Name: Tanay Mukherjee

STA 9891

NOTE: Code for replication with the output has been shared in the appendix section.

(a) What are the top 10 words associated with positive reviews?

refreshing, wonderfully, 7, haunting, fi, noir, kung, captures, rare, apartment

(b) What are the top 10 words associated with negative reviews?

unfunny, forgettable, poorly, sci, waste, disappointment, worst, pointless, lousy, fu

(c) How can we identify the review in the training set that is hardest to classify? Find it and present the review.

We can identify the hardest review to classify by looking for probability which is close to 0.50.

The most difficult review to categorize has the rank 17508

? ? ? is in general a well ? ? we know well from books and movies this movie or this story don't work and i felt its not ? mistake the cast isn't good the actors are over ? and making ? ? the costumes are so clean and ? that everything even ? clothes look fake and for the serious ? who thinks twice this movie can be seen as a comedy instead of mystery drama the actor playing ? ? is doing a nice job but nothing fantastic the scenes are as said before perfect and looking fake the story is not very ? although a mystery of murder but who cares about the death of a ? and ? ? ? woman in the ? desert the ? is not likable

(d) Find and show the most positive review and the most negative review.

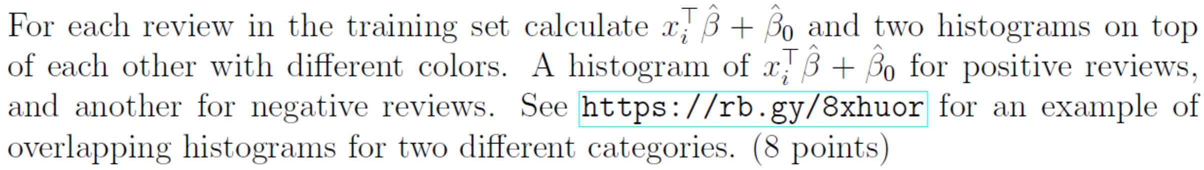
(i) Most Positive review has the rank 9158

