CS-432/532 Introduction to Web Science: Assignment #9: Document Filtering

Due on Thursday, April 21, 2016

Dr. Michael L. Nelson

Plinio Vargas
pvargas@cs.odu.edu

Contents

List of Figures

Listings

List of Tables

Problem 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 Approach

The blog claimed for this assignment was: `http://www.insidehoops.com/blog/`

A great deal of code implementation for this assignment was obtained from [1]. The textbook used Python 2.7 and SQLite as a working environment. Although it would be easier to use similar platform to complete the work, personal preference was a factor in modifying the working environment using Python 3.4 and PostgreSQL.

Some table schema were also modify to facilitate and expedite data input. For example, [1] had two tables: **feature counts**(fc) and **category count**(cc), where the category field was a VARCHAR field. In our implementation, a new table was added and the field typed was modified to INT. Below is the list and schema of database tables:

category_tb: contains list of all categories

Table 1: Category_tb Schema

Attname	Type
cat_id	integer
description	character varying

attname contains table field names. Type column specifies field type: e.g. integer, date, etc.

training_tb: captures the terms for each of the entries in the blog. It also includes actual category and fisher classifier category.

Table 2: Training_tb Schema

Attname	Type
trn_id	integer
words	character varying
cat_id	integer
fisher_cat	integer
title	character varying

attname contains table field names. Type column specifies field type: e.g. integer, date, etc.

cc: similar to [1] contains category training count for the blog. At difference from [1] category field is an integer type instead of character varying.

Table 3: CC Schema

Attname	Type
category	integer
count	integer

attname contains table field names. Type column specifies field type: e.g. integer, date, etc.

fc: similar to [1] contains feature (or term) counts for entries in the blog. At difference from [1] category field is an integer type instead of character varying.

Table 4: FC Schema

Attname	Type
feature	character varying
category	integer
count	integer

attname contains table field names. Type column specifies field type: e.g. integer, date, etc.

Raw entry data was saved in four files with the following format: www-insidehoops-com-blog-feed-rss2-paged-%s. Subsequent program execution required the utilization of those files in order to maintain data integrity since blog sites are dynamic. Line

Listing 1: Saving Raw Data: Predictor.py

```
    indata = (lambda x: 'www-insidehoops-com-blog-feed-rss2-paged-%s' % x)
    x = 0

    # generating 100 entries from blog
    print('Generating 100 blog entries ...')
    while total_entries < 100:
        x += 1
        title, wc, entries = getwordcounts(indata(x))

        """
        # write raw data
        print('Writing raw data ....')
        file = open(output(x), 'w')
        page = requests.get(url_blog(x)).text
        file.write(page)
        file.close()
        """
```

The inside loop was commented in order to prevent re-writing those files.

1.2 Solution

1.2.1 Selected Blog

`http://www.insidehoops.com/blog/`

1.2.2 Blog Entries Classification

Table 5: Category_tb Data

Cat_id	description
1	roster
2	awards
3	injury
4	record setting
5	advice
6	strategy
7	staff
8	standings

cat_id is an integer field linking its value to a category description. Description column is self-explained.

Problem 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 fisherprob() for the actual category.

2.1 Approach

2.1.1 Manually Entry Classification

To accomplish this task a file was created (see Listing 5) containing SQL **UPDATE** for each blog entry. *Predictor.py* (line 96-98 in Listing 2) calls *docclass* library where a function executes the update for each line contained in the classification text file (lines 108-112 Listing 6).

Listing 2: Classification Blog Entry in: Predictor.py

```
96      # set the actual categories for entries
97      print('Updating actual categories ...')
98      c1.set_categories('categories.txt')
```

Listing 3: Update Classification Entry in: docclass.py

```
108   def set_categories(self, file):
109     with open(file, mode='r') as infile:
110       for sql_text in infile:
        self.con.execute(sql_text)
    self.conn.commit()
```

To enter our training data we read from *training_tb* iterating for all rows which id is less than 51. See Listing 4 below:

Listing 4: Entering Training Data: Predictor.py

```
    # train the first 50 entries of 100 blog entries
    print('Training first 50 entries from the blog ....')

    for id, text, category in c1.get_data('select trn_id, words, cat_id from training_tb order by trn_id;'):
        if id < 51:
            print(id, text, category)
            c1.train(text, category)
```

