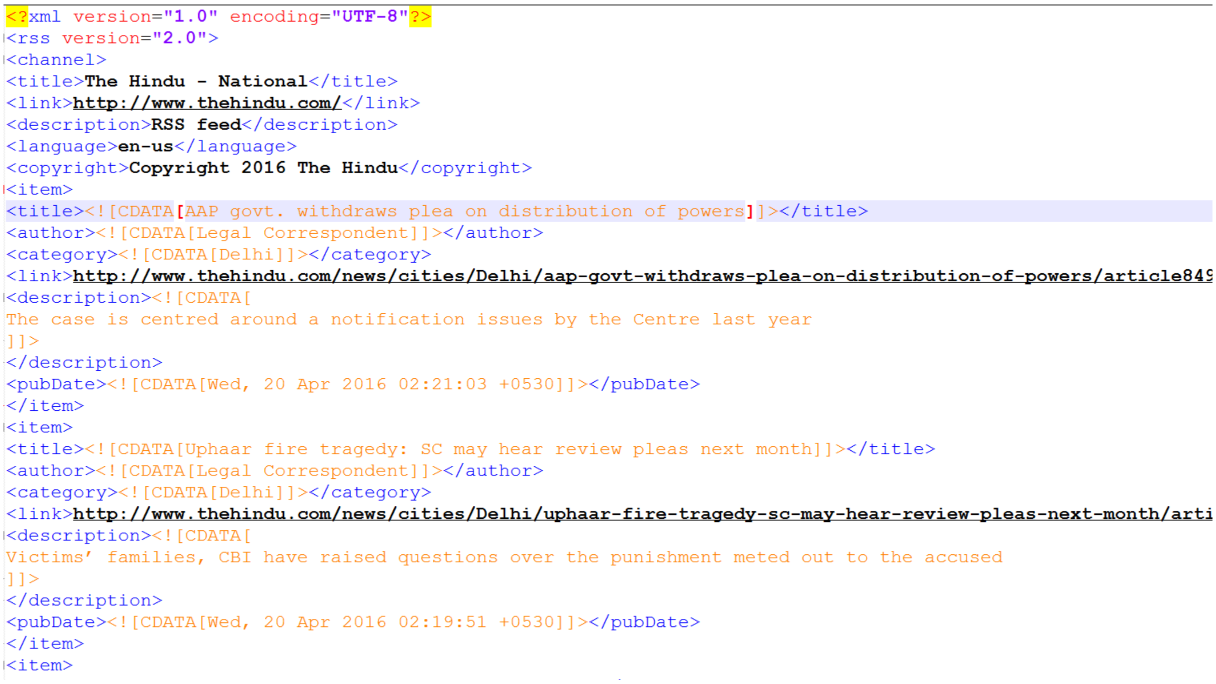
```
1.  Choose a blog or a newsfeed (or something similar with an Atom
or RSS feed).  Every student should do a unique feed, so please
‘‘claim’’ the feed on the class email list (first come, first served).
It should be on a topic or topics of which you are qualified to
provide classification training data.  Find something with at least
100 entries (or items if RSS).

Create between four and eight different categories for the entries
in the feed:

examples:

work, class, family, news, deals

liberal, conservative, moderate, libertarian

sports, local, financial, national, international, entertainment

metal, electronic, ambient, folk, hip-hop, pop

Download and process the pages of the feed as per the week 12
class slides.

Be sure to upload the raw data (Atom or RSS) to your github account.
```

1.1 Solution

1. I started doing this problem by first searchin for a blog which gives me minimum 100 feeds.
2. It became very tough for me to find a blog like that, Finally I have found this page “http://www.thehindu.com/news/natio with minimum of 100 rss feeds.
3. There were more than 100 rss feeds, but I picked only 100 from them and saved it in an xml file. Sample xml file can be seen in fig1.
4. Once I got these feeds I have read all of those titles and description of each feed and then I categorized them into the following categories.
5. Accident- Anything that is attacked,killed,injured,any loss happened,any illness or any one arrested. All these situations are categorized into accident category.
6. law- Any anouncement from supreme or high courts,any legal statements are categorized into law category.
7. politics- Any minister,government and its anouncements and activities are categorized into politics category.
8. elections- Any votings,campaigns,candidates,polls are categorized into elections category.
9. entertainment- Any anouncement regarding films,any happy moments,movies,songs,celebrations are categorized into entertainment category.
10. others- All those which do not come in above categories are categorized into others category.
11. transportation- Any information regarding buses,trains,travel details,traffic are categorized into transportation category.

