INTRODUCTION TO WEB SCIENCES:
Assignment 9

Manoj Chandra Kompalli

21 April 2016

Contents

1 Question 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).

1.1 Approach

A blog with more than 100 feeds was really a huge task. Most blogs had around 30 feeds. I had actually tried using a blog about astronomy. I could not categorize the items easily in that blog because most of the items were related.

- I have used a newspaper blogs sports page.
- The Newspaper is Indian Express. As it is an Indian Newspaper, all the sports articles are about the sports followed in India.
- Cricket is big in India and most articles are based on it.
- Football is a world sport so I found most articles about it.
- As the Rio Olympics are approaching, some data was about it.
- Tennis is also a big sport and currently many events are taking place all over the worls
- Hockey is iIndia's national sport and so some articles could be of that
- For sports like WWE, Mototsports I have the category Others

1.2 Input blog data

```
This XML file does not appear to have any style information associated with it. The document tree
<rss xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:wfw="http://wellformedweb.org/
<channel>
<title>The Indian ExpressSports - The Indian Express</title>
<atom:link href="http://indianexpress.com/section/sports/feed/" rel="self" type="application/rss+x
<link>http://indianexpress.com</link>
<description>
Latest News, Breaking News Live, Current Headlines, India News Online
</description>
<lastBuildDate>Thu, 21 Apr 2016 21:09:15 +0000</lastBuildDate>
<language>en</language>
<sy:updatePeriod>hourly</sy:updatePeriod>
<sy:updateFrequency>1</sy:updateFrequency>
<generator>http://wordpress.com/</generator>
<cloud domain="indianexpress.com" port="80" path="/?rsscloud=notify" registerProcedure="" protocol
<image>
<url>
http://0.gravatar.com/blavatar/efe0300e7f891c5c802ed340f6b20b67?s=96&d=http%3A%2F%2Fs2.wp.com%2Fi%
</url>
<title>Sports - The Indian Express</title>
<link>http://indianexpress.com</link>
</image>
<atom:link rel="search" type="application/opensearchdescription+xml" href="http://indianexpress.co
<atom:link rel="hub" href="http://indianexpress.com/?pushpress=hub"/>
<item>
<title>
Pune boy to appear in world in-line hockey Championship at New Zealand
</title>
<link>
http://indianexpress.com/article/sports/hockey/iihf-pune-boy-to-appear-in-world-in-line-hockey-cha
</link>
<comments>
http://indianexpress.com/article/sports/hockey/iihf-pune-boy-to-appear-in-world-in-line-hockey-cha
</comments>
<pubDate>Fri, 22 Apr 2016 01:40:06 +0000</pubDate>
<lastBuildDate>Fri, 22 Apr 2016 01:40:06 +0000</lastBuildDate>
<dc:creator>
<![CDATA[ Samreen Sayyed ]]>
</dc:creator>
```

Figure 1: Blog data in xml file containing 100 items

2 Question 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 Approach

I have started off by training the first 50 words. I have used methods like classify, fisherprob, cprob from docclass.py and feedfilter.py . Initially most predictions were incorrect but as the training progressed there were good number of hits mostly in categories like cricket and football.

- I have extracted fisherprob and cprob using the functions from docclass.
- Listings show Fisher method used which predicts category based on the entry.
- Based on the first 50 tarained entries, next 50 categories were predicted.
- I am showing title,actual,predicted for the first 50 feeds.
- I am displaying Title,actual,predicted,,cprob,fprob for the last 50 blog items

2.2 Tables

2.2.1 Manual Classification

No.	Title	Predicted	Actual
1	Pune boy to appear in world in-line hockey Champi-onship at New Zealand	None	hockey
2	Leicester City and Spurs dominate Players of the Year selection	hockey	football
3	Chris Gayle readies to unwind after birth of daughter Blush	hockey	cricket
4	Virat Kohli, Ishant Sharma, Ajinkya Rahane, Irfan Pathan indulge in charitable work	hockey	cricket
5	Leicester Citys Jamie Vardy accepts FA charge, awaits decision on ban	football	football
6	Rafael Nadal beats Albert Montanes to reach Barcelona Open quarterfinals	hockey	tennis
7	Louis Oosthuizen joins Vijay Singh, Adam Scott to give Rio Olympics golf a miss	hockey	olympics
8	IPL 2016: Sunrisers Hyderabad tame the Lions in their own backyard	football	cricket
9	It took guts to pick Marcus Rashford, says Manchester United boss Louis Van Gaal	football	football
3	Chris Gayle readies to unwind after birth of daughter Blush	hockey	cricket

