Assignment 9

CS 432

Miranda SMITH

2017

**Question 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:*

*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. Create a table with 100 rows, like:*

*title classification*

*----- --------------*

*Ric Ocasek - 80s*

*"Something To Grab*

*For" (forgotten song)*

*Weezer - "Pinkerton" alternative*

*(LP Review)*

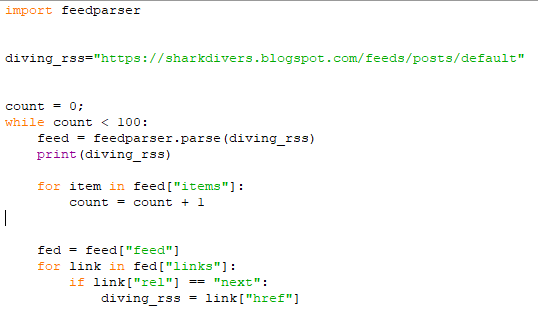
*Schon & Hammer - 80s*

*"No More Lies"*

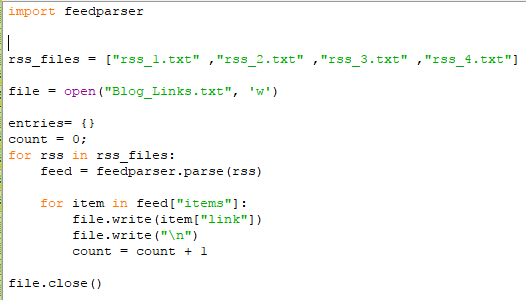
*(forgotten song)*

*etc. This is your "ground truth" (or "gold standard") data.*

I chose the feed from http://sharkdivers.blogspot.com/

I got 4 rss feeds from the ‘get\_rss\_feed.py’ program. It went through each rss feed counting the number of entries and going to the next RSS feed until there were 100 entries. It just printed out the links and I downloaded them to rss\_1.txt, rss\_2.txt, rss\_3.txt and rss\_4.txt.

I went through each RSS file with ‘get\_entry\_links.py’ and printed them out to a file.



I then manually went to each link and categorized the content. The categories went in the file ‘Blog\_links\_Categories.txt’

The categorizes I chose were

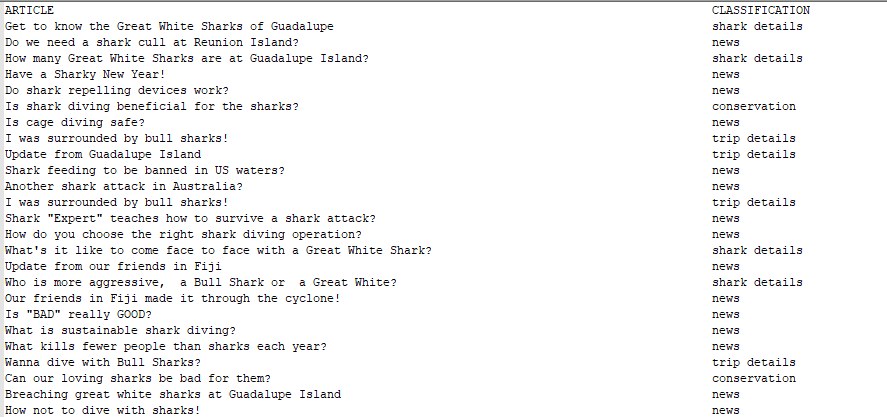
Shark Details

Trip Details

News

Conservation

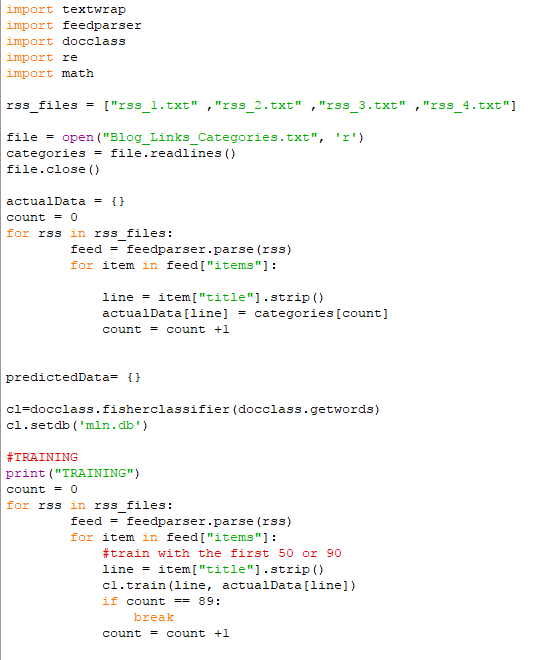
I then created a program ‘getGroundTruthTable.py’ to create the required table. It just linked each category to the correct feed title and formatted the output.

The entirety of the output in in ‘GroundTruthTable.txt’.

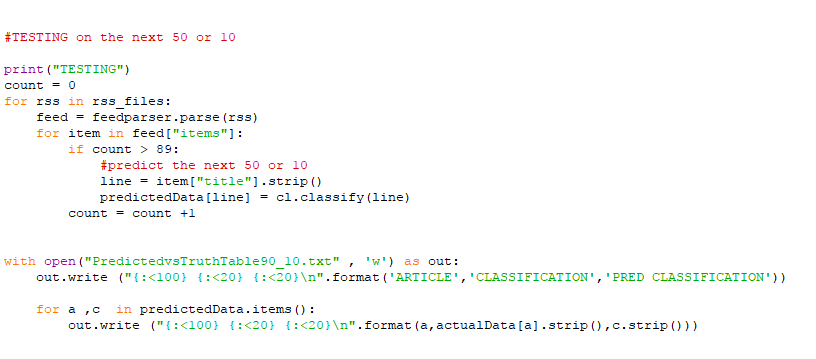
**Question 2 & 3**

*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. Assess the performance of your classifier in each of your categories by computing precision, recall, and F-measure. Use the "macro-averaged" label based method. For example, if you have 5 categories (e.g., 80s, metal, alternative, electronic, cover), you will compute precision, recall, and F-measure for each category, and then compute the average across the 5 categories. Repeat, but use the first 90 entries to train your classifier and the last 10 entries for testing.*

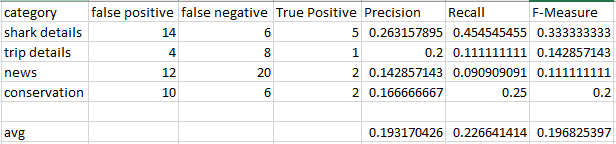
I created the program ‘getPredictedTable.py’ to use the fisher method and create the required table.

I link the category with the entry of the RSS feed. Then train the fisher method with either the first 50 or 90 entries, depending on the run. I only use the title of the entry as the data, because that is how assignment 8 was done.

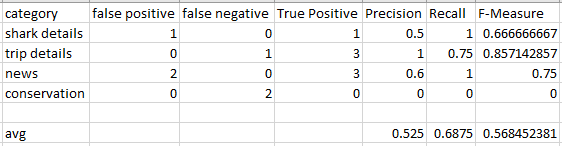
The next part of the program predicts the remaining entries and creates a table. It outputs a file called PredictedvsTruthTable50\_50 or 90\_10



I used an excel work book ‘PredictedvsTruthTable’ to calculate the precision, recall and F-measure for both files.

For the 50-50 split: 

For the 90-10 split:



The 90-10 split is 2.9 times better than the 50 -50 split for F-Measure.