1.2 Outputs

Sample Blog XML file



```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
<channel>
<title>The Hindu - National</title>
<link>http://www.thehindu.com/</link>
<description>RSS feed</description>
<language>en-us</language>
<copyright>Copyright 2016 The Hindu</copyright>
<item>
<title><![CDATA[AAP govt. withdraws plea on distribution of powers]]></title>
<author><![CDATA[Legal Correspondent]]></author>
<category><![CDATA[Delhi]]></category>
<link>http://www.thehindu.com/news/cities/Delhi/aap-govt-withdraws-plea-on-distribution-of-powers/article849
<description><![CDATA[
The case is centred around a notification issues by the Centre last year
]]>
</description>
<pubDate><![CDATA[Wed, 20 Apr 2016 02:21:03 +0530]]></pubDate>
</item>
<item>
<title><![CDATA[Uphaar fire tragedy: SC may hear review pleas next month]]></title>
<author><![CDATA[Legal Correspondent]]></author>
<category><![CDATA[Delhi]]></category>
<link>http://www.thehindu.com/news/cities/Delhi/uphaar-fire-tragedy-sc-may-hear-review-pleas-next-month/arti
<description><![CDATA[
Victims' families, CBI have raised questions over the punishment meted out to the accused
]]>
</description>
<pubDate><![CDATA[Wed, 20 Apr 2016 02:19:51 +0530]]></pubDate>
</item>
<item>
```

Figure 1: Sample list of Blog Feeds

2 Problem 2

```
2.  Manually classify the first 50 entries, and then classify (using
the fisher classifier) the remaining 50 entries.

Create a table with the title, predicted category, actual category,
and cprob() and fisherprob() for the actual category.
```

2.1 Solution

1. In order to classify the feeds into the respective categories I need to train the data first.
2. So what I did first is to divide the data into two parts. First 50 feeds are for training the code and the next 50 are for getting automatic classifications based on training.
3. This is done with the help of docclass and feedfilter codes which are taken as a reference from Programming collective Intelligence textbook.
4. Later by using the code I figured out to get the cprob() and fisherprob() values for the feeds.
5. I have saved all those outputs and can be seen in the following tables below.
6. Sample output for the first 50 feeds can be seen in the fig2 and the sample output for the last 50 feeds can be seen in the fig3/.

2.2 Code Listing

```

import feedparser
import re
import math
import docclass
# Takes a filename or URL of a blog feed and classifies the entries
def read(feed,classifier):

  splitRegexp = re.compile( r"<[^>]+>" )

  num=0
  Get feed entries and loop over them
  f=feedparser.parse(feed)
  print
  print '----- Begin manual classification (training) -----'
  for entry in f['entries'][0:50]:
    num=num +1
    # Print the contents of the entry
    title=entry['title'].encode('utf-8').replace("'","")
    print 'Title:      '+ title

    description = splitRegexp.sub( "", entry[ "description" ] )

    print description #entry['description'].encode('utf-8')

    # Combine all the text to create one item for the classifier
    #fulltext='%s\n%s\n%s' % (entry['title'],entry['publisher'],entry['description'])
    fulltext='%s\n%s' % (entry['title'],entry['description'])
    # Remove apostrophes
    fulltext=fulltext.replace("'","")
    # Print the best guess at the current category
    predicted=str(classifier.classify(fulltext))
    print 'Predicted category: ', predicted

    # Ask the user to specify the correct category and train on that
    actual=raw_input('Actual category: ')
    feature=None
    classifier.train(fulltext, actual)

    # Save the manual classifications
    # num, entry, feature, predicted, actual, cprob=None
    classifier.manualClassdb(num, title, feature, predicted, actual)

#def autoClassify(feed,classifier):
  num=50
  print '----- Begin automatic classification -----'
  # Get feed entries and loop over them
  f=feedparser.parse(feed)
  for entry in f['entries'][50:100]:
    num=num+1
    # Print the contents of the entry
    title=entry['title'].encode('utf-8').replace("'","")
    print 'Title:      '+ title
    description = splitRegexp.sub( "", entry[ "description" ] )

    print description #entry['description'].encode('utf-8')

    # Combine all the text to create one item for the classifier
    #fulltext='%s\n%s\n%s' % (entry['title'],entry['publisher'],entry['description'])
    fulltext='%s\n%s' % (entry['title'],entry['description'])
    fulltext=fulltext.replace("'","")
    # Print the best guess at the current category
    predicted=str(classifier.classify(fulltext))
    print 'Predicted: ', predicted

    # Ask the user to specify the correct category
    actual=raw_input('Enter actual category: ')
    feature=raw_input('Enter string classifier: ')

    #classifier.train(entry,cl)
    # probability the item should be in this category
    cp=round(classifier.cprob(feature,predicted),3)
    print 'cprob: ', str(cp)
    fischerprob1=round(classifier.fisherprob(feature,predicted),4)
    print 'fisherprob: ', str(fischerprob1)
    # Save the trained classifications
    # num, entry, feature, predicted, actual, cprob(feature, predicted)
    classifier.autoClassdb(num, title, feature, predicted, actual, cp)
   # entryfeatures(entry)
  #return classifier

def entryfeatures(entry):
  splitter=re.compile('\\W*')
  f={}

  # Extract the title words and annotate
  titlewords=[s.lower() for s in splitter.split(entry['title'])
          if len(s)>2 and len(s)<20]
  for w in titlewords: f['Title:'+w]=1

  # Extract the description words
  descriptionwords=[s.lower() for s in splitter.split(entry['description'])
          if len(s)>2 and len(s)<20]

  # Count uppercase words
  uc=0
  for i in range(len(descriptionwords)):
    w=descriptionwords[i]
    f[w]=1
    if w.isupper(): uc+=1

    # Get word pairs in description as features
    if i<len(descriptionwords)-1:
      twowords=' '.join(descriptionwords[i:i+1])
      f[twowords]=1

  # Removed: Keep creator and publisher whole
  #f['Publisher:'+entry['publisher']]=1

  # UPPERCASE is a virtual word flagging too much shouting
  if float(uc)/len(descriptionwords)>0.3: f['UPPERCASE']=1
  print f
  return f

def main():
  cl=docclass.fisherclassifier(docclass.getwords)
  cl.setdb('dpaladhi.db')
  read('my_data.xml',cl)
main()
```

