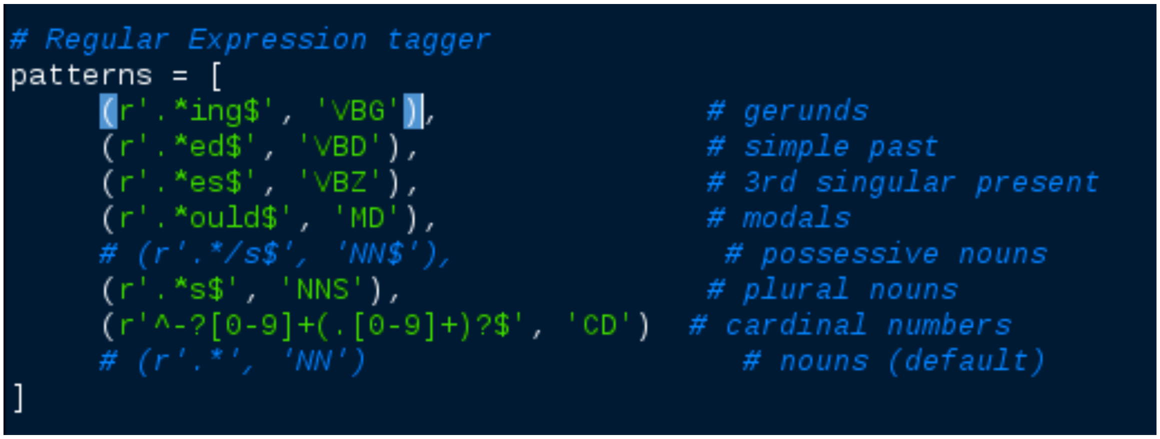
Josh Weeks
CSCI 4140
Project Part 1 Report

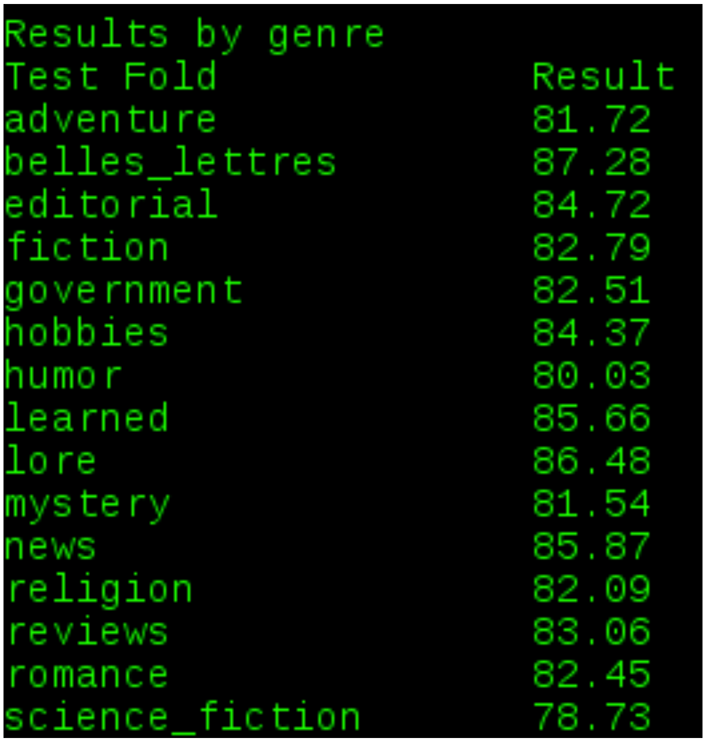
The regular expression tagger that I used looks as follows:



```
# Regular Expression tagger
patterns = [
     (r'.*ing$', 'VBG'),                # gerunds
     (r'.*ed$', 'VBD'),                 # simple past
     (r'.*es$', 'VBZ'),                 # 3rd singular present
     (r'.*ould$', 'MD'),                # modals
     # (r'.*/s$', 'NN$'),                # possessive nouns
     (r'.*s$', 'NNS'),                  # plural nouns
     (r'^-?[0-9]+(.[0-9]+)?$', 'CD')  # cardinal numbers
     # (r'.*', 'NN')                       # nouns (default)
]
```

I was motivated to choose these patterns based on some comparisons of tagged sentences. Here are some examples of sentences that use the part of speech.
An example of a gerund is “Running is good exercise.”
An example of a simple past is “He lived in Fiji in 1976.”
An example of a 3rd singular present is “He walks to the park daily.”
An example of a modal is “I can speak English.”
An example of a plural noun is “The boys were throwing baseballs.”
A cardinal number is a number denoting quantity such as ‘one’, ‘two’, ‘three’, etc.