4	Virat Kohli, Ishant Sharma, Ajinkya Rahane, Irfan Pathan indulge in charitable work	hockey	cricket
5	Leicester Citys Jamie Vardy accepts FA charge, awaits decision on ban	football	football
6	Rafael Nadal beats Albert Montanes to reach Barcelona Open quarterfinals	hockey	tennis
7	Louis Oosthuizen joins Vijay Singh, Adam Scott to give Rio Olympics golf a miss	hockey	olympics
8	IPL 2016: Sunrisers Hyderabad tame the Lions in their own backyard	football	cricket
9	It took guts to pick Marcus Rashford, says Manchester United boss Louis Van Gaal	football	football
17	IPL 2016: RCB replace injured Samuel Badree with uncapped South African Tabraiz Shamsi	cricket	cricket
18	Gagan Narang, Chain Singh to vie for final spot in Mens 50m rifle prone	cricket	others
19	Cristiano Ronaldo downplays injury, says all good	cricket	football
20	Jurgen Klopp formula starting to work magic for Liverpool, says James Milner	cricket	football
21	Uncle Virat Kohli congratulates Chris Gayle on birth of daughter Blush	cricket	cricket
22	Rio Games count down starts with Olympia torch lighting	cricket	olympics
23	Ive slept with more than 500 women while on tours for West Indies, former seamer Tino Best reveals in upcoming book	hockey	cricket
24	We, humans, failed you: Virat Kohli tweets on Shaktimans death	cricket	cricket
25	I just want to win, says Englands Ben Stokes	cricket	cricket
26	Leicester City fairytale sprinkles magic dust on rival fans	football	football
27	Jamie Vardy was very unlucky: Roy Hodgson	football	football
28	Chyna, former WWE womens champ, passes away	cricket	others
29	Liverpool thrash 10-man Everton 4-0 in Merseyside derby	football	football
30	Neymar to play for Brazil at Rio Games 2016, not Copa America	cricket	football
31	IPL 2016: KXIP opt for Dharamsala as second home venue, leave BCCI, HPCA confused	cricket	cricket
32	Will look to continue creating history: Dipa Karmakar	cricket	others
33	IPL 2016: Krunal Pandya knows his bowling well, says Rohit Sharma	cricket	cricket
34	Chris Gayle becomes father to girl Blush	cricket	cricket
35	Cristiano Ronaldo needs more rest after injury scare, says Zinedine Zidane	football	football
36	Matteo Darmian shines as Manchester United beat Crystal Palace 2-0	football	football

37	ICC to discuss structure, scheduling of bilateral series	cricket	cricket
38	Madras High Court dismisses PIL against BCCI	cricket	cricket
39	Inzamam-ul-Haq took up Pakistan chief selector job on condition of time to prepare squad for World Cup 2019	cricket	cricket
40	Barcelona snap losing streak to thump Deportivo 8-0; Atletico, Real Madrid win to keep league race alive	cricket	football
41	IPL 2016: Water woes brew in Jaipur	cricket	cricket
42	Chip off old block: Rahul Dravids son Samit scores 125 for school team	cricket	cricket
43	FIFA U-17 World Cup tournament director buoyed by Narendra Modis encouragement	cricket	football
44	Jack Wilshere will be available after West Brom, says Arsene Wenger	cricket	football
45	Bayern Munich keeper Manuel Neuer signs contract extension to 2021	cricket	football
46	Jose Mourinho, Claudio Ranieri to manage at Old Trafford for charity	football	football
47	Usain Bolts main goal: to defend the titles, do a threepeat at Rio Olympics 2016	cricket	olympics
48	I want to go and get more goals for Tottenham Hotspur, says Harry Kane	football	football
49	Bombay High Court allows BCCI to hold May 1 IPL 2016 MI vs RPS match in Pune	cricket	cricket
50	Athletics report: Tourist hurdler, tired athlete and a bumbling walker	cricket	others