Listing 1: Python Code for feedfilter

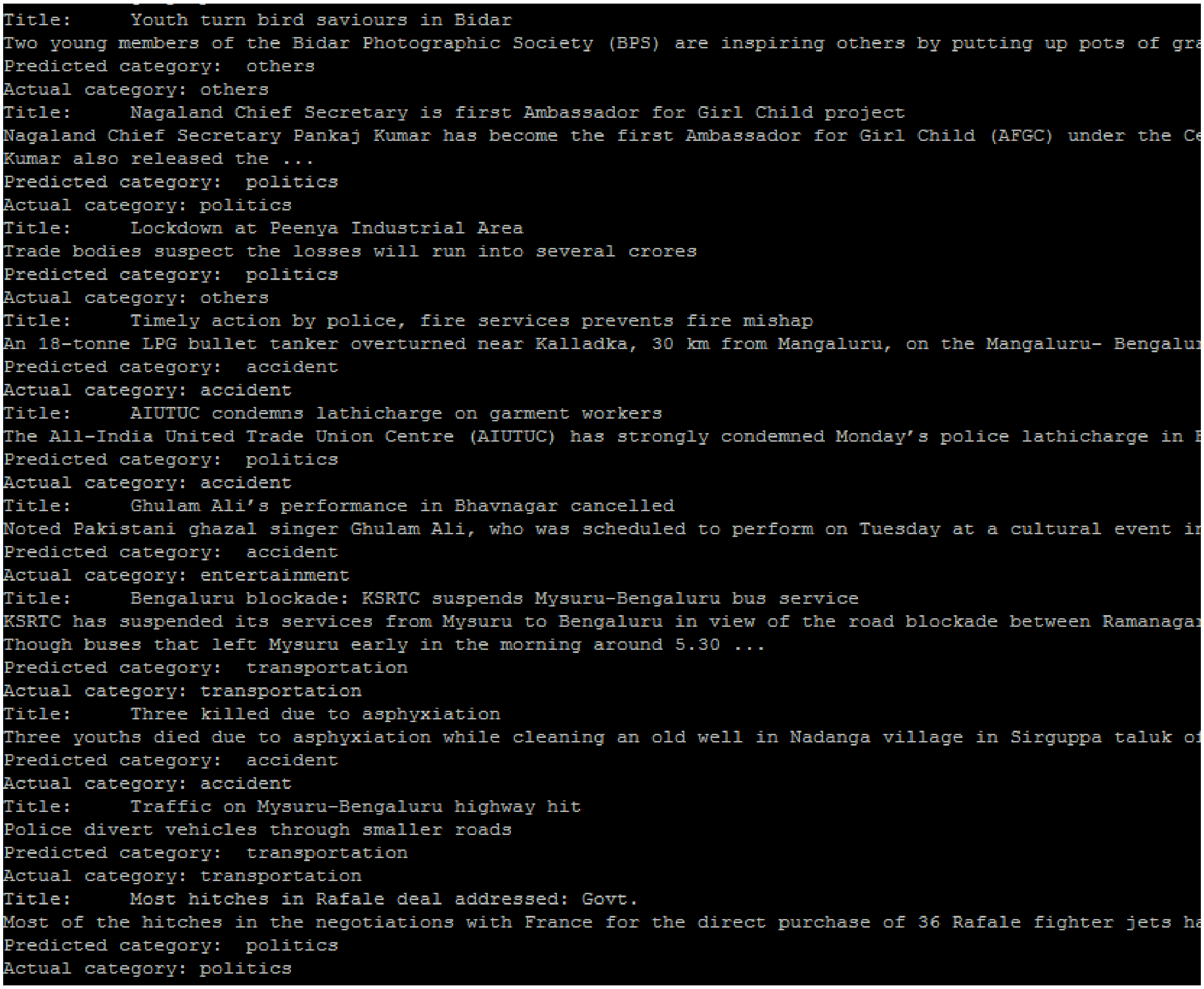
2.3 Code Listing

```
#from pysqlite2 import dbapi2 as sqlite
from sqlite3 import dbapi2 as sqlite
import re
import math

def getwords(doc):
  splitter=re.compile('\\W*')
  ## Remove all the HTML tags
  doc=re.compile(r'<[^>]+>').sub('',doc)
  # Split the words by non-alpha characters
  words=[s.lower() for s in splitter.split(doc)
          if len(s)>2 and len(s)<20]

  # Return the unique set of words only
  return dict([(w,1) for w in words])

class classifier:
  def __init__(self,getfeatures,filename=None):
    # Counts of feature/category combinations
    self.fc={}
    # Counts of documents in each category
    self.cc={}
    ## extract features for classification
    self.getfeatures=getfeatures

  def setdb(self,dbfile):
    self.con=sqlite.connect(dbfile)
    self.con.execute('create table if not exists rss(num, entry, feature, predicted, actual, cprob)')
    self.con.execute('create table if not exists fc(feature,category,count)')
    self.con.execute('create table if not exists cc(category,count)')
    # remove old data from previous sessions
    # self.con.execute('delete from rss')
    # self.con.execute('delete from fc')
    # self.con.execute('delete from cc')

  def manualClassdb (self,num, entry, feature, predicted, actual):
    self.con.execute("insert into rss values ('%s','%s', '%s', '%s','%s', '%s')"
                        % (num, entry, feature, predicted, actual, None))
    self.con.commit()

  def autoClassdb (self,num, entry, feature, predicted, actual, cp):
    self.con.execute("insert into rss values ('%s','%s', '%s', '%s','%s', '%s')"
                        % (num, entry, feature, predicted, actual, cp))
    self.con.commit()
  ## Increase the count of a feature/category pair
  def incf(self,f,cat):
    count=self.fcount(f,cat)
    if count==0:
      self.con.execute("insert into fc values ('%s','%s',1)"
                        % (f,cat.lower()))
    else:
      self.con.execute(
        "update fc set count=%d where feature='%s' and category='%s'"
        % (count+1,f,cat.lower()))

  ## The number of times a feature has appeared in a category
  def fcount(self,f,cat):
    res=self.con.execute(
      'select count from fc where feature="%s" and category="%s"'
      %(f,cat)).fetchone()
    if res==None: return 0
    else: return float(res[0])

  ## Increase the count of a category
  def incc(self,cat):
    count=self.catcount(cat)
    if count==0:
      self.con.execute("insert into cc values ('%s',1)" % (cat.lower()))
    else:
      self.con.execute(”update cc set count=%d where category=’%s’”
                       % (count+1,cat))

  ## The number of items in a category
  def catcount(self,cat):
    res=self.con.execute(’select count from cc where category=”%s”’
                         %(cat)).fetchone()
    if res==None: return 0
    else: return float(res[0])

  ## The list of all categories
  def categories(self):
    cur=self.con.execute(’select category from cc’);
    return [d[0] for d in cur]

  ## The total number of items
  def totalcount(self):
    res=self.con.execute(’select sum(count) from cc’).fetchone();
    if res==None: return 0
    return res[0]


  ## The train method takes an item(document) and a classification.
  ## It uses the getfeatures function to the break the item into its
  ## separate features. It then calls incf to increase the counts for
  ## this classification for every feature. Finally, it increases
  ## the total count for this classification.
  def train(self,item,cat):
    features=self.getfeatures(item)
    # Increment the count for every feature with this category
    for f in features:
      self.incf(f,cat)

    # Increment the count for this category
    self.incc(cat)
    self.con.commit()