2.2 Solution

2.2.1 Blog Entry Classification

Listing 5: Blog Entry Classification File: categories.txt

```
UPDATE training_tb SET cat_id = 5 WHERE trn_id = 1
UPDATE training_tb SET cat_id = 6 WHERE trn_id = 2
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 3
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 4
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 5
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 6
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 7
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 8
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 9
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 10
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 11
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 12
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 13
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 14
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 15
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 16
UPDATE training_tb SET cat_id = 5 WHERE trn_id = 17
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 18
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 19
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 20
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 21
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 22
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 23
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 24
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 25
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 26
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 27
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 28
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 29
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 30
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 31
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 32
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 33
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 34
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 35
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 36
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 37
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 38
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 39
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 40
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 41
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 42
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 43
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 44
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 45
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 46
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 47
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 48
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 49
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 50
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 51
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 52
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 53
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 54
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 55
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 56
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 57
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 58
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 59
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 60
UPDATE training_tb SET cat_id = 5 WHERE trn_id = 61
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 62
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 63
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 64
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 65
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 66
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 67
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 68
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 69
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 70
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 71
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 72
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 73
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 74
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 75
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 76
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 77
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 78
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 79
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 80
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 81
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 82
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 83
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 84
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 85
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 86
UPDATE training_tb SET cat_id = 8 WHERE trn_id = 87
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 88
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 89
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 90
UPDATE training_tb SET cat_id = 3 WHERE trn_id = 91
UPDATE training_tb SET cat_id = 2 WHERE trn_id = 92
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 93
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 94
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 95
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 96
UPDATE training_tb SET cat_id = 1 WHERE trn_id = 97
UPDATE training_tb SET cat_id = 7 WHERE trn_id = 98
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 99
UPDATE training_tb SET cat_id = 4 WHERE trn_id = 100
```

2.2.2 Fisher Method Results

Table 6: Fisher-Result

Title	Predicted Category	Actual Category	Fisherprob()
Charlie Villanueva suggests Russell Westbrook take his dancing to a club	advice	advice	1.0000
Matt Barnes says Grizzlies need to change gameplan	strategy	strategy	1.0000
Jamal Crawford wins 2015-2016 NBA Sixth Man of Year award	awards	awards	0.9844
Suns keep Earl Watson as head coach	staff	staff	0.9170
Chris Kaman may play increased playoff role for Blazers	standings	standings	0.9846
Nets hire Kenny Atkinson as head coach	staff	staff	0.8642
Kawhi Leonard wins 2015-2016 NBA Defensive Player of Year award	awards	awards	0.7989
D-League hires Brad Walker as head of basketball operations	staff	staff	0.9544
J.J. Barea will miss Mavs-Thunder Game 2	injury	injury	0.9290
Thunder player health update	injury	injury	0.9410
Cavs win Game 1 vs Pistons	standings	standings	0.9508
J.J. Barea health in question for Game 2	injury	injury	0.8392
Frank Vogel, Doc Rivers named NBA Coaches of Month for April, 2016	awards	awards	0.9955
Norman Powell, Karl-Anthony Towns named NBA Rookies of Month for April 2016	awards	awards	0.9988
LeBron James, James Harden named NBA Players of Month for April 2016	awards	awards	0.9886
Nuggets announce front office and basketball title changes	staff	staff	0.9254
Nuggets promote Herb Livsey to Lead Scout	advice	advice	0.9994
Washington Wizards not keeping Randy Wittman as head coach	staff	staff	0.9696
Sacramento Kings not keeping George Karl as head coach	staff	staff	0.9539
Ryan Gomes named D-League Impact Player of the Year	awards	awards	0.8780
Pistons sign Lorenzo Brown	roster	roster	0.6605
Cleveland Cavaliers sign Dahntay Jones	roster	roster	0.9448
Very rough Nets season almost over	standings	standings	0.9785
Phoenix Suns buy Bakersfield Jam D-League team	staff	staff	0.9720
Heat sign Dorell Wright	roster	roster	0.9707
Tristan Thompson now Cavs starting center	roster	roster	0.3987
Quinn Cook wins D-League Rookie of Year award	awards	awards	0.8270
DeMarre Carroll feels he is ready for more minutes	roster	roster	0.7733
Jrue Holiday has surgery for orbital wall fracture	injury	injury	0.9706
NBA records set by the Warriors so far	record setting	record setting	0.9841
Big night in NBA West Playoff race	record setting	record setting	0.8809
Boston Celtics sign guard John Holland	roster	roster	0.7994
Paul Millsap, Karl-Anthony Towns named NBA Players of Week	awards	awards	0.9770
Warriors now have 72 wins	record setting	record setting	0.9925
Miami Heat sign Briante Weber	roster	roster	0.9951
Sixers hire Bryan Colangelo as President of Basketball Operations	staff	staff	0.9931
Pelicans sign James Ennis for rest of season	roster	roster	0.7905
Brandon Knight has surgery for sports hernia	injury	injury	0.8316
Warriors achieve 70-win NBA season, and counting	record setting	record setting	0.9930
Notes from around the NBA: April 7	standings	standings	0.9719
Spurs set NBA record for home wins	record setting	record setting	0.8209
Historic Spurs vs Warriors matchup tonight	record setting	record setting	0.9793
Sixers sign Christian Wood for rest of season	roster	roster	0.8354
Grizzlies waive Ryan Hollins, sign Xavier Munford	roster	roster	0.8771
Alec Burks should return soon for Jazz	roster	roster	0.8069
Chemistry between Derrick Rose and Jimmy Butler can use improvement	roster	roster	0.6476
Emmanuel Mudiay will likely play for Nuggets summer league team	roster	roster	0.8421
Unclear if Kristaps Porzingis will play again this season	roster	roster	0.6691
Josh Richardson, Karl-Anthony Towns named NBA Rookies of Month for March	awards	awards	0.9998
LeBron, Westbrook named NBA Players of Month for March	awards	awards	0.9343
Security shakeup at Pelicans arena	injury	staff	0.2303
Blake Griffin set to return for Clippers	roster	roster	0.1440
Danilo Gallinari will not return this season	roster	injury	0.0070
NBA notes from around the league, March 31	awards	record setting	0.0135
DeAndre Jordan is above 70Detroit Pistons on a hot streak	standings	standings	0.1395
Spurs set an NBA record for home wins	record setting	record setting	0.1306
Grizzlies sign Jordan Farmar for rest of season	roster	roster	0.6814
Rodney Hood got picture and autograph from Kobe Bryant after game	standings	standings	0.2943
Nic Batum gets triple-double vs Sixers	roster	record setting	0.0270
Carmelo Anthony speaks about kid running onto court to hug him	roster	advice	0.1346
Seth Curry getting more minutes from Kings	roster	standings	0.0044
Unclear if DeMarre Carroll will play again this season	roster	roster	0.4020
Pelicans announce injuries to Ryan Anderson, Norris Cole, Alonzo Gee, Jrue Holiday	injury	injury	0.1343
Giannis Antetokounmpo may play point guard next season for Bucks	roster	roster	0.2881
Hassan Whiteside big for Heat off bench	roster	roster	0.0477
LeBron James, Klay Thompson named NBA Players of Week	awards	awards	0.4563
D.J. Stephens named D-League Performer of Week	awards	awards	0.3405
Historical Warriors NBA season continues	record setting	record setting	0.7068
LeBron James keeps climbing NBA all-time historical lists	standings	record setting	0.0308

