A9 Submission

Breon Day

Cs432

Spring 2017

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:

The blog i choose for this assignment was https://www.theguardian.com/books/rss i felt the category of books and literature would be easy enough to classify, my chosen classifications being **Social Commentary, Ranking and Comparisons, Author Spotlight, Review, News, Other.**
I parsed the blog and generated titles summaries and a combination of the two for training data

```
for e in d.entries:
#    numEntries=numEntries+1

    if 'summary' in e:
        summary = e.summary
    else:
        summary = e.description



    parsedtitle, parsedsummary, combinedsumtitle =parsePost(summary,e.title)

    outFile1.write(parsedtitle)
    outFile1.write('\n')
    outFile2.write(parsedsummary)
    outFile2.write('\n')
    outFile3.write(combinedsumtitle)
    outFile3.write('\n')
```

For creation of the table i took my created titles and uploaded them into a google sheets classified them as ground truth and created a pdf from the title/ classifications
Full view:
https://github.com/BreonDay/cs532-s17/blob/master/Submissions/A9/Src/Question1/Ground%20Truth%20Data%20-%20ground%20truth.pdf

Titles	Classifications
Tory MP's complaint that prize for writers of colour was unfair to whites dismissed	News
The story behind F Scott Fitzgerald's lost short stories	Author Spotlight
The Correspondence by JD Daniels Review – blackly comic verve	Review
'What is a heart? You have an organ in your body and you have a symbol of love'	Author Spotlight
The Courage of Hopelessness by Slavoj Žižek Review – how the big hairy Marxist would change the world	Review
James Patterson writing true-crime book about Aaron Hernandez	News
Frequent readers make the best lovers, say dating-app users	Other
White Tears by Hari Kunzru Review – a satire of cultural appropriation	Review
Make it your hone: the ebook that you are forced to edit as you read	Other
Why Brit crime fiction is paying international dividends	Other
How eBooks lost their shine: 'Kindles now look clunky and unhip'	News
Beano legend Leo Baxendale dies aged 86	News
Screen fatigue' sees UK ebook sales plunge 17% as readers return to print	News
New William Gibson novel set in a world where Hillary Clinton won	Review
International prize for Arabic fiction goes to Mohammed Hasan Alwan	Author Spotlight
Robert Pirsig: Zen and the Art of Motorcycle Maintenance author dies aged 88	News
Wellcome science book prize goes to story of a heart transplant	News
Author Kuki Gallmann shot by raiders on her ranch in Kenya	News
Not just William: Richmal Crompton's adult fiction republished	Author Spotlight
Bill O'Reilly's publisher stands by him after Fox sacking	News
Proust's complaint about neighbours' loud sex among treasures in French sale	News
Stella prize 2017: Heather Rose's The Museum of Modern Love wins award	Author Spotlight
Bana Alabed, seven-year-old Syrian peace campaigner, to publish memoir	Author Spotlight
From Atwood's assault to Pynchon's paper bag: the best author cameos	Ranking And Comparisons
The age of anxiety: what does Granta's best young authors list say about America?	Social Commentary
Cannery Row may be sentimental but it is far from shallow	Review
Tips, links and suggestions: what are you reading this week?	Other
Poem of the week: In the Evening by Anna Akhmatova	Other
Enough David Foster Wallace, already! We need to read beyond our bubbles	Social Commentary
Empty satire: the regrettable rise of blank-paged books in the Trump era	Social Commentary
John Steinbeck's Tortilla Flat is not for 'literary slummers'	Author Spotlight
Call me British, American, Jewish, Londoner – just don't call me patriotic | Will Self	Social Commentary
Plath's letters probably won't harm Hughes's reputation | Rafia Zakaria	News
A pint of Sarah Perry, please: the literary food tie-ins we want to try	Ranking And Comperisions
The new age of Ayn Rand: how she won over Trump and Silicon Valley	News
The riddle of Donald Trump: how a man of few words reached the pinnacle of power	Social Commentary
Don't say divorce, say special relationship: the thorny language of Brexit	Social Commentary
Can you judge a book by its odour?	Other
Why Ruby Tandoh has been cooking up a storm	Other
Lose the plot: why you should skip to the end of books	Other
Neil Gaiman on American Gods, Norse Mythology and more – books podcast	Review
Durga Chew-Bose: 'I don't really believe in writing as catharsis'	Author Spotlight
Priestdaddy by Patricia Lockwood Review – a dazzling comic memoir	Review
The 100 best nonfiction books of all time: No 64 – Walden by Henry David Thoreau (1854)	Ranking And Comperisions
Top 10 terrible houses in fiction	Ranking And Comperisions
Translating Agatha Christie into Icelandic: 'One clue took 10 years'	Author Spotlight
The Nordic Guide to Living 10 Years Longer by Dr Bertil Marklund – digested read	Other
Dava Sobel: 'If you enjoy detective mysteries, you would love rummaging through archives'	Author Spotlight
Primo Levi's If This is a Man at 70	Review
The Mesmerist by Wendy Moore Review – the doctor who put London in a trance	Review
Move Fast and Break Things by Jonathan Taplin Review – the damage done by Silicon Valley	Review
Hostage by Guy Delisle Review – held captive by every frame	Review
The Good Bohemian: The Letters of Ida John Review – the Bloomsbury group laid bare	Review
The Shortest History of Germany Review – probing an enigma at the heart of Europe	Review
East London Review – a journey through a smartphone lens	Review
Strange Labyrinth by Will Ashon Review – summoning the spirits of Epping Forest	Review
Deviate: The Science of Seeing Differently by Beau Lotto Review – why we need brain control	Review