Table 1: Entries classified manually

2.2.2 Last 50 items using fisher classifier

No.	Title	Predicted	Actual	CProb	FisherProb
51	Have to improve my landings, says Dipa Karmakar	cricket	Olympics	0	0.083
52	Sports federations concerned about venues for Rio Games 2016	cricket	Olympics	0	0.25
53	KKR put KXIP back in their place bottom	cricket	cricket	0	0.5
54	We have to win, we need to win, says Louis Van Gaal	football	football	0	0.833
55	PSG silent over Jose Mourinho reports	football	football	0	0.5
56	PM Modi lauds Dipa Karmakar for her determination	football	football	0	0.5
57	Pablo Zabaleta casts doubt over Manchester City future	football	football	0	0.5

58	From what Ive heard Antonio Contes disciplined, says Cesc Fabregas	football	football	0	0.5
59	Former German FA boss cleared over Qatar cancer comment	football	football	0	1
60	Formula 1 should stop tinkering with rules: Mercedes boss Toto Wolff	cricket	others	0	0.5
61	Vincent Kompany is working 100 percent with normality, says Manuel Pellegrini	football	football	0	0.75
62	The performance Tottenham Hotspur showed was perfect, says Mauricio Pochettino	football	football	0	0.75
63	SRH vs MI, IPL 2016: David Warner played a great innings, says Tim Southee	cricket	cricket	0	0.955
64	Next objective is a medal in Rio Olympics 2016, says Dipa Karmakar	others	olympics	0	0.083
65	IPL 2016, SRH vs MI: Barinder Sran fined for inappropriate conduct	cricket	cricket	0	1
66	Ex-CBI chief moves High Court over BCCI-ICC revenue model	cricket	cricket	0	0.75
67	Suresh Raina, wife Priyanka expecting first child	cricket	cricket	0	0.75
68	PCB disappointed after West Indies refuse to tour Pakistan in 2016	cricket	cricket	0	0.75
69	South Africa players against day-night Tests, feel disadvantaged	cricket	cricket	0	0.5
70	Spurs hit Stoke City for a four, cut Leicester Citys lead	football	football	0	0.875
71	Novak Djokovic, Serena Williams wins Laureus World Sportsman and Sportswoman of the Year awards	football	tennis	0	0.5
72	Formula One allows Pirelli more track days to test tyres for 2017	cricket	others	0	0.25
73	After win in Monte Carlo, Rafael Nadal turns sights to Barcelona	cricket	tennis	0	0.5
74	At Rio Olympics, India plans to promote Make in India initiative	olympics	olympics	0	0.87

75	One-state, one-vote in BCCI will lead to internal politics, BCA tells Supreme Court	cricket	cricket	0	0.9
76	Dipa Karmakar vaults into history books after qualifying for Rio Olympic Games 2016	others	Olympics	0	0.25
77	IPL 2016, SRH vs MI: In David vs Goliath, Warner the difference	cricket	cricket	0	0.75
78	Charged with improper conduct, Jamie Vardy could face extended ban	football	footbal	0	0.5
79	Newcastle United hoping Manchester City will be distracted by Champions League semi-final	football	football	0	0.5
80	Barcelonas slump baffles Carles Puyol, Luis Figo, Raul	football	football	0	0.5
81	Azlan Shah Cup: A few takeaways for India mens hockey team	cricket	hockey	0	0.25
82	IPL 2016, KXIP vs RPS: All is well that Glenn Maxwell ends	cricket	cricket	0	0.5
83	IPL 2016: Who said what about DDs win over RCB	cricket	cricket	0	0.833
84	Bengaluru FC retain I-League title with one game to play	football	football	0	1
85	Rafael Nadal is King of Clay once again with Monte Carlo title	football	tennis	0	0.25
86	Leaving Afghanistans coach job costs Inzamam-ul-Haq rupees 4 lakhs	cricket	cricket	0	0.75
87	Leicester City, minus Jamie Vardy, salvage 2-2 draw with West Ham United	football	football	0	0.666
88	Jurgen Klopp makes 10 changes as Liverpool beat Bournemouth	football	football	0	0.833
89	IPL 2016, RCB vs DD: DD beat RCB by 7 wickets	cricket	cricket	0	0.5
90	Rajkot police orders inquiry into celebratory firing at Ravindra Jadejas wedding	cricket	cricket	0	0.5
91	Roy Hodgson plays down Andy Carrolls Euro 2016 chances	football	football	0	0.5
92	Sebastian Vettel hits out at torpedo Daniil Kvyat	others	others	0	0.5
93	Engine problems force Lewis Hamilton to back of China grid	football	other	0	0.5