The genre (category) evaluation ranged from 78.73% to 87.28%. The results from the evaluation were as follows:

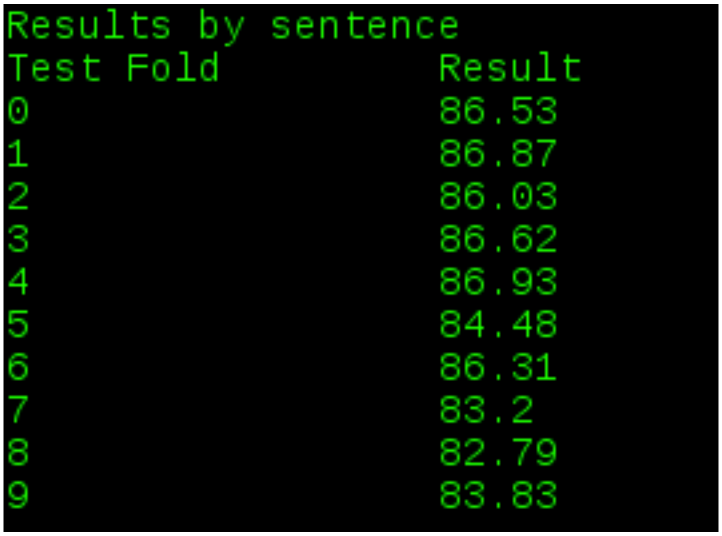


```
Results by genre
Test Fold             Result
adventure             81.72
belles_lettres        87.28
editorial             84.72
fiction               82.79
government            82.51
hobbies               84.37
humor                 80.03
learned               85.66
lore                  86.48
mystery               81.54
news                  85.87
religion              82.09
reviews               83.06
romance               82.45
science_fiction       78.73
```

Science fiction is the lowest percentage because they make up most of the words that they use; some nouns may not show up in the corpus tagger. News had a decent percentage because they try to use grammar correctly and most of the words are easily tagged. If I started to combine the genres together, I believe I could find some interesting trends.

Data split by source (fileid) evaluations took a lot longer to produce results and the results that it did produced were very high. I think this happened because there are 500 file ids. The file ids are set to fold ten times. This causes the categories to be mixed up if you are shuffling the data. A science fiction movie may be have played a role in achieving the news percentage in that situation.

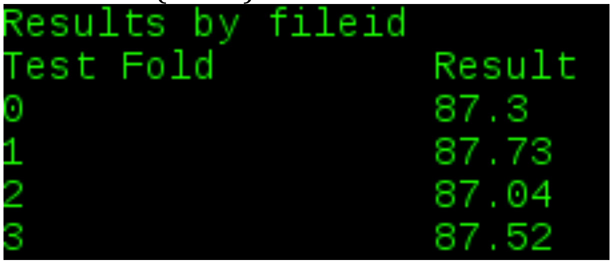
The evaluation results of splitting by sentence are as follows:



```
Results by sentence
Test Fold           Result
0                   86.53
1                   86.87
2                   86.03
3                   86.62
4                   86.93
5                   84.48
6                   86.31
7                   83.2
8                   82.79
9                   83.83
```

Simply splitting the data into sentences in 10 folds shows some good results. From ~82% to ~87%, we see that the evaluations are fairly consistent. It also ran the quickest of the loops. I did not have to search for file ids or categories, which saves some time. I think that the way that I chose to do genre and source made my loops much slower. In the future, I plan to learn how I can make them more efficient.

The source(file id) evaluation results are as follows:



```
Results by fileid
Test Fold           Result
0                   87.3
1                   87.73
2                   87.04
3                   87.52
```

I believe that although the file ids approach takes the longest to run, it is the most accurate. If the data is shuffled then it would not be as effective because some of each category would be in every other category. The results were between 87% and 88%. The other two methods had one fold be near 88% but the average of the file id evaluations was much better than the others. I plan to make this loop much faster in the future because it does give the best results and will be useful in my future.