  ## Probability is a number between 0 and 1, indicating
  ## the likelihood of an event. You calculate the probability of
  ## a word in a particular category by dividing the number of
  ## times the word appears in a document in that category
  ## by the total number of documents in the category.
  def fprob(self,f,cat):
    if self.catcount(cat)==0: return 0

    # The total number of times this feature appeared in this
    # category divided by the total number of items in this category
    return self.fcount(f,cat)/self.catcount(cat)

  def weightedprob(self,f,cat,prf,weight=1.0,ap=0.5):
    # Calculate current probability
    basicprob=prf(f,cat)

    # Count the number of times this feature has appeared in
    # all categories
    totals=sum([self.fcount(f,c) for c in self.categories()])

    # Calculate the weighted average
    bp=((weight*ap)+(totals*basicprob))/(weight+totals)
    return bp




class naivebayes(classifier):

  def __init__(self,getfeatures):
    classifier.__init__(self,getfeatures)
    self.thresholds={}

  def docprob(self,item,cat):
    features=self.getfeatures(item)

    # Multiply the probabilities of all the features together
    p=1
    for f in features: p*=self.weightedprob(f,cat,self.fprob)
    return p

  def prob(self,item,cat):
    catprob=self.catcount(cat)/self.totalcount()
    docprob=self.docprob(item,cat)
    return docprob*catprob

  def setthreshold(self,cat,t):
    self.thresholds[cat]=t

  def getthreshold(self,cat):
    if cat not in self.thresholds: return 1.0
    return self.thresholds[cat]

  def classify(self,item,default=None):
    probs={}
    # Find the category with the highest probability
    max=0.0
    for cat in self.categories():
      probs[cat]=self.prob(item,cat)
      if probs[cat]>max:
        max=probs[cat]
        best=cat

    # Make sure the probability exceeds threshold*next best
    for cat in probs:
      if cat==best: continue
      if probs[cat]*self.getthreshold(best)>probs[best]: return default
    return best

## This function will return the probability that an item with the
## specified feature belongs in the specified category, assuming there
## will be an equal number of items in each category.
class fisherclassifier(classifier):
  def cprob(self,f,cat):
    # The frequency of this feature in this category
    clf=self.fprob(f,cat)
    if clf==0: return 0

    # The frequency of this feature in all the categories
    freqsum=sum([self.fprob(f,c) for c in self.categories()])

    # The probability is the frequency in this category divided by
    # the overall frequency
    p=clf/(freqsum)

    return p


  def fisherprob(self,item,cat):
    # Multiply all the probabilities together
    p=1
    features=self.getfeatures(item)
    for f in features:
      p*=(self.weightedprob(f,cat,self.cprob))

    # Take the natural log and multiply by -2
    fscore=-2*math.log(p)
    fprobvalue=self.invchi2(fscore,len(features)*2)
    #print fprobvalue

    # Use the inverse chi2 function to get a probability
    return fprobvalue

  ## Inverse chi-squared function
  def invchi2(self,chi, df):
    m = chi / 2.0
    sum = term = math.exp(-m)
    for i in range(1, df//2):
        term *= m / i
        sum += term
    return min(sum, 1.0)

  def __init__(self,getfeatures):
    classifier.__init__(self,getfeatures)
    self.minimums={}

  def setminimum(self,cat,min):
    self.minimums[cat]=min

  def getminimum(self,cat):
    if cat not in self.minimums: return 0
    return self.minimums[cat]

  def classify(self,item,default=None):
    # Loop through looking for the best result
    best=default
    max=0.0
    for c in self.categories():
      p=self.fisherprob(item,c)
      # Make sure it exceeds its minimum
      if p>self.getminimum(c) and p>max:
        best=c
        max=p
    return best
```