94	Positive start by Indian shooters in Rio World Cup	others	others	0	0.5
95	Harbhajan Singh can take his villa whenever he wants, says Amrapali CMD	football	cricket	0	0.125
96	IPL 2016: KKR beat SRH by eight wicket, Gambhir smashes unbeaten 90	cricket	cricket	0	0.5
97	Mitchell Starc ties the knot with girlfriend Alyssa Healy	cricket	cricket	0	0.5
98	IPL 2016 preview: Struggling MI face RCB challenge at Wankhede Stadium	cricket	cricket	1	0.955
99	IPL 2016, KXIP vs KKR: KXIP seek home comfort against KKR	cricket	cricket	1	0.833
100	IPL 2016, RCB vs DD: Quintal de Kock gives Delhi Daredevils heavyweight scalp	cricket	cricket	1	0.5

Table 2: Entries classified using Fisher Classifier

2.3 Code Listing

2.3.1 feedfilter.py

```
import docclass
import feedparser
import re
import math

# 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

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

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


```

```
28
29
30
31    # Combine all the text to create one item for the classifier
32      #fulltext='%s\n%s\n%s' % (entry['title'],entry['publisher'],entry['summary'])
33
34      fulltext='%s\n%s' % (entry['title'],entry['summary'])
35
36    # Remove apostrophes
37
38      fulltext=fulltext.replace("'","")
39
40    # Print the best guess at the current category
41
42      predicted=str(classifier.classify(fulltext))
43      print 'Predicted category: ', predicted
44
45      # Ask the user to specify the correct category and train on that
46
47      actual=raw_input('Actual category: ')
48      feature=None
49      classifier.train(fulltext, actual)
50
51      # Save the manual classifications
52      # num, entry, feature, predicted, actual, cprob=None
53
54      classifier.manualClassdb(num, title, feature, predicted, actual)
55
56  #def autoClassify(feed,classifier):
57    num=50
58    print '----- Begin automatic classification -----'
59    # Get feed entries and loop over them
60    f=feedparser.parse(feed)
61    for entry in f['entries'][50:100]:
62      num=num+1
63      # Print the contents of the entry
64      title=entry['title'].encode('utf-8').replace("'","")
65      print 'Title:     '+ title
66      summary = splitRegexp.sub( "", entry[ "summary" ] )
67
68      print summary #entry['summary'].encode('utf-8')
69
70      # Combine all the text to create one item for the classifier
71      #fulltext='%s\n%s\n%s' % (entry['title'],entry['publisher'],entry['summary'])
72      fulltext='%s\n%s' % (entry['title'],entry['summary'])
73      fulltext=fulltext.replace("'","")
74      # Print the best guess at the current category
75      predicted=str(classifier.classify(fulltext))
76      print 'Predicted: ', predicted
77
78      # Ask the user to specify the correct category
79      actual=raw_input('Enter actual category: ')
80      feature=raw_input('Enter string classifier: ')
81
82      #classifier.train(entry,cl)
83      # probability the item should be in this category
84      cp=round(classifier.cprob(feature,predicted),3)
```

```
    fp=round(classifier.fisherprob(feature,predicted),3)
    # print 'cprob: ', str(cp)
    print 'fisherprob: ', str(fp)

    # Save the trained classifications
    # num, entry, feature, predicted, actual, cprob(feature, predicted)
    classifier.autoClassdb(num, title, feature, predicted, actual, cp)
  #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 summary words
  summarywords=[s.lower() for s in splitter.split(entry['summary'])
          if len(s)>2 and len(s)<20]

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

    # Get word pairs in summary as features
    if i<len(summarywords)-1:
      twowords=' '.join(summarywords[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(summarywords)>0.3: f['UPPERCASE']=1

  return f
```

