Project 1 report

Song Li(sol315)

October 27, 2016

1 Task 1

For task 1, I used the nltk library and the sklearn library. Overall, it's very easy to use.The result is:

	Bayes model	Rocchio model
Train Time	0.00406s	0.0147s
Test Time	0.00141s	0.00443s
Accuracy	0.866	0.71
Precision	0.866	0.947
Call back	1.0	0.8602

The basic steps of this task is: parse json file, normalize words, get tokens, get TF-IDF table, train and predict
What's interesting here is, all of the labels generated by Bayes model is 1, so the call back rate is 1.0

2 Task 2

For task 2, Firstly I used the POS tag to tag every reviews, then I get the frequency of every NN and NNS, use them as the attributes.
In the optimization part, I just keep all of the alpha and number letters, all of them are lower letters. What's more, I wanted to use a word list to filter the "useless" NN and JJ, but I can't find a list which really works

3 Task 3

For task 3. At first, I just do a combination of frequent attributes and the JJ appear in the same reviews. Find the most frequent 50s.
For the optimization in this part. I changed the "all" JJ s to the nearest one. Like "This computer is very good and screen is beautiful", I just keep the nearest JJ, so the result will be (computer, good) and (screen, beautiful). This optimization is kind of useful and helps a lot.