Listing 2: Python Code for docclass

2.4 Outputs

Output 1



```
Title:     Youth turn bird saviours in Bidar
Two young members of the Bidar Photographic Society (BPS) are inspiring others by putting up pots of gra
Predicted category:  others
Actual category: others
Title:     Nagaland Chief Secretary is first Ambassador for Girl Child project
Nagaland Chief Secretary Pankaj Kumar has become the first Ambassador for Girl Child (AFGC) under the Ce
Kumar also released the ...
Predicted category:  politics
Actual category: politics
Title:     Lockdown at Peenya Industrial Area
Trade bodies suspect the losses will run into several crores
Predicted category:  politics
Actual category: others
Title:     Timely action by police, fire services prevents fire mishap
An 18-tonne LPG bullet tanker overturned near Kalladka, 30 km from Mangaluru, on the Mangaluru- Bengalur
Predicted category:  accident
Actual category: accident
Title:     AIUTUC condemns lathicharge on garment workers
The All-India United Trade Union Centre (AIUTUC) has strongly condemned Monday's police lathicharge in B
Predicted category:  politics
Actual category: accident
Title:     Ghulam Ali's performance in Bhavnagar cancelled
Noted Pakistani ghazal singer Ghulam Ali, who was scheduled to perform on Tuesday at a cultural event in
Predicted category:  accident
Actual category: entertainment
Title:     Bengaluru blockade: KSRTC suspends Mysuru-Bengaluru bus service
KSRTC has suspended its services from Mysuru to Bengaluru in view of the road blockade between Ramanagar
Though buses that left Mysuru early in the morning around 5.30 ...
Predicted category:  transportation
Actual category: transportation
Title:     Three killed due to asphyxiation
Three youths died due to asphyxiation while cleaning an old well in Nadanga village in Sirguppa taluk of
Predicted category:  accident
Actual category: accident
Title:     Traffic on Mysuru-Bengaluru highway hit
Police divert vehicles through smaller roads
Predicted category:  transportation
Actual category: transportation
Title:     Most hitches in Rafale deal addressed: Govt.
Most of the hitches in the negotiations with France for the direct purchase of 36 Rafale fighter jets ha
Predicted category:  politics
Actual category: politics
```

