CS475

Assignment 10

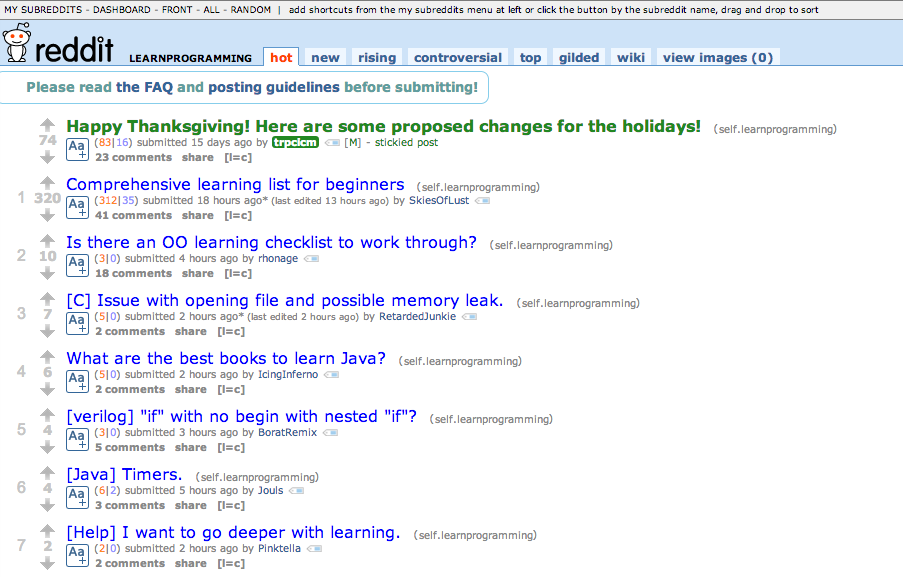
Fall 2013

Onapha Rattanachottiteepakorn

**Question 1**: Choose a blog or a newsfeed (or something similar with an Atom or RSS feed). 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.

For the testing data, I chose the subreddit r/learnprogramming from the website Reddit. At first, I tried to grab the rss feed from a Tumblr blog but the feed only provided 20 entries so I chose Reddit since their API allows max-item limitation. It was easy for me to identify each entry with a label, as the questions tend to be very specific about the language the poster is asking about. The categories included ‘c’, ‘cpp’, ‘csharp’, ‘java’, ‘sql’, ‘python’, and ‘others’. Other posts that ask for general advise for programming were labeled as ‘general’.

The xml file used for this assignment can be found in the same directory as this document.

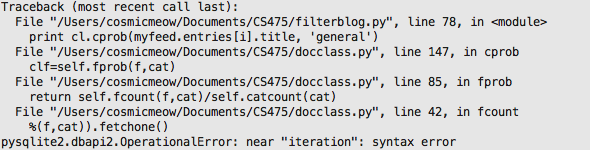


**Figure 1:** r/learningprogramming, filled with questions with specific language attached to most of the post titles

**Question 2**: Manually classify the first 50 entries, and then classify (using the fisher classifier) the remaining 50 entries. Report the cprob() values for the 50 titles as well. From the title or entry itself, specify the 1-, 2-, or 3-gram that you used for the string to classify. Do not repeat strings; you will have 50 unique strings.

For the first 50 entries, I manually classified the title of each of the entries by only looking at the title. The parser could not find entries[i].content from the parsed xml file so I had to classify entries[i].title for each of the entries rather than their contents. : (

I included the cprob function for the entries, however they all printed out as 0 with errors that I couldn’t comprehend. I tried to debug the process by only include one category (general in Figure 2). The program printed out 0 for all entries with the additional errors seen in the figure.

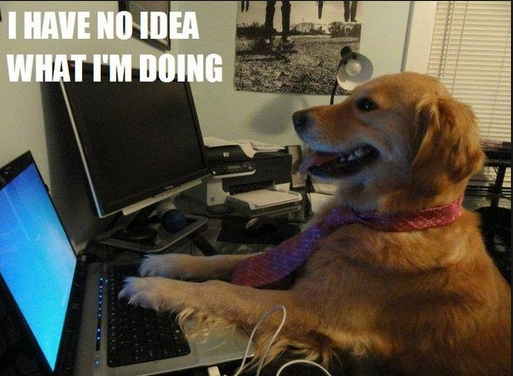


**Figure 2**: Errors from finding the chi probability.

**Question 3**: Assess the performance of your classifier in each of your categories by computing precision and recall. Note that the definitions are slightly different in the context of classification.

This question requires complete knowledge on the assignment and the general Document Classification itself. Reading over the lecture, I gained wonderful knowledge from the great Professor Nelson, whom I am eternally grateful for his amazing class. As far as this semester goes, I have to say that Professor Nelson’s class is the most entertaining and most useful for real work environment.

As for the answer for this class, please see Figure 3 on the next page.

**Figure 3**: This is me.

Happy Holidays.