Assignment 9 Report

Professor Nelson

By Tyler Medina

5/1/17

<div align="center">**Part 1**</div>

1.1 Netflix RSS Feed

I originally claimed the ESPN feed, but I decided to change to a Netflix rss. My original plan was to classify news feeds into their respective sports categories. Categories such as nfl, mlb, nba, nhl, etc. The reason I decided to change rss feeds was that the ESPN feed only contained the 25 most recent stories, and I needed at least 100 rss feeds for this assignment. The Netflix rssfeed I decided to use contains the top 100 movie/series on Netflix. The rss contains links that link to the Netflix dvd page with the film's genre, summary, reviews, and recommendations.

1.2 Categories

For this assignment, the movies/series will be classified into categories based on their genre. These categories are displayed in the figure below:

Categories
Action
Drama
Horror
Comedy
Family

I was debating adding an "other" category, but I determined that these categories would encompass the majority, if not all, of the movies to be classified. This would make the other category relatively useless. The trickiest parts would be categorizing movies that have multiple dramas. For example, a romantic comedy can be a drama and a comedy. The classifier could label this romantic comedy a drama, and although this may technically be correct, a film that focuses on making people laugh should be labeled a comedy.

1.3 Ground Truth

Below is the gold standard model that the classifier will strive to be:

Gold Standard:

Title	Category
Game of Thrones	action
Sully	drama
Star Trek Beyond	action
Ghostbusters	comedy
Jason Bourne	action
The Magnificent Seven	action
Free State of Jones	drama
Money Monster	drama
Hell or High Water	drama
The Legend of Tarzan	action
The Accountant	action
Deepwater Horizon	drama
Hacksaw Ridge	drama
Central Intelligence	comedy
X-Men: Apocalypse	action
Independence Day: Resurgence	action
Manchester by the Sea	drama
Suicide Squad	action
The Jungle Book (2016)	family
Bad Moms	comedy
Now You See Me 2	action
Captain America: Civil War	action
Homeland	drama
The Nice Guys	comedy
The Infiltrator	drama

Movie	Genre
Outlander	drama
Florence Foster Jenkins	drama
Mechanic: Resurrection	action
The Huntsman: Winters War	drama
The Shallows	horror
War Dogs	drama
Finding Dory	family
A Hologram for the King	drama
Miss Peregrines Home for Peculiar Children	family
Blood Father	action
Neighbors 2: Sorority Rising	comedy
Whiskey Tango Foxtrot	comedy
Batman v Superman: Dawn of Justice	action
Eye in the Sky	drama
The Man Who Knew Infinity	drama
The Boss	comedy
Passengers	drama
Dont Breathe	horror
Sausage Party	comedy
Jack Reacher: Never Go Back	action
Warcraft	action
Mothers Day	comedy
The BFG	family
Ben-Hur	action
Captain Fantastic	comedy
Snowden	drama
The Duel	drama
Criminal	action
The Meddler	comedy
Joy	drama
Moonlight	drama
Arrival	drama
Petes Dragon	family
The Finest Hours	action

Petes Dragon	family
The Finest Hours	action
11/22/63	drama
Wild Oats	comedy
Café Society	comedy
Hello, My Name Is Doris	comedy
Mike and Dave Need Wedding Dates	comedy
Miracles from Heaven	drama
The Revenant	drama
13 Hours: The Secret Soldiers of Benghazi	action
Bridge of Spies	drama
Race	drama
The Conjuring 2	horror
Hunt for the Wilderpeople	family
My Big Fat Greek Wedding 2	comedy
Bridget Joness Baby	comedy
The Divergent Series: Allegiant	action
10 Cloverfield Lane	drama
London Has Fallen	action
Alice Through the Looking Glass	family
Genius	drama
The Night Manager	drama
House of Cards	drama
I Am Wrath	action
Teenage Mutant Ninja Turtles: Out of the Shadows	action
Kubo and the Two Strings	family
Love & Friendship	comedy
Our Kind of Traitor	drama
The Purge: Election Year	action
The Hateful Eight	drama
Deadpool	action
Storks	family
Concussion	drama
Ice Age: Collision Course	family

ext file

Ice Age: Collision Course	family
Nerve	drama
Brooklyn	drama
Dark Horse	drama
Keeping Up with the Joneses	comedy
The Intern	comedy