Figure 2: Sample outputs for first 50 feeds

Title	**Predicted**	**Actual**
AAP govt. withdraws plea on distribution of powers	none	politics
Uphaar fire tragedy: SC may hear review pleas next month	politics	law
Where the mind is not without fear	politics	elections
Buddhadeb on poll circuit	politics	elections
If we win, there will be common programme	elections	elections
Nitish Kumar will be next PM: Lalu	elections	elections
Union govt. allots Rs. 800 crore to clean up polluted lakes in garden	elections	politics
Will get Kohinoor back, says Centre	elections	politics
All members of mob equally guilty: HC	elections	law
Govt. yet to pay farmers in Punjab for procured wheat	politics	politics
SC asks NAAC to hear grievances of deemed varsities	elections	law
Developed countries must tax coal for climate fund	politics	politics
Kirpals body arrives in India	politics	accident
Centre introducing chaos: HC	politics	politics
Trading bloc to India: Cut tariffs or exit FTA talks	politics	others
PM invokes Vajpayee, moots development of Kashmir	politics	politics
Handwara rejoices as Army bunkers are dismantled	politics	accident
Two youths create oases for birds as Bidar sizzles	politics	others
Centre rejects T.N. proposal to free Rajiv Gandhi killers	politics	politics
Rajans choice of words could have been better	law	politics
As an alumnus, I feel hurt over JNU controversy: Nirmala	politics	politics
Woman Maoist killed in Gadhchiroli encounter	accident	accident
Insult to God to have unauthorised places of worship: SC	law	law
I have not been formally approached for Atulya Bharat: Amitabh Bachchan	politics	entertainment
Odisha to provide free drinking water to urban poor	politics	politics

Table 1: Manually classified first 25 entries

Title	**Predicted**	**Actual**
India to insist written commitment from Pak on NIA team visit	entertainment	accident
TN diocese sued for reinstating convicted priest	politics	accident
Massive effort to be launched for water conservation: Modi	politics	politics
Mahaveer Jayanti celebrated	politics	entertainment
Forty five more fire stations to be set up in Odisha: Patnaik	politics	others
LDF promises to free Vigilance, create 25 lakh jobs	politics	politics
Movement of buses affected between Tumakuru, Bengaluru	politics	transportation
Explaining Ola and Ubers surge pricing	elections	transportation
Trains between Mysuru, Bengaluru packed to capacity	transportation	transportation
Tirupati temple deposits 1,311 kg gold in bank	politics	others
Mob ransacks Hebbagodi police station	entertainment	accident
EC silent on complaints against AIADMK: Pon Radhakrishnan	politics	politics
Kalaburagi, Bidar record highest temperature	others	others
Women-only bus service launched in Kashmir	entertainment	transportation
Goa Government to bring monkey-hunting tribe to mainstream	politics	politics
Youth turn bird saviours in Bidar	others	others
Nagaland Chief Secretary is first Ambassador for Girl Child project	politics	politics
Lockdown at Peenya Industrial Area	politics	others
Timely action by police, fire services prevents fire mishap	accident	accident
AIUTUC condemns lathicharge on garment workers	politics	accident
Ghulam Alis performance in Bhavnagar cancelled	accident	entertainment
Bengaluru blockade: KSRTC suspends Mysuru-Bengaluru bus service	transportation	transportation
Three killed due to asphyxiation	accident	accident
Traffic on Mysuru-Bengaluru highway hit	transportation	transportation
Most hitches in Rafale deal addressed: Govt.	politics	politics