2. Train the Fisher classifier on the first 50 entries (the "training set"), then use the classifier to guess the classification of the next 50 entries (the "test set"). Create a table with 50 rows, like title actual predicted ----- ------ --------- Donnie Iris - 80s 80s "Ah! Leah!" (Forgotten Song) Black Sabbath - metal metal "Vol. 4" (LP Review) Catherine Wheel - alternative metal "Ferment" (LP Review) Assess the performance of your classifier in each of your categories by computing precision, recall, and F-measure.

Training was handled by my fisher.py program

First i opened the files containing my titles summaries combinedtitlesummaries and my newly created ground truth classification file

```python
# coding: utf-8
import docclass

#open file where titles are stored
outFile1= open ('titles3.txt','r')
#open file where categories are stored
outFile2=open('categories2.txt','r')
outFile3 = open('combinedsumtitle3.txt', 'r')



count=0
titles=[]
categories=[]
predictions=[]
amountTrained=0
summaryTitles=[]
trainingCount1=50
trainingCount2=90
maxTrainingData=100
remainingClassifications=0

#populate lists with files

for entry in outFile1:
    titles.append(entry)
for entry in outFile2:
    categories.append(entry)
for entry in outFile3:
    summaryTitles.append(entry)
```

Training was then handled by denoting the amount you wanted to train i have one commented out at all times Namely 50 and 90 this setup gave me the data for questions 2 and 3

```
#summaryTitles
cl=docclass.fisherclassifier(docclass.getwords)
#delete min.db file in project after every run
cl.setdb('mln.db')


#train the first   entries in the summaryTitle txt file
for entry in summaryTitles:
    #make sure one or the other is commented out at all times
    #train first 50
    if count<trainingCount1:
    #train first 90
    #if count < trainingCount2:
        docclass.mytraining(cl,entry,categories[count])
        count+=1

# classify the remaing x entries
predictionsToDO=maxTrainingData-count
predictionsDone=0
while count<maxTrainingData:
    print count
    prediction=cl.classify(summaryTitles[(count)])
    predictions.append(prediction)
    count+=1

# check the results
#while predictionsDone  <predictionsToDO:
#    print(predictionsDone)
#    print(predictionsToDO)
 #   print(maxTrainingData-predictionsDone)
#    print titles[(maxTrainingData-predictionsDone)-1]
 #   print predictions[(predictionsDone)]
 #   predictionsDone+=1

#open file where title+summarys are stored
if len(predictions)==50:
    outFile4 = open('50predictions2.txt', 'wb')
if len(predictions)==10:
    outFile5 = open('10predictions2.txt', 'wb')

print len(predictions)
for prediction in predictions:
    if len(predictions)==50:
        outFile4.write(prediction)
    elif len(predictions)==10:
        outFile5.write(prediction)
```