Part 2

2.1 Train First 50 Entries

Feedfilter.py parses the entries in the xml file. I manually enter the actual category and the train() function in docclass.py trains the fisher classifier to predict the category of future inputs.

```python
def classify(self,item,default=None):
    probs={}
    max=0.0
    for cat in self.categories():
        probs[cat]=self.prob(item,cat)
        if probs[cat]>max:
            max=probs[cat]
            best=cat

    for cat in probs:
        if cat==best: continue
        if probs[cat]*self.getthreshold(best)>probs[best]: return default
    return best

class fisherclassifier(classifier):
    def cprob(self,f,cat):
        clf=self.fprob(f,cat)
        if clf==0: return 0

        freqsum=sum([self.fprob(f,c) for c in self.categories()])

        p=clf/(freqsum)

        return p


    def fisherprob(self,item,cat):
        p=1
        features=self.getfeatures(item)
        for f in features:
            p*=(self.weightedprob(f,cat,self.cprob))

        fscore=-2*math.log(p)
```

leng

2.2 Results of 50 training entries

The following figures show the results of the predicted categories, followed by the actual categories:

```
Game of Thrones                             None       action
Sully                                       action     drama
Star Trek Beyond                            action     action
Ghostbusters                                drama      comedy
Jason Bourne                                action     action
The Magnificent Seven                       action     action
Free State of Jones                         action     drama
Money Monster                               drama      drama
Hell or High Water                          action     drama
The Legend of Tarzan                        action     action
The Accountant                              action     action
Deepwater Horizon                           comedy     drama
Hacksaw Ridge                               action     drama
Central Intelligence                        action     comedy
X-Men: Apocalypse                           drama      action
Independence Day: Resurgence                action     action
Manchester by the Sea                       action     drama
Suicide Squad                               action     action
The Jungle Book (2016)                      action     family
Bad Moms                                    action     comedy
Now You See Me 2                            action     action
Captain America: Civil War                  drama      action
Homeland                                    drama      drama
The Nice Guys                               action     comedy
The Infiltrator                             action     drama
Inferno                                     drama      action
The Secret Life of Pets                     action     family
The Girl on the Train                       comedy     drama
Me Before You                               drama      drama
Outlander                                   action     drama
Florence Foster Jenkins                     action     drama
Mechanic: Resurrection                      drama      action
The Huntsman: Winters War                   drama      drama
The Shallows                                drama      horror
```

ext file   length : 9403   line:

Title		
The Shallows	drama	horror
War Dogs	drama	drama
Finding Dory	action	family
A Hologram for the King	action	drama
Miss Peregrines Home for Peculiar Children	drama	family
Blood Father	family	action
Neighbors 2: Sorority Rising	action	comedy
Whiskey Tango Foxtrot	drama	comedy
Batman v Superman: Dawn of Justice	action	action
Eye in the Sky	action	drama
The Man Who Knew Infinity	action	drama
The Boss	action	comedy
Passengers	drama	drama
Dont Breathe	action	horror
Sausage Party	drama	comedy
Jack Reacher: Never Go Back	action	action
Warcraft	family	action

50 automatic classifications

Title	Prediction	Actual
Mothers Day	comedy	comedy
The BFG	family	family
Ben-Hur	action	action
Captain Fantastic	family	comedy
Snowden	drama	drama
The Duel	action	drama
Criminal	drama	action
The Meddler	action	comedy
Joy	action	drama

Title	Genre 1	Genre 2
Joy	action	drama
Moonlight	drama	drama
Arrival	drama	drama
Petes Dragon	family	family
The Finest Hours	drama	action
11/22/63	action	drama
Wild Oats	drama	comedy
Café Society	action	comedy
Hello, My Name Is Doris	drama	comedy
Mike and Dave Need Wedding Dates	family	comedy
Miracles from Heaven	drama	drama
The Revenant	drama	drama
13 Hours: The Secret Soldiers of Benghazi	action	action
Bridge of Spies	comedy	drama
Race	action	drama
The Conjuring 2	comedy	horror
Hunt for the Wilderpeople	family	family
My Big Fat Greek Wedding 2	action	comedy
Bridget Joness Baby	action	comedy
The Divergent Series: Allegiant	drama	action
10 Cloverfield Lane	action	drama
London Has Fallen	action	action
Alice Through the Looking Glass	action	family
Genius	drama	drama
The Night Manager	action	drama
House of Cards	action	drama
I Am Wrath	drama	action
Teenage Mutant Ninja Turtles: Out of the Shadows	action	action
Kubo and the Two Strings	family	family
Love & Friendship	action	comedy
Our Kind of Traitor	comedy	drama
The Purge: Election Year	action	action
The Hateful Eight	drama	drama
Deadpool	action	action

ext file   length : 940

Deadpool	action	action
Storks	family	family
Concussion	drama	drama
Ice Age: Collision Course	action	family
Nerve	drama	drama
Brooklyn	action	drama
Dark Horse	action	drama
Keeping Up with the Joneses	drama	comedy
The Intern	action	comedy

2.3 Analyzing the results

The precision, recall, and F-Measure were calculated for each of the categories in the figure below:

Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

F-Measure = 2 * P*R / (P+R)

Category	Precision	Recall	F-Measure
Action	0.28	0.58	0.38
Drama	0.48	0.38	0.42
Horror	0.00	0.00	0.00
Comedy	0.17	0.05	0.08
Family	0.36	0.45	0.40
Average	0.26	0.29	0.26

The classifier wasn't very accurate after being trained with fifty movies. It's f-measure score was only .40. Part of the reason the score is low is that the horror genre was scarcely represented in the data set and should be viewed as an outlier. There were only two horror films in the 100 movies and none of them were true positives under this category. This brings the average of all the scores down.

