Assignment 9 Report

Professor Nelson

By Tyler Medina

5/1/17

**Part 1**

**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:

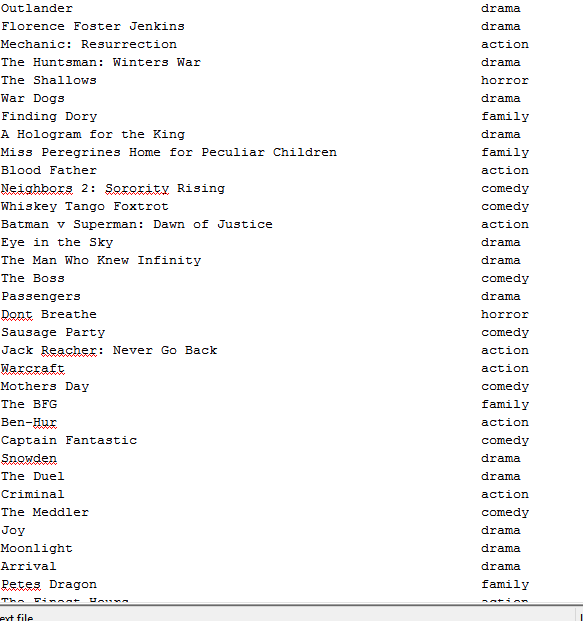
|  |
| --- |
| **Categories** |
| Action |
| Drama |
| Horror |
| Comedy |
| Family |

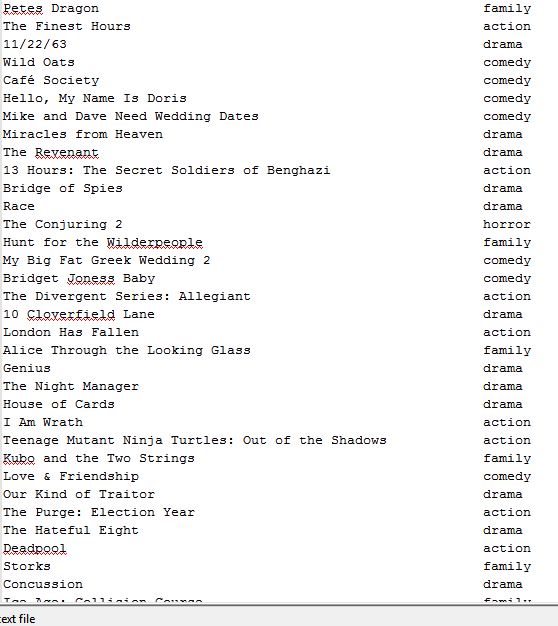
I was debating adding an “other” category, but I determined that these categories would encompass the majority, if not all, of the movies to be classified. This would make the other category relatively useless. The trickiest parts would be categorizing movies that have multiple dramas. For example, a 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.