? by now you've probably heard a bit about the new disney ? of ? classic film ? castle in the sky during late summer of ? disney released ? ? service on video which included a ? of the ? ? saying it was due out in ? it's obviously way past that year now but the ? has been finally ? and it's not ? castle in the sky just castle in the sky for the ? since ? is not such a nice word in spanish even though they use the word ? many times throughout the ? you've also probably heard that world ? ? joe ? who ? the movie originally went back to ? the excellent music with new ? ? came out before my ? ? and after ? of the ? of the wind which began studio ? and it's long ? of hits and in my opinion i think it's one of ? best films with a powerful lesson ? inside this two hour and four minute gem ? castle in the sky is a film for all ages and i ? everyone to see it br br for those ? with castle in the ? story it begins right at the start and doesn't stop for the next two hours the ? is so ? and ? ? you see ? true vision and believe me it's one fantastic one the film begins with ? a girl with one ? past as she is being held ? by the government on an ? ? holds the key to ? the castle in the sky and a long lost ? the key to ? is a ? ? she has which is ? by many ? the government the military and the air ? group the ? gang who ? and ? later ? soon the ? attack the ship and she ? during the ? she falls a few ? feet but the fall is soft and thanks to her ? as she ? down from the sky ? an ? boy who ? by working in the ? sees ? and ? her the two become fast friends but thanks to her ? the two get caught up in one huge ? ride as the ? gang and government try to capture ? one action sequence after another we learn all of the character's ? and ? as we build to the emotional and action ? climax which will surely please all with it's fantastic animation and wonderful dialogue plus somewhat ? surprise i think this film i s simply remarkable and does hold for the two hour and four minute run time the story is wonderful as we ? into ? ? animation which has no ? the setting of the film is a ? of many time ? it does seem to take place at the end of the ? but it is some ? ? which has ? technology and ? ? is also surprisingly a funny film the film has ? of hilarious moments almost ? to the drama and action the film holds i think the funniest part is a fight scene where ? boss faces off against a ? and soon after a ? breaks out it's funny as we see the men compare their strength and the music fits right in with it perfectly br br now let's talk about how the ? ? an excellent cast give some great performances to bring these characters to life teen ? james van ? ? plays the hero ? who has a much more ? voice then in the japanese version where in the original he ? more ? either way i think his voice is a nice fit with ? anna ? the young oscar winner from the ? plays ? this is also a nice performance but the voice is a bit ? she doesn't stay true to one accent at times she sounds as american as ? ? but at other times she sounds like someone from new ? the performance i most enjoyed however was of ? ? who played ? ? not only is this an excellent performance but the voice and emotion she gives the character really brings it to life if there was ever a live action ? movie g d ? she would be the one to play her you can just imagine her in the role well somewhat luke ? himself mark ? is ? and this is another top rate ? performance you may be familiar with ? from a long line of voice work after he did the original star wars movies but he ? ? to full evil his voice sounds like his regular voice and mix of the ? who he played for many episodes on the animated batman series ? out the cast is voice character actor jim ? who does a great ? job as the general and andy dick and ? ? as members of the ? gang br br now let me talk about what really makes this ? special joe ? ? ? music for those who have never heard of him mr ? does the music and like all of ? films the music is very memorable each of his ? has it's own ? which fits the particular film perfectly now these new ? he has done are more american like which i think was the ? of the new ? don't ? the classic ? of the japanese version are still here in great form the score to me sounds to be ? like this is a hollywood ? it has more power it has more ? it's ? and ? the film's ? the first seconds where we are introduced to the ? has some new music i am not sure but i believe when we first saw the ? there was no music at all but a majority of the music has new ? and more background music to enjoy things seem very ? in a powerful scene the music is more ? then in the original versions in a ? scene it's more ? overall i think many of you will be ? with the new ? an ? i highly did myself and personally think it helps ? the film i ? the new score over the old one and i hope disney will release or ? the music ? to a full ? soundtrack br br another plus side to the ? is that the story remains ? and much of the original japanese lines are ? in ? i'm sure a few lines where changed and this is the same way lines have been changed but a majority are close or exactly the original lines and dialogue ? has written i was afraid some excellent lines would be ? but they were there ? some new lines have been added as well which help out but i am not sure whether to consider this a good thing or a bad thing disney did not ? the ending song it was in japanese i was ? when they did completely new songs for the ? ? but with this version it's the original song in japanese so i guess it's good it's still the original but bad since a majority of people seeing this ? speak english br br there is a big down side to this ? and it deals with how the voices match the character's ? of course in any ? it won't be perfect but i think in ? and ? the ? of lines to match were much better executed and disney had a little bit more time with this one some of the time everything ? perfect some of the time it doesn't ? match and in a rare case someone says something and the ? don't move at all there's a scene where ? ? and her mouth doesn't move one bit br br as far as things about the film itself these are my thoughts i thought the most amazing part of ? was the animation from the opening sequence to the ending the animation is so ? and ? you just have to watch in ? you see the true nature of each character true detail to their face with extreme close ups and action you have to give a ? of credit for the effort that these ? put into this film everything is so well done and beautifully hand drawn it's like a moving piece of art and to think this was done in the mid ? the animation is quite different from disney ? has it's own ? ? which is very different but very good and after all these years the ? look as ? as ever ? also has ? of action sequences lots of plane ? plus a few on ground these sequences are so well done and so intriguing it's scary that they are ? to a big budget action film and the finale is just something you must see the sound effects are pure and classic and fit ? guns ? and everything else well and like all ? films each one ? on a different theme i g ? ? this one has a great a lesson on ? and power people don't realize how ? can take over you and how having too much power isn't good people are obsessed with power and are ? and the main ? ? ? shows this br br all in all ? castle in the sky was a great film to begin with and is now ? for the most part i am glad a more mainstream audience now have the chance to see this classic animated film in all it's ? with a great voice cast who put a lot into the film with the excellent ? musical score from joe ? disney has done a nice job on this ? and is quite worthy though i think the voices ? the ? better in the ? and ? ? disney ? castle in the sky is still a great ? and is worth the long ? because now more can ? a fantastic film