2.3.2 prog.py

```
import feedfilter

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

if __name__ == "__main__":
  main()
```

2.3.3 docclass.py

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

def getwords(doc):
  splitter=re.compile('\\W*')
  #print doc
  ## 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

```

```
157   def getthreshold(self,cat):
158     if cat not in self.thresholds: return 1.0
159     return self.thresholds[cat]
160 
161   def classify(self,item,default=None):
162     probs={}
163     # Find the category with the highest probability
164     max=0.0
165     for cat in self.categories():
166       probs[cat]=self.prob(item,cat)
167       if probs[cat]>max:
168         max=probs[cat]
169         best=cat
170 
171     # Make sure the probability exceeds threshold*next best
172     for cat in probs:
173       if cat==best: continue
174       if probs[cat]*self.getthreshold(best)>probs[best]: return default
175     return best
176 
177 ## This function will return the probability that an item with the
178 ## specified feature belongs in the specified category, assuming there
179 ## will be an equal number of items in each category.
180 class fisherclassifier(classifier):
181   def cprob(self,f,cat):
182     # The frequency of this feature in this category
183     clf=self.fprob(f,cat)
184     if clf==0: return 0
185 
186     # The frequency of this feature in all the categories
187     freqsum=sum([self.fprob(f,c) for c in self.categories()])
188 
189     # The probability is the frequency in this category divided by
190     # the overall frequency
191     p=clf/(freqsum)
192 
193     return p
194 
195 
196   def fisherprob(self,item,cat):
197     # Multiply all the probabilities together
198     p=1
199     features=self.getfeatures(item)
200     for f in features:
201       p*=(self.weightedprob(f,cat,self.cprob))
202 
203     # Take the natural log and multiply by -2
204     fscore=-2*math.log(p)
205 
206     # Use the inverse chi2 function to get a probability
207     return self.invchi2(fscore,len(features)*2)
208 
209   ## Inverse chi-squared function
210   def invchi2(self,chi, df):
211     m = chi / 2.0
212     sum = term = math.exp(-m)
213     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
```

3 Question 3:

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

3.1 Approach

Precision ,Recall and F measure are based on TP(True positive), True Negative(TN), False Positive(FP), False Negative(FN)

- Precision is the fraction of retrieved instances that are relevant
- Precision=TP/(TP+FP)
- Recall is the fraction of relevant instances that are retrieved
- Recall=TP/(TP+FN)
- F-Measure is a measure of a test's accuracy. It considers both the precision and the recall
- F1=2TP/(2TP+FP+FN)

3.2 Tables

Category	**TP**	**TN**	**FP**	**FN**
cricket	19	6	1	24
olympics	1	0	4	45
football	16	4	0	30
tennis	0	0	3	47
hockey	0	0	1	49
others	1	1	3	45

Table 3: TP TN FP FN values of different categories

Category	**Precision**	**Recall**	**F1**
cricket	0.95	0.44	0.6031
olympics	0.2	0.02173	0.03921
football	1	0.3478	0.5161
tennis	0	0	0
hockey	0	0	0
others	0.25	0.02173	0.04

Table 4: Precision Recall and F1 values for each category

References

[1] F-Measure wiki: https://en.wikipedia.org/wiki/F1_score.

[2] Precision Recall Wiki: https://en.wikipedia.org/wiki/Precision_and_recall.

[3] Programming Collective Intelligence : https://github.com/manojchandrak/Programming-Collective-Intelligence.

[]