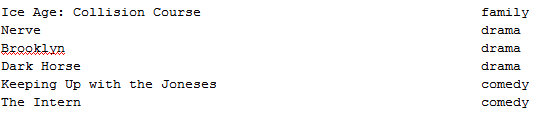
**1.3 Ground Truth**

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





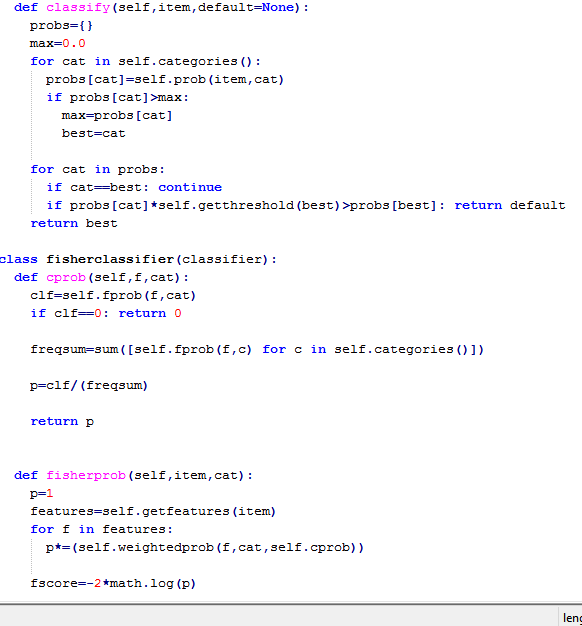




**Part 2**

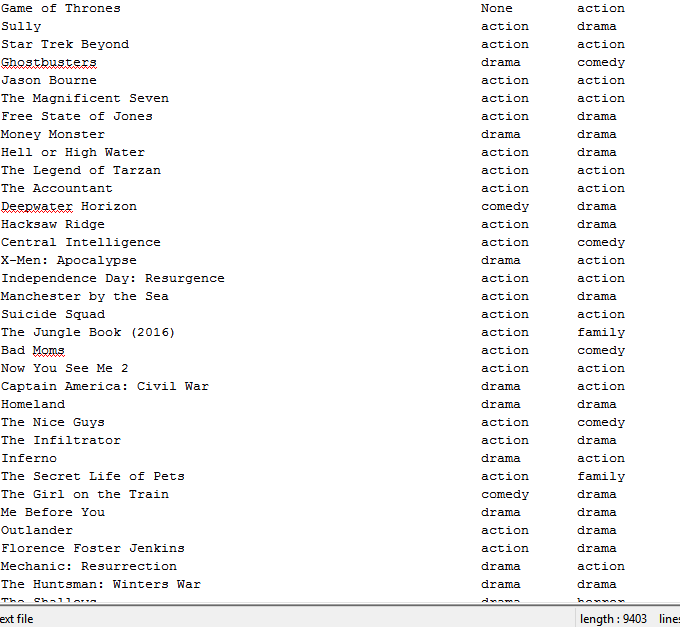
**2.1 Train First 50 Entries**

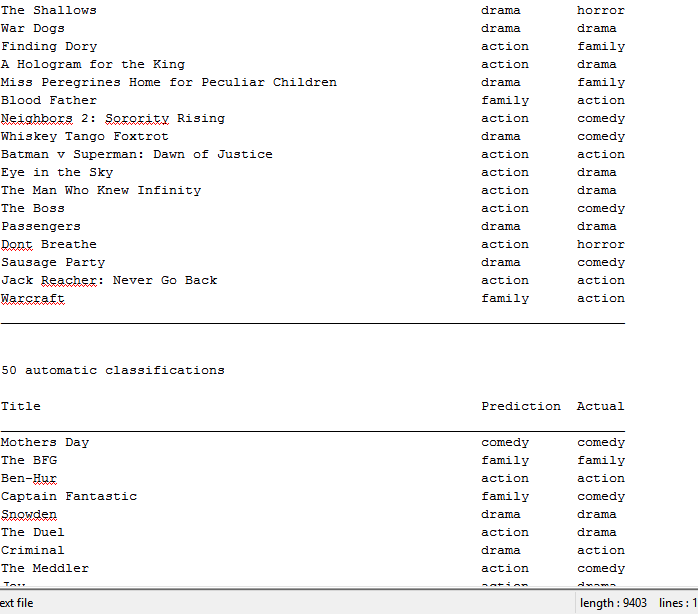
Feedfilter.py parses the entries in the xml file. I manually enter the actual category and the train() function in docclass.py trains the fisher classifier to predict the category of future inputs.



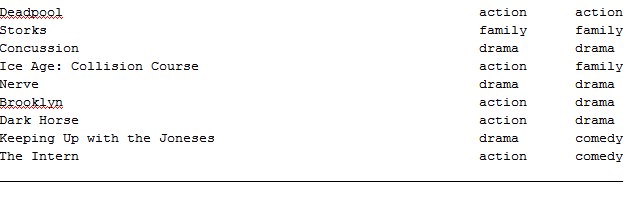
**2.2 Results of 50 training entries**

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









**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)

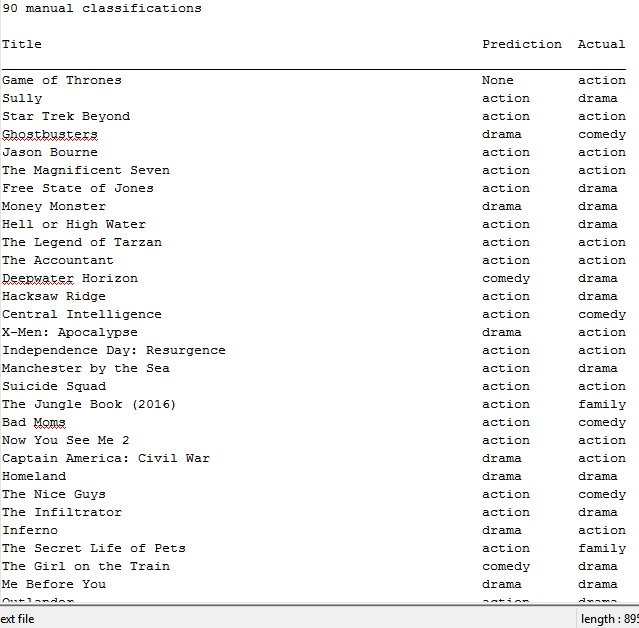
|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Precision** | **Recall** | **F-Measure** |
| Action | 0.28 | 0.58 | 0.38 |
| Drama | 0.48 | 0.38 | 0.42 |
| Horror | 0.00 | 0.00 | 0.00 |
| Comedy | 0.17 | 0.05 | 0.08 |
| Family | 0.36 | 0.45 | 0.40 |
| **Average** | 0.26 | 0.29 | 0.26 |

