CS 432/532: Web Science

Spring 2017

Assignment 9

Michael Micros

Instructor: Michael L. Nelson

Old Dominon University

Norfolk, Virginia

May 1, 2017

Problem 1: Choosing a blog and classifying entries

I originally "claimed" the headlines newsfeed from foxsports.com. Apparently, foxsports uses rss and not atom, which created a lot of difficulties for me. After battling with pagination for a while I decided to switch to a sports blog that I discovered, which coincidentally is very interesting and I believe ties in with this class to a certain degree. This blog is :

http://thehothand.blogspot.com/

This blog is writen by Texas Tech Professor Alan Reifman and its topic is statistical analysis of streaks in sports. In order to get the xml for 100 entries from the blog the following cURL command was used:

```
#! /bin/bash

curl -L "http://thehothand.blogspot.com/feeds/posts/default?max-results=100" > blogFeeds.xml
```

"blogFeeds.txt" was uploded to github. Additionally, the classification of the entries of the feed are in "classifications.txt".

```
1  Title                        Classification
2  ---------------------------------------------
3  UConn Women Go for 100       basketball
4  Straight Wins
5
6  Columbus Blue Jackets        hockey
7  Seek to Tie NHL Record
8  for Longest Winning
9  Streak
10
11 Gotta "Love" It -- 34        basketball
12 Points, 8-of-10 on Treys,
13 in First Quarter Alone
14
15 Michael Phelps Looks to      olympics
16 Extend Olympic-Gold
17 Streaks
18
19 What's Up (Or In this        baseball
20 Case, Down) With the
21 Cubs?
22
23 Serena Williams Just         other
24 Keeps Winning Grand Slam
25 Titles
26
27 Obscure Baseball-Card        baseball
28 Find: Walt Dropo, Co-
29 Record Holder for Hits in
30 Consecutive At-Bats (12)
31
```

Figure 1: Sample of classifyed entries from the "hothand" blog.

The categories that I created were:

- basketball
- baseball
- hockey
- olympics
- other

It is important to note that about 80% of this blog deals with streaks in basketbal and baseball. Therefore, I had to put tennis, footbal(suprizingly) and soccer in the "others" category, so as to limit the number of categories but also because there were so few entries in these categories.

Problem 2: Training and Testing (50-50)

Making use of the code provided by the "Programming Collective Intelligence" textbook we train the Fisher classifier on the first 50 entries and test it on the final 50. The feedfilter and docclass modules were imported to perform this. The code for this question and question 3 is "part1.py". The only parameter changed is the number of entries to use for training.The classifications for all the entries were saved in "class.txt" and loaded to the script.

```
num = 50

def q1Train(f,cat,num,classifier):
  # Get feed entries and loop over them
  print "--- TRAINING  ---"
  for entry in f['entries']:
    i = f.entries.index(entry)
    if( i < num):
      print
      print '-----'
      print 'Title:     '+entry['title'].encode('utf-8')
      fulltext='%s\n%s' % (entry['title'],entry['summary'][0])
      print i
      classifier.train(fulltext, cat[i])

  for i in classifier.categories():
    classifier.setminimum(i,.16)

def q1Test(f,cat,num,classifier):
  print "--- TESTING  ---"
  #h = open("table10.txt", "w")
  h = open("table50.txt", "w")
  h.write("title,  pred, act\n")

  for entry in f['entries']:
    i = f.entries.index(entry)
```

Figure 2: Part of "part1.py" code that trains and tested the Fisher classifyer.

The precision, and f-measure were calculated using the "assess" function displayed in Fig 4.

Precision: 0.88
Recal: 0.96
F-Measure: 0.92
The classification results were saved in "table50.txt" and uploaded to github.

```
title,  pred, act
A Look Back at Kevin Durant's , basketball, basketball
Lionel Messi's Recent Soccer S, other, other
12 (Or Is It 14?) Straight Tie, other, other
Hot and Cold Starts to Current, baseball, baseball
Louisville's Hancock and Michi, basketball, basketball
Aftermath of Heat Streak, basketball, basketball
NHL: Penguins Win 14th Straigh, hockey, hockey
Miami Heat Winning Streak Ende, basketball, basketball
Comparing This Year's Miami He, basketball, basketball
Heat Keeps Winning, Nuggets Lo, basketball, basketball
Bizarre Game Sends Heat Win St, basketball, basketball
Miami Again Turns Up Heat in F, basketball, basketball
Heat Uses 28-4 Spurt vs. Toron, basketball, basketball
Texas Tech Pitcher Trey Masek:, baseball, baseball
LeBron James's Minutes Played , basketball, basketball
Heat Wave Has Miami Winning St, basketball, basketball
Blackhawks' Undefeated Streak , hockey, hockey
Hot 3-Point Shooting in Housto, basketball, basketball
Hot Hand Meta-Analysis Publish, other, other
Northern Illinois's Shooting G, basketball, basketball
Titans' Unusual Streak of Four, other, other
Bobcats End Losing Streak, Cli, None, basketball
Tennessee Men's Three-Point Sh, basketball, basketball
Texas Tech Football Defense No, other, other
```

Figure 3: Part of the results of the 50 entries the classifier was tested on.

```
def assess(f, num):
  tp, fp, fn = 0, 0, 0
  for i in range(num, 100):
    if(f.entries[i].pred == f.entries[i].cls):
      tp +=1
    if(f.entries[i].pred != f.entries[i].cls):
      fp +=1
    if(not f.entries[i].pred):
      fn +=1

  precision = float(tp)/(tp + fp)
  recall = float(tp)/(tp +fn)

  fmeasure = 2*(precision*recall)/(precision+recall)
  print "\n----------------------------------------"
  print precision, recall
  print fmeasure
  print "----------------------------------------\n"

glTrain(f, cat, num, cl)
```

Figure 4: The assess function that calculates precision, recall and f-measure

Problem 3: Training and Testing (90-10)

The same script used earlier is modifyed to train on the first 90 entries and test on the final 10.
The results calculated for a 90-10 split were saved in "table10.txt".

Precision: 0.6
Recal: 0.86
F-Measure: 0.71

The fact that the 90-10 split performs worse than the 50-50, I beleive has to do with the nature of the entries. Most of the data fall under the category of basketball and baseball. Additionally, most of the entries that belong in the "olympics" category are in the last 10 entries that we test. I feel that we have fallen into one of those exceptions where the training set does not apply nicely to the testing data. A 10-fold cross-validation would probably provide more realistic results.

```
-----
Title:     US Target Shooter Medals in Fifth Straight Olympiad
95
Guess: hockey
Actual: olympics

-----
Title:     Is There a Hot Hand in Olympic Archery?
96
Guess: other
Actual: olympics

-----
Title:     Great Olympic Streaks: Basketball
97
Guess: olympics
Actual: olympics

-----
Title:     Great Olympic Streaks: Gymnastics
98
Guess: olympics
Actual: olympics

-----
Title:     Great Olympic Streaks: Women's Swimming
99
Guess: olympics
Actual: olympics

----------------------------------------
0.6 0.857142857143
0.705882352941
----------------------------------------
```

Figure 5: table10.txt

Figure 6: table10.txt