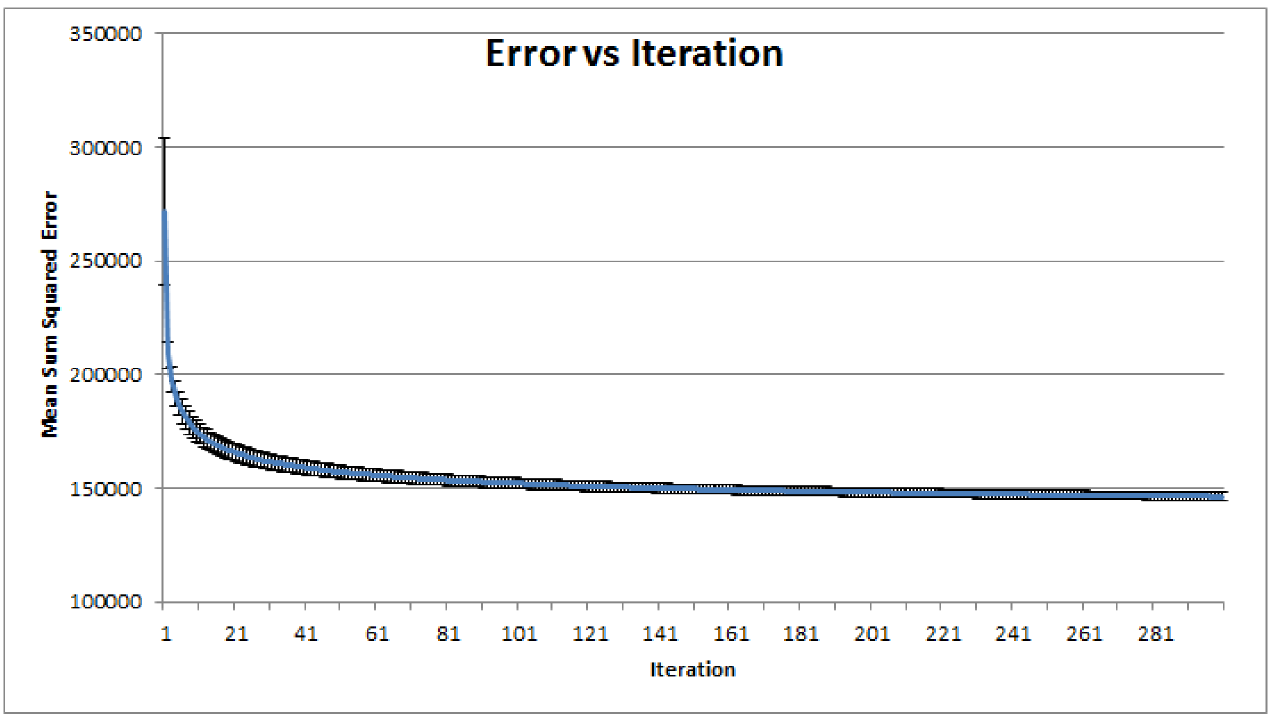
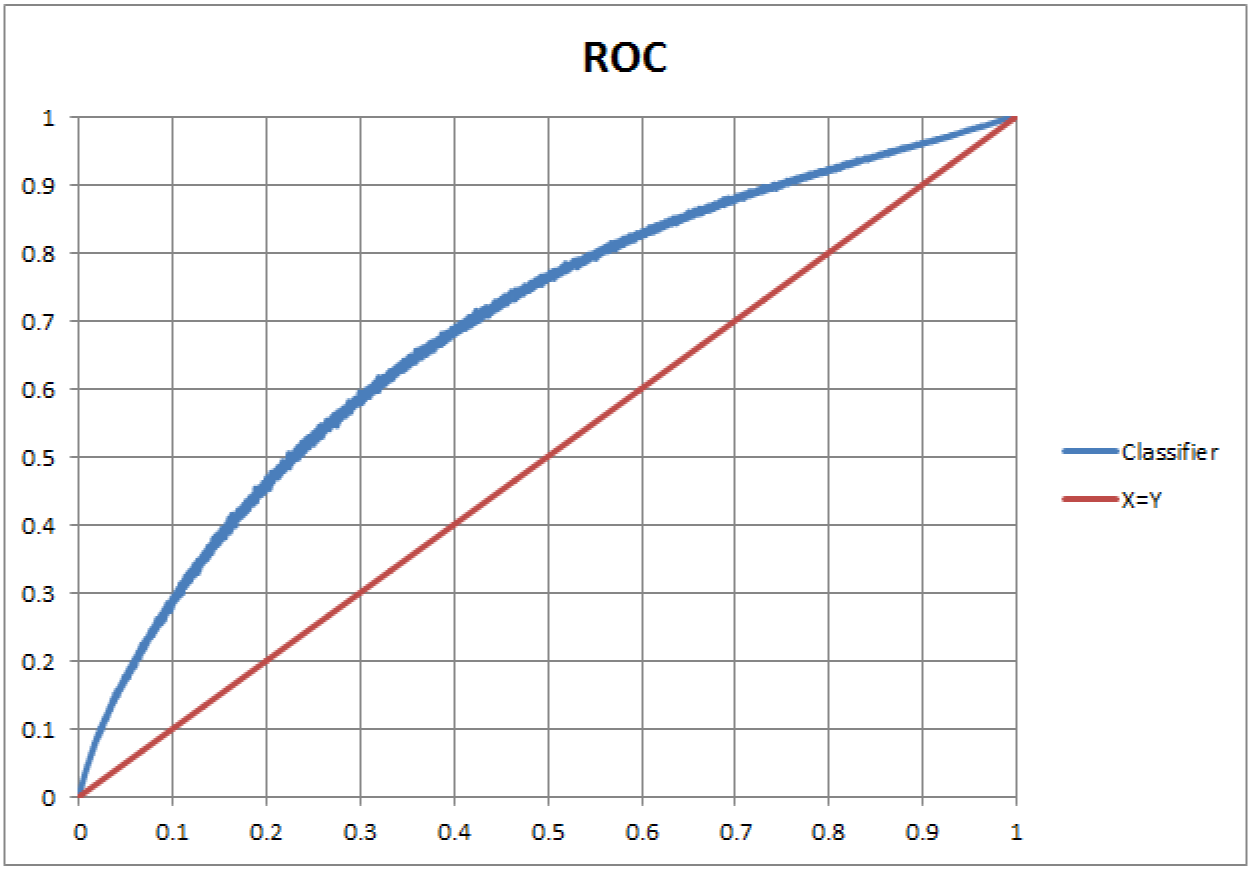
Project group: Eunkwang Joo and Shaddi Hasan