(ii) Most Negative review has the rank 19639

? ? 3 starts as a group of ? ? men steal a ? ? developed to ? the dead while trying to escape the man is shot at the ? ? holding the ? is ? the man gets some of the green ? on a ? on his hand which soon after turns him into a flesh eating ? zombie within hours the ? area is ? with the flesh ? ? on the look out for fresh victims ? ? ? his army ? find themselves in big trouble as they stop to help ? ? ring her friend ? ? ? who has been ? by zombie ? general ? is in ? of the situation has to stop the zombie ? from ? throughout the whole world but will he his men ? br br this italian produced film was to be directed by italian zombie gore film ? ? ? but the story goes he ? a ? therefore couldn't finish the film so producer ? ? asked second ? director ? ? writer ? ? to step in complete the film apparently ? ? did more than just finish it they actually ? a lot of the footage ? shot added a lot of their own ? 3 ended up as nearly a straight 50 50 ? the script by ? is an absolute mess none of it is well thought out is just as stupid as it gets the scenes of zombie ? ? people are not only ? ? but the whole idea is just absurd the zombies themselves have no ? whatsoever look at the scene where ? is on the ? the zombies are slow as they ? along but then look at the scene earlier on where she was ? by the zombie with the ? because that one runs around like it's on ? then for no ? explanation about 10 minutes before the film ? the zombies suddenly develop the ability to speak which also looks ? there are so many things wrong with ? 3 scene after scene of terribly thought out ? directed action awful character's really dull broken english dialogue which doesn't make sense half the time then there's the embarrassing scene where the zombie head inside the ? suddenly ? the ability to fly through the air ? ? ? the scene when the ? in white ? ? at the end are about to kill ? roger but instead of using their ? ? they decide to try kill them by hand even when ? ? up a gun himself they still ? to use their ? when ? starts to shoot them all they still ? to use their ? it's one of the most ? handled scenes ever put to film then there's the end where ? takes off in the ? but can't rest it down on the ground for literally a few seconds to pick his buddy up then a ? of zombies suddenly ? up from under some ? of ? what since when did zombies hide themselves yet alone under ? of ? this all may sound ? but believe me it's not it's a really bad film that is just boring ? simply doesn't work on any level as a piece of entertainment except for a few ? laughs br br it's hard to know who was responsible for what exactly but none of the footage is particularly well shot it has a bland ? feel about it for some reason the makers have tried to ? every scene in ? the problem is they clearly only had one ? machine you can see that at one ? of the screen the ? is ? ? as it is coming straight out of the machine ? out as it ? across the scene since a lot of it is set during the day it doesn't add any sort of atmosphere whatsoever when they do get it right the ? is ? ? across the screen it just looks like they shot the scene on a ? day the direction is poor with no ? it just looks feels bottom of the ? stuff even the blood gore isn't up to much there's a gory hand ? at the start a scene when something ? out of a ? ? ? a ? woman what actually took her ? off in the ? by the way why didn't it take the ? off the guy who ? in to save her a few ok looking zombies is as gory as it gets for anyone hoping to see a gore ? the likes of which ? ? ? up during the late 70's early 80's will be very disappointed there aren't any decent ? scenes no ? no stand out ? ? very little gore at all br br ? the film is poor the special effects are cheap looking the cinematography is dull the music is terrible the locations are bland it has rock bottom production values this was actually shot in the ? to keep the cost down to a ? the entire film is obviously dubbed the acting still looks awful though the english version seems to have been written by someone who doesn't understand the language that well br br ? 3 is not a sequel to ? classic zombie gore ? ? 2 ? it has nothing to do with it at all apart from the cash in title i'm sorry but ? 3 is an amateurish mess of a film it's boring it makes no sense it's not funny enough to be entertaining it lacks any decent gore one to avoid