From the slides and the stack overflow information i was able to use my data to compute precision recall and thus f measure
Q2 macro fmeasure 50 predictions

	true positives	false positive	false negatives	precision	recall	f-measure
news	2	11	1	0.1538461538	0.6666666667	0.25
author spotlight	0	0	5	0	0	0
review	10	2	27	0.8333333333	0.2702702703	0.4081632653
Ranking and comparison	0	0	2	0	0	0
social commentary	1	5	1	0.1666666667	0.5	0.25
other	0	18	1	0	0	0
macro	2.6	3.6	7.2	0.2307692308	0.2873873874	0.1816326531

Q3 macro f measure 10 predictions

	true positives	false positive	false negatives	precision	recall	f-measure
news	0	0	2	0	0	0
author spotlight	0	0	3	0	0	0
review	2	8	0	0.2	1	0.3333333333
Ranking and comparison	0	0	0	0	0	0
social commentary	0	0	2	0	0	0
other	0	0	1	0	0	0
macro	0.3333333333	1.333333333	1.333333333	0.03333333333	0.1666666667	0.05555555556

Strangely enough nearly all metrics decreased as the training data increased i can only guess it would be due to the large number of reviews that were in the later half of the training data and it caused my trainer to get greedy and attempt to only classify reviews

4. Rerun question 3, but with "10-fold cross validation". What was the change, if any, in precision and recall (and thus F-Measure)?

I handled 10-fold cross validation in Cross.py
Similiar to fisher.py i created the lists from my combined summary titles and ground truth categories
Then i split them based on a value n creating sublists from each at n*10 and n-1*10 and my parts to be classified in a similiar vein then i simply ran it n times

```
count=0

count2=0
#change this variable every one to get the values 1=0-9 10=90-99
n=10

filename="CrossValidation" + str(n*10)+".txt"
outFile4=open(filename,"wb")
cl=docclass.fisherclassifier(docclass.getwords)
#delete min.db file in project after every run
cl.setdb('mln2.db')

#create cross 10 sublists
summariesToBeClassified=  summaryTitles[((n-1)*10):(n*10)]
sublist2=  summaryTitles[:(n-1)*10]
sublist3=  summaryTitles[(n)*10:]
trainingdata_Entries= sublist2+sublist3

#
categoriesToBeClassified=  categories[((n-1)*10):(n*10)]
sublist5=  categories[:(n-1)*10]
sublist6=  categories[(n)*10:]
trainingdata_Categories= sublist5+sublist6

# train it based on what was given
#for entry in trainingdata_Entries:
#    if count < 90:
#        docclass.mytraining(cl, entry, trainingdata_Categories[count])
#        count += 1
while count < 90:
        docclass.mytraining(cl, trainingdata_Entries[count], trainingdata_Categories[count])
        count += 1
count =0
```

Q4 fmeasure average

average	precision	recall	f-measure
news	0	0	0
author spotlight	0.2	0.1	0.1333333333
review	0.4691666667	0.9	0.6167985393
Ranking and comparison	0	0	0
social commentary	0	0	0
other	0.1	0.05	0.06666666667
macro	0.1281944444	0.175	0.1361330899

And as we can see above the macro value .136 lies between the 50 prediction

	true positives	false positive	false negatives	precision	recall	f-measure
news	2	11	1	0.1538461538	0.6666666667	0.25
author spotlight	0	0	5	0	0	0
review	10	2	27	0.8333333333	0.2702702703	0.4081632653
Ranking and comparison	0	0	2	0	0	0
social commentary	1	5	1	0.1666666667	0.5	0.25
other	0	18	1	0	0	0
macro	2.6	3.6	7.2	0.2307692308	0.2873873874	0.1816326531

And the 10 prediction

	true positives	false positive	false negatives	precision	recall	f-measure
news	0	0	2	0	0	0
author spotlight	0	0	3	0	0	0
review	2	8	0	0.2	1	0.3333333333
Ranking and comparison	0	0	0	0	0	0
social commentary	0	0	2	0	0	0
other	0	0	1	0	0	0
macro	0.3333333333	1.333333333	1.333333333	0.03333333333	0.1666666667	0.05555555556