Our implementation proceeds in phases. In the first phase, we read in the tokenized data and construct our set of features and our set of observations. We chose to only use terms that appeared more than 1000 times in the dataset; this yielded around 10,000 features. After reading in the tokens, we produce two matrices: a matrix containing all the reviews and their associated terms (the document matrix) and a vector containing the ratings for every document. We also produce maps between feature ID's and terms.

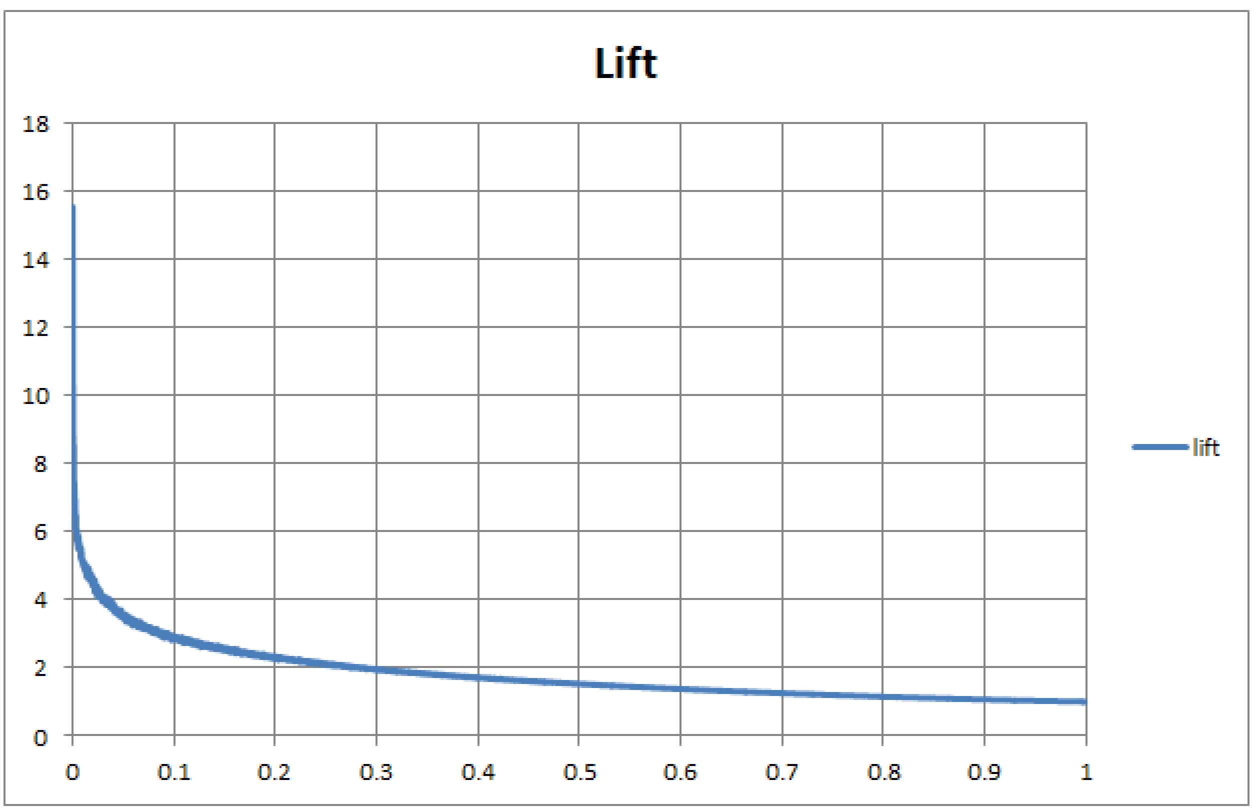
Our second phase is training. To perform 10-fold cross validation, we split the original document matrix and ratings vector into 10 blocks, each with approximately 10,000 reviews. Each fold proceeds identically. To train our model, we initially start with a β vector (our per-term weights) of all zeros. We train in blocks of reviews: for each block, we compute the L_2 error with ridge regularization, and then update β accordingly. Updating β for all blocks in our training set constitutes one iteration of our classifier (9 updates to β). We used a Γ value of 0.0000005, and a λ value of 0.1. We chose these values based on trial-and-error, as they seemed to provide the fastest decrease in error. We trained our model for 300 iterations. We computed the mean error-vs-iteration across our ten folds; we present it below. Note, the y-axis on the chart begins at 100000.