Values were obtained from running Predictor.py

Problem 2 continued on next page. . .

Table 7: Fisher-Result Cont..

Title	Predicted Category	Actual Category	Fisherprob()
Raptors going for win No. 50	record setting	standings	0.0047
Paul Millsap gets stitches for cut on head	roster	standings	0.0003
Pistons sign Lorenzo Brown to second 10-day contract	roster	roster	0.2508
Sixers waive Sonny Weems, sign Christian Wood to 10-day contract	roster	roster	0.7356
Blake Griffin begins serving suspension	roster	injury	0.0325
Grizzlies sign Xavier Munford to second 10-day contract	roster	roster	0.2127
Nets sign Henry Sims to second 10-day contract	roster	roster	0.2287
Chandler Parsons undergoes season-ending knee surgery	injury	injury	0.5346
Pelicans sign Jordan Hamilton to 10-day contract	roster	roster	0.2855
Nuggets sign Axel Toupane to multi-year contract	roster	roster	0.6341
Meyers Leonard having season-ending shoulder surgery	roster	injury	0.2155
Huge James Harden season stats also include a ton of turnovers	roster	standings	0.0440
Alan Anderson day-to-day with groin injury	roster	injury	0.0235
Jazz enjoy big win road vs Rockets	record setting	standings	0.0105
Help coming soon for the Pistons	standings	roster	0.0098
Festus Ezeli ahead of schedule in knee rehab	roster	injury	0.0496
Huge Jazz-Rockets game tonight	record setting	standings	0.0030
Timberwolves sign Greg Smith for remainder of season	roster	roster	0.5951
No reason to think Phil Jackson will eventually coach the Knicks	roster	staff	0.3192
Jahlil Okafor undergoes knee surgery	roster	injury	0.3707
Chandler Parsons out with knee injury	roster	injury	0.0172
Terrico White named D-League Performer of Week	awards	awards	0.4837
Notes from around the NBA: March 21	awards	record setting	0.0211
Lots of NBA triple-doubles this season	roster	record setting	0.0038
Justise Winslow aiming for better 3-point shot	roster	roster	0.5353
Pistons guard Jodie Meeks remains out	roster	roster	0.2034
Grizzlies sign Jordan Farmar to 10-day contract	roster	roster	0.3493
Kings co-owner lists house for $35 million	roster	staff	0.3687
Stephen Curry jokes about awful game vs Spurs	record setting	record setting	0.0629
Draymond Green wants Warriors to beat Bulls all-time season record	record setting	record setting	0.3840

Values were obtained from running Predictor.py

Problem 3

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

3.1 Approach

We use formulation from [2] to solve this problem. First, *Predictor.py* retrieves all categories (line 108-149), and then for each category initializes True Positive (TP) and False Positive (FP) values.