Table 2: Manually classified next 25 entries from 25 to 50 entries

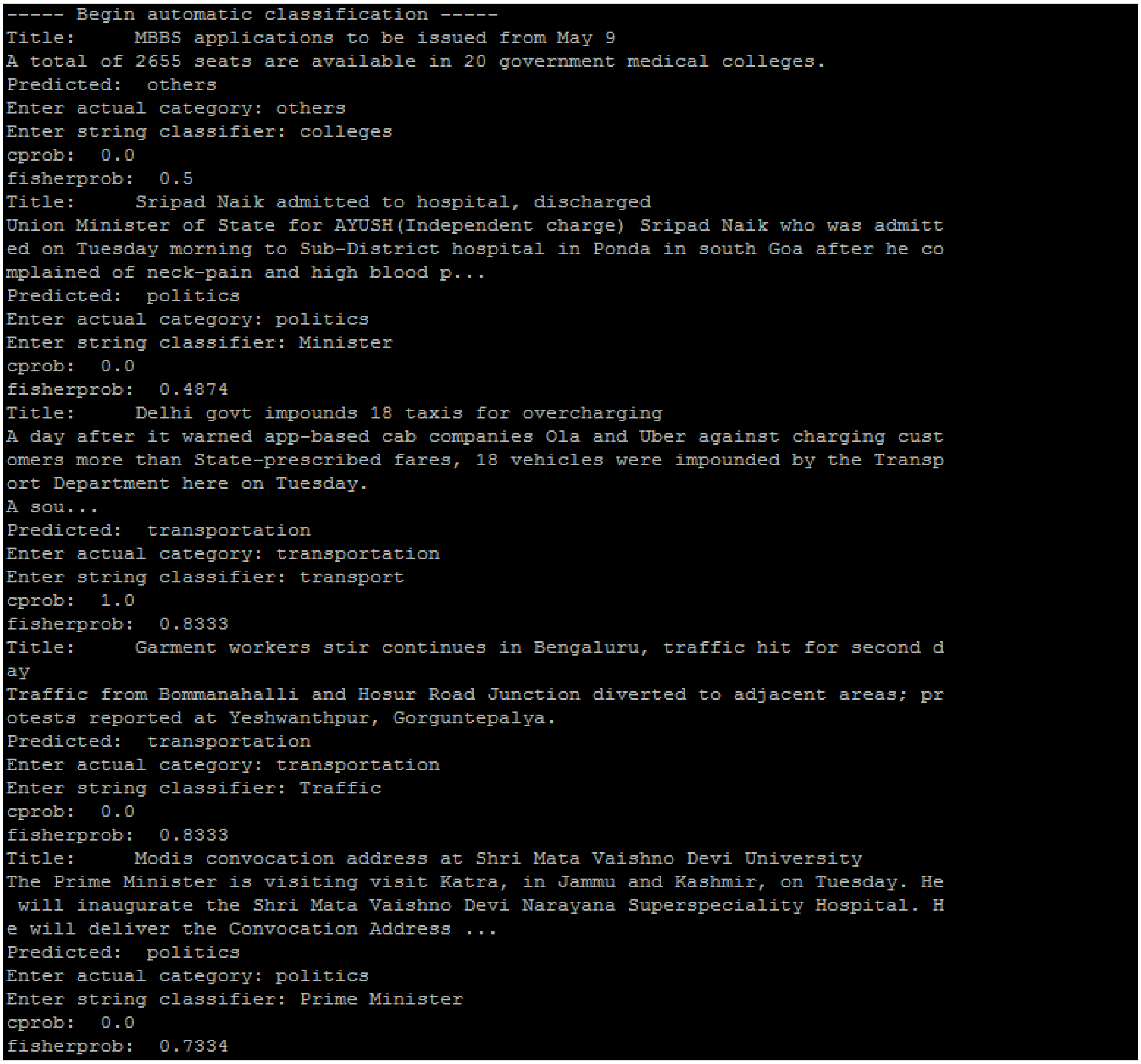
Title	**Predicted**	**Actual**	**Cprob**	**fisherprob**
MBBS applications to be issued from May 9	others	others	0.0	0.5
Sripad Naik admitted to hospital, discharged	politics	politics	0.0	0.4874
Delhi govt impounds 18 taxis for overcharging	transportation	transportation	1.0	0.8333
Garment workers stir continues in Bengaluru, traffic hit for second day	transportation	transportation	0.0	0.8333
Global economic situation grim, worrisome: Jaitley	transportation	others	0.0	0.5
Pak troops violate ceasefire	politics	politics	0.0	0.7334
Local youth held in Handwara molestation case	accident	accident	0.0	0.5
Adopt Periyars Self-Respect principles, Bhagwan tells deprived classes	politics	accident	0.0	0.5
Unstable academic calendar has made students life messy	politics	politics	0.0	0.9
Cops told to be on alert	entertainment	others	0.0	0.5
Children need protection from heat and dehydration	others	accident	0.0	0.5
Bengaluru Today	transportation	others	0.0	0.5
Waive farm loans even if it means borrowing advance	politics	politics	0.0	0.9
Kabini backwaters, a paradise for animals during drought	others	others	0.0	0.5
New pay still on paper for gazetted officers	others	others	0.0	0.5
Four killed in accident	accident	accident	0.0	0.5
Case registered against polytechnic staff for beating student	politics	accident	0.0	0.5
Training programme on tilapia fish farming	politics	others	0.0	0.5966
PEW busts spurious liquor-manufacturing unit	accident	accident	0.0	0.5
Tremors rock Andaman islands	entertainment	accident	0.0	0.5
JD(S) wins Hassan local body by-election	entertainment	elections	0.0	0.25
Where the grass is green even in blazing summer	transportation	others	0.0	0.5
Tamil Nadu Assembly elections Poll diary	accident	elections	0.0	0.5
Vasan to start his campaign from Papanasam	transportation	elections	0.0	0.25
An occasion to build new relationships	entertainment	entertainment	0.0	0.5