The final phase is validation. We performed validation on each of the left out blocks of data in our ten trials. We simply multiply each validation set by the corresponding β, and compare to the known values of these. We present the results of our validation below. Our mean AUC score across all validations was 0.69. Our 1% lift score was 5.0. We present the average ROC of all our validation experiments below; this is computed by averaging the TPR across all ten validation sets for each 0.01% increase in FPR. We also show the corresponding lift plot.



ROC
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Classifier
X=Y



Lift
18
16
14
12
10
8
6
4
2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
lift

30 most positive and negative features are as below. Since we did not remove stopwords before training the sets, we created another list without stop words. Interestingly, many stopwords are removed from top positive word list, yet top 30 negative words list was not changed at all. That means there was no stopwords in negative words list. Another interesting point is that there are a handful of Spanish words in positive words list. A couple of possible hypotheses can be drawn from it; if Spanish speakers are likely to leave positive reviews, or Spanish books are excellent.

<Top 30 positive and negative words excluding stop words>

Positive Words	Weight		Negative Words	Weight
condition	0.539220273		waste	-0.932127416
que	0.398371726		poorly	-0.745070159
excellent	0.383721173		disappointing	-0.666487336
awesome	0.366281718		disappointment	-0.6291821
arrived	0.315301955		worst	-0.613406837
book	0.31330052		boring	-0.545012176
en	0.310523629		disappointed	-0.458413661
libro	0.309729159		useless	-0.456440419
es	0.288584977		garbage	-0.421924919
y	0.287229359		trash	-0.40516749
great	0.269056618		awful	-0.392378926
pleased	0.261166781		fails	-0.372941524
outstanding	0.251193821		skip	-0.360766917
para	0.250852048		hoping	-0.360411286
el	0.250788033		ridiculous	-0.35848403
delivery	0.245177329		terrible	-0.3472296
loved	0.238798827		sorry	-0.338045597
thank	0.234022632		unfortunately	-0.336393267

thanks	0.232853383		lacks	-0.330875188
timely	0.229029328		tedious	-0.322208315
una	0.224982768		stupid	-0.318461746
hooked	0.219346598		misleading	-0.318363011
de	0.218777731		repetitive	-0.297471374
superb	0.217557445		drivel	-0.292259187
boo	0.215776235		errors	-0.289391369
amazing	0.212290674		unless	-0.288588792
informative	0.209878638		bother	-0.285527825
invaluable	0.208554819		pathetic	-0.285136402
turner	0.205427036		shallow	-0.283490002
fantastic	0.204718933		dull	-0.278490335

<Top 30 positive and negative words>

Positive Words	Weight		Negative Words	Weight
the	1.486795068		waste	-0.932127416
and	0.959186673		poorly	-0.745070159
a	0.66863215		disappointing	-0.666487336
condition	0.539220273		disappointment	-0.6291821
to	0.480947256		worst	-0.613406837
this	0.447643697		boring	-0.545012176
que	0.398371726		disappointed	-0.458413661
excellent	0.383721173		useless	-0.456440419
awesome	0.366281718		garbage	-0.421924919

arrived	0.315301955		trash	-0.40516749
book	0.31330052		awful	-0.392378926
en	0.310523629		fails	-0.372941524
libro	0.309729159		skip	-0.360766917
es	0.288584977		hoping	-0.360411286
y	0.287229359		ridiculous	-0.35848403
great	0.269056618		terrible	-0.3472296
of	0.262968689		sorry	-0.338045597
pleased	0.261166781		unfortunately	-0.336393267
outstanding	0.251193821		lacks	-0.330875188
para	0.250852048		tedious	-0.322208315
el	0.250788033		stupid	-0.318461746
delivery	0.245177329		misleading	-0.318363011
loved	0.238798827		repetitive	-0.297471374
thank	0.234022632		drivel	-0.292259187
thanks	0.232853383		errors	-0.289391369
timely	0.229029328		unless	-0.288588792
una	0.224982768		bother	-0.285527825
hooked	0.219346598		pathetic	-0.285136402
de	0.218777731		shallow	-0.283490002
superb	0.217557445		dull	-0.278490335