Listing 6: Calculating Precision and Recall: Predictor.py

```
108     # determine best class
109     tp = {}
110     fp = {}
111     cat_description = []
112 
113     # get all categories
114     cat_list = c1.get_data(’select cat_id, description from category_tb order by
            cat_id;’)
115     for cat in cat_list:
```

```
116         cat_description.append(cat[1])
117         cat = cat[0]
118         tp[cat] = 0
119         fp[cat] = 0
120 
121     print('Determining best class using Fisher classifier ....')
122     true_category = []
123     for category in c1.get_data('select cat_id from training_tb order by trn_id;')
        :
124         true_category.append(category[0])
125 
126     outfile = open('fisher-result.txt', 'w')
127     outfile.write('\\begin{table}[!htbp]\n')
128     outfile.write('\\caption{Fisher-Result} \label{tab:fisher-result}\n')
129     outfile.write('\\begin{center}\n')
130     outfile.write('\\vspace{-5mm}\n')
131     outfile.write('\\begin{tabular}{ l l l l}\n')
132     outfile.write('\\hline\n')
133     outfile.write('Title & Predicted Category & Actual Category & Fisherprob()
        \\\\\ \n')
134     outfile.write('\\hline\n')
135 
136     # initialize confusion matrix
137     c_matrix = [[0 for i in range(8)] for j in range(8)]
138     k = 0
139     for title in all_titles:
140         k += 1
141         fisher_cat = c1.classify(" ".join(wordcounts[title]))
142         c1.update_fisher_cat(k, fisher_cat)
143         print(k, title, true_category[k - 1], fisher_cat)
144         if fisher_cat == true_category[k - 1]:
145             c_matrix[fisher_cat - 1][fisher_cat - 1] += 1
146             tp[fisher_cat] += 1
147         else:
148             fp[fisher_cat] += 1
149             c_matrix[true_category[k - 1] - 1][fisher_cat - 1] += 1
```

Next, an array with the true classification values are kept in object *true_category* (lines 122-124). We used a matrix (confusion matrix) to calculate precision and recall. We also write the values for Table 6 and Table 7 in lines 123-134 and inside of the loop lines 151-152 as the fisher probability is being calculated.

Problem 3 continued on next page. . .

3.2 Solution

Table 8: Precision vs Recall

Cat	TP	FP	Precision	FN	Recall	F-measure
Roster	29	14	0.6744	30	0.9667	0.7945
Awards	13	3	0.8125	13	1.0000	0.8966
Injury	7	1	0.8750	14	0.5000	0.6364
Record Setting	10	3	0.7692	16	0.6250	0.6897
Advice	2	0	1.0000	3	0.6667	0.8000
Strategy	1	0	1.0000	1	1.0000	1.0000
Staff	8	1	0.8889	11	0.7273	0.8000
Standing	6	2	0.7500	12	0.5000	0.6000

Values were obtained from running Predictor.py

In the ideal world, precision and recall values will be close to 1. Given that we have a very small amount of training data (50 entries) there were not sufficient data to infer a good prediction. It is noted that the most frequent categories: **Roster** and **Awards** have high recall values. The latest one has also a high precision.

Problem 4

Redo the questions above, but with the extensions on slide 27 and pp. 136--138.

4.1 Approach

4.2 Solution

Problem 5 - Extra Credit

A 1:1 split for training:test data typically not a good split; 5:1 or even 10:1 is preferable. We also typically use something called “10-fold cross validation” to make sure we spread the training out and don’t “overfit” on a particular sequence of training data.
Rerun questions 2 & 3, but manually classifying all 100 documents, then using 90 for training and 10 for testing. Use 10-fold cross validation and generate the table from Q2, but this time with the average of all 10 values. What was the change, if any, in precision and recall (and thus F-Measure)?

5.1 Approach

5.2 Solution

References

[1] Segarn, Toby. Programming Collective Intelligence. *Building Smart Web 2.0 Application.* (pp 29-53). Sebastopol, CA: O'Reilly Media.

[2] Text Mining, Analytics & More. (n.d.) Retrieved April 21, 2016, from http://www.text-analytics101.com/2014/10/computing-precision-and-recall-for.html