Table 3: Automatically classified 25 entries from 50 to 75 entries

Title	**Predicted**	**Actual**	**Cprob**	**fisherprob**
Bears nocturnal adventure triggers beehive of activity	accident	accident	0.0	0.5
CCTVs to be installed in parts of Karimnagar	entertainment	politics	0.0	0.25
Balineni scoffs at rumours	elections	politics	0.0	0.752
Monkey dies after being attacked by dog	transportation	transportation	0.0	0.8333
Troubled by tradition	politics	politics	0.0	0.5966
Red rebels kill Odisha villager	accident	accident	0.0	0.5
Punjab CM approves Rs 750 crore for roads	politics	politics	0.0	0.3289
14 cases reported	entertainment	elections	0.0	0.25
Dolphin washed ashore in Kilakkarai	politics	accident	0.0	0.5966
Fill vacancies in High Court	politics	law	0.0	0.2904
engagements	law	entertainment	0.0	0.5
Remembered only during elections, they harbour no high hopes	others	others	0.0	0.6407
Three-time MLA to contest in Cumbum	accident	elections	0.0	0.5
TMC fields candidates for Vilathikulam, Srivaikuntam	transportation	elections	0.0	0.5
High Court reserves order on Virbhadras children plea	politics	law	0.0	0.2904
Are there no rights violations in Union Territories, SC asks Centre	politics	politics	0.0	0.728
Animation film to promote brand Amaravati	others	entertainment	0.0	0.5
Seeking divine help to garner votes	politics	politics	0.0	0.75
Kochis public transport to take new route	transportation	transportation	1.0	0.8333
Ragi gruel centre launched for traffic police	politics	transportation	0.0	0.1667
Cinema	entertainment	entertainment	0.0	0.75
Engagements	accident	entertainment	0.0	0.5
Martyrs chronicle: bringing together bits and pieces	accident	accident	0.0	0.5
Vasan to start his campaign from Papanasam	accident	accident	0.0	0.5
Nine arrested on murder charge	accident	accident	1.0	0.8333

Table 4: Automatically classified next 25 entries from 75 to 100 entries

Output 2



```
----- Begin automatic classification -----
Title:     MBBS applications to be issued from May 9
A total of 2655 seats are available in 20 government medical colleges.
Predicted:  others
Enter actual category: others
Enter string classifier: colleges
cprob:  0.0
fisherprob:  0.5
Title:     Sripad Naik admitted to hospital, discharged
Union Minister of State for AYUSH(Independent charge) Sripad Naik who was admitt
ed on Tuesday morning to Sub-District hospital in Ponda in south Goa after he co
mplained of neck-pain and high blood p...
Predicted:  politics
Enter actual category: politics
Enter string classifier: Minister
cprob:  0.0
fisherprob:  0.4874
Title:     Delhi govt impounds 18 taxis for overcharging
A day after it warned app-based cab companies Ola and Uber against charging cust
omers more than State-prescribed fares, 18 vehicles were impounded by the Transp
ort Department here on Tuesday.
A sou...
Predicted:  transportation
Enter actual category: transportation
Enter string classifier: transport
cprob:  1.0
fisherprob:  0.8333
Title:     Garment workers stir continues in Bengaluru, traffic hit for second d
ay
Traffic from Bommanahalli and Hosur Road Junction diverted to adjacent areas; pr
otests reported at Yeshwanthpur, Gorguntepalya.
Predicted:  transportation
Enter actual category: transportation
Enter string classifier: Traffic
cprob:  0.0
fisherprob:  0.8333
Title:     Modis convocation address at Shri Mata Vaishno Devi University
The Prime Minister is visiting visit Katra, in Jammu and Kashmir, on Tuesday. He
 will inaugurate the Shri Mata Vaishno Devi Narayana Superspeciality Hospital. H
e will deliver the Convocation Address ...
Predicted:  politics
Enter actual category: politics
Enter string classifier: Prime Minister
cprob:  0.0
fisherprob:  0.7334
```

Figure 3: Sample output for last 50 feeds

3 Problem 3

```
3.  Assess the performance of your classifier in each of your
categories by computing precision, recall, and F-measure.
```

3.1 Solution

1. In this I need to calculate the precision, recall and F-measure for each of the category.
2. For doing this I have used the following formula.

$$Precesion = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2TP}{2TP + FP + FN}$$

3. TN is the true negative that is there is the respective category is not present in predicted and actual fields.
4. TP is the true positive that is there is the respective category is present in both predicted and actual fields.
5. FN is the False negative that is there is the respective category is present in actual and not present in predicted field.
6. FP is the False positive that is there is the respective category is not actual and is present in predicted field.

Category	**TN**	**TP**	**FN**	**FP**
accident	35	7	5	3
law	47	0	2	1
politics	33	9	1	7
elections	43	0	6	1
entertainment	4	2	3	5
others	37	4	7	2
transportation	40	3	1	6

Table 5: TN, TP, FN, FP values for each category

Category	**Precision**	**Recall**	**F-Measure**
accident	0.7	0.5833	0.6363
law	0	0	0
politics	0.5625	0.9	0.6923
elections	0	0	0
entertainment	0.2857	0.4	0.3333
others	0.6667	0.3636	0.4705
transportation	0.3333	0.75	0.4615

Table 6: Precision, Recall and F-Measure values for each category

Bibliography

[1] The Hindu, National, http://www.thehindu.com/news/national/, 2016

[2] Wikipedia,Precision and recall,https://en.wikipedia.org/wiki/Precision_and_recall, 2016

[3] Wikipedia, F1 score, https://en.wikipedia.org/wiki/F1_score, 2016