Part 3

3.1 Results

The method of training the classifier is similar to above, but we stop at 90 instead of 50 this time:

```
90 manual classifications

Title                               Prediction  Actual
_____
Game of Thrones                     None        action
Sully                               action      drama
Star Trek Beyond                    action      action
Ghostbusters                        drama       comedy
Jason Bourne                        action      action
The Magnificent Seven               action      action
Free State of Jones                 action      drama
Money Monster                       drama       drama
Hell or High Water                  action      drama
The Legend of Tarzan                action      action
The Accountant                      action      action
Deepwater Horizon                   comedy      drama
Hacksaw Ridge                       action      drama
Central Intelligence                action      comedy
X-Men: Apocalypse                   drama       action
Independence Day: Resurgence        action      action
Manchester by the Sea               action      drama
Suicide Squad                       action      action
The Jungle Book (2016)              action      family
Bad Moms                            action      comedy
Now You See Me 2                    action      action
Captain America: Civil War          drama       action
Homeland                            drama       drama
The Nice Guys                       action      comedy
The Infiltrator                     action      drama
Inferno                             drama       action
The Secret Life of Pets             action      family
The Girl on the Train               comedy      drama
Me Before You                       drama       drama
Outlander                           action      drama
```
ext file   length : 89!

Outlander	action	drama
Florence Foster Jenkins	action	drama
Mechanic: Resurrection	drama	action
The Huntsman: Winters War	drama	drama
The Shallows	drama	horror
War Dogs	drama	drama
Finding Dory	action	family
A Hologram for the King	action	drama
Miss Peregrines Home for Peculiar Children	drama	family
Blood Father	family	action
Neighbors 2: Sorority Rising	action	comedy
Whiskey Tango Foxtrot	drama	comedy
Batman v Superman: Dawn of Justice	action	action
Eye in the Sky	action	drama
The Man Who Knew Infinity	action	drama
The Boss	action	comedy
Passengers	drama	drama
Dont Breathe	action	horror
Sausage Party	drama	comedy
Jack Reacher: Never Go Back	action	action
Warcraft	family	action
Mothers Day	comedy	comedy
The BFG	family	family
Ben-Hur	action	action
Captain Fantastic	family	comedy
Snowden	drama	drama
The Duel	action	drama
Criminal	drama	action
The Meddler	action	comedy
Joy	action	drama
Moonlight	drama	drama
Arrival	drama	drama
Petes Dragon	family	family
The Finest Hours	drama	action

ext file   length : 8953   lines : 113

The Finest Hours	drama	action
11/22/63	action	drama
Wild Oats	drama	comedy
Café Society	comedy	comedy
Hello, My Name Is Doris	family	comedy
Mike and Dave Need Wedding Dates	comedy	comedy
Miracles from Heaven	family	drama
The Revenant	action	drama
13 Hours: The Secret Soldiers of Benghazi	action	action
Bridge of Spies	drama	drama
Race	drama	drama
The Conjuring 2	comedy	horror
Hunt for the Wilderpeople	family	family
My Big Fat Greek Wedding 2	comedy	comedy
Bridget Joness Baby	action	comedy
The Divergent Series: Allegiant	drama	action
10 Cloverfield Lane	action	drama
London Has Fallen	action	action
Alice Through the Looking Glass	drama	family
Genius	action	drama
The Night Manager	drama	drama
House of Cards	action	drama
I Am Wrath	drama	action
Teenage Mutant Ninja Turtles: Out of the Shadows	action	action
Kubo and the Two Strings	family	family
Love & Friendship	comedy	comedy
Our Kind of Traitor	drama	drama
The Purge: Election Year	action	action

10 automatic classifications

Title	Prediction	Actual
The Hateful Eight	drama	drama
Deadpool	action	action
Storks	comedy	family
Concussion	drama	drama
Ice Age: Collision Course	action	family
Nerve	drama	drama
Brooklyn	action	drama
Dark Horse	action	drama
Keeping Up with the Joneses	drama	comedy
The Intern	family	comedy

By using the same method as part 2, we can calculate the precision, recall, and f-measure and put it into the table below:

Category	Precision	Recall	F-Measure
Action	0.33	0.59	0.42
Drama	0.46	0.41	0.43
Horror	0.00	0.00	0.00
Comedy	0.36	0.26	0.30
Family	0.29	0.36	0.32
Average	0.29	0.32	0.29

Using 90 movies as training data provides similar results. The difference with this data set, is that the Comedy genre is more represented. There are more false positives, false negatives, and true positives for this category. Although this lowers the scores of the other categories, it increases the average precision, recall, and f-measure. This suggests that the classifier becomes more accurate the more it is trained. Repeating the assignment with a larger data set should confirm this hypothesis.