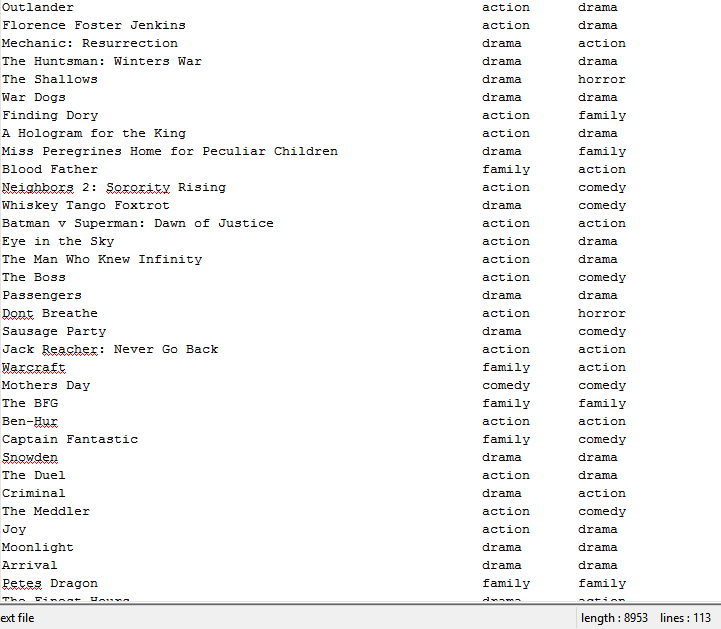
The classifier wasn’t very accurate after being trained with fifty movies. It’s f-measure score was only .40. Part of the reason the score is low is that the horror genre was scarcely represented in the data set and should be viewed as an outlier. There were only two horror films in the 100 movies and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

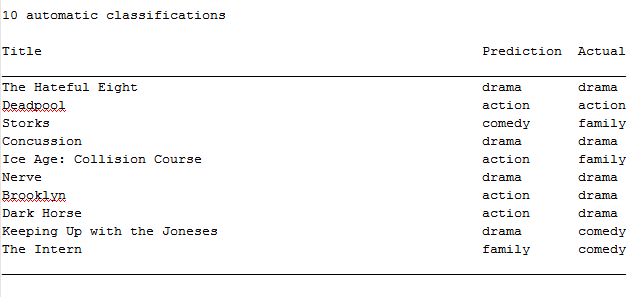
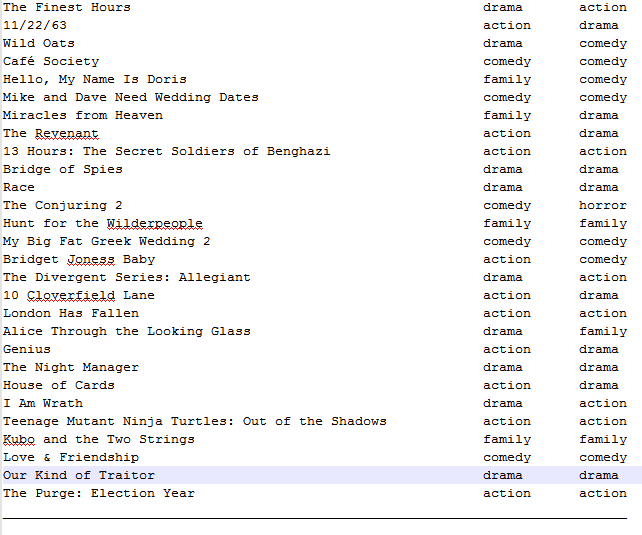
**Part 3**

**3.1 Results**

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

50 this time: 





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.