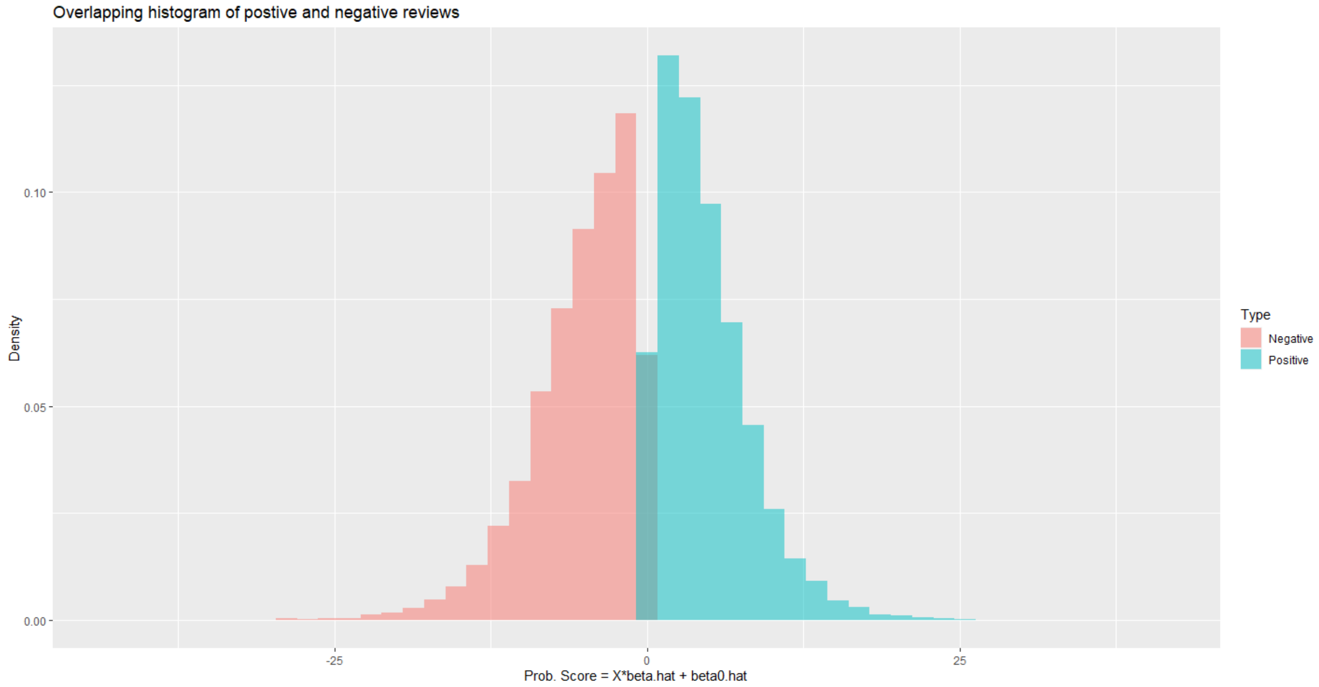
(e)



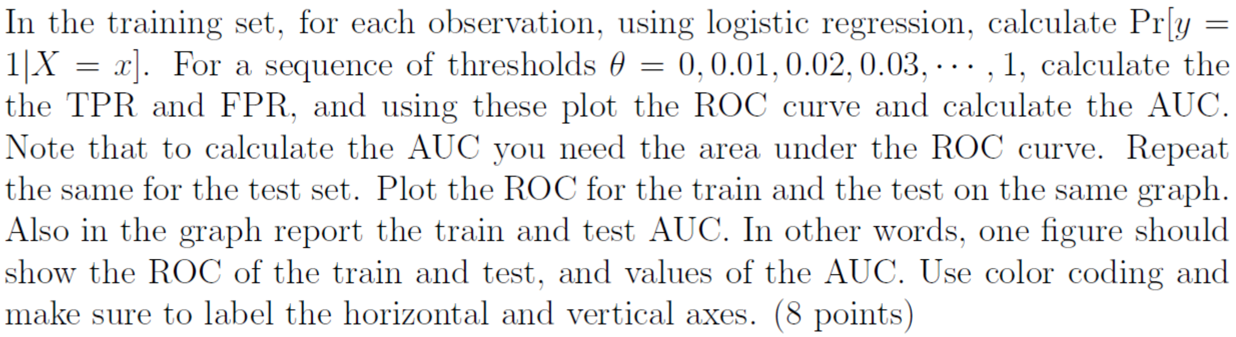
For each review in the training set calculate $x_i^\top \hat{\beta} + \hat{\beta}_0$ and two histograms on top of each other with different colors. A histogram of $x_i^\top \hat{\beta} + \hat{\beta}_0$ for positive reviews, and another for negative reviews. See `https://rb.gy/8xhuor` for an example of overlapping histograms for two different categories. (8 points)

We calculate the value of Xtrain*Beta_hat + Beta$_0$_hat and then categorize the values for positive and negative reviews, and then plot them together in a histogram as below:

I used the bin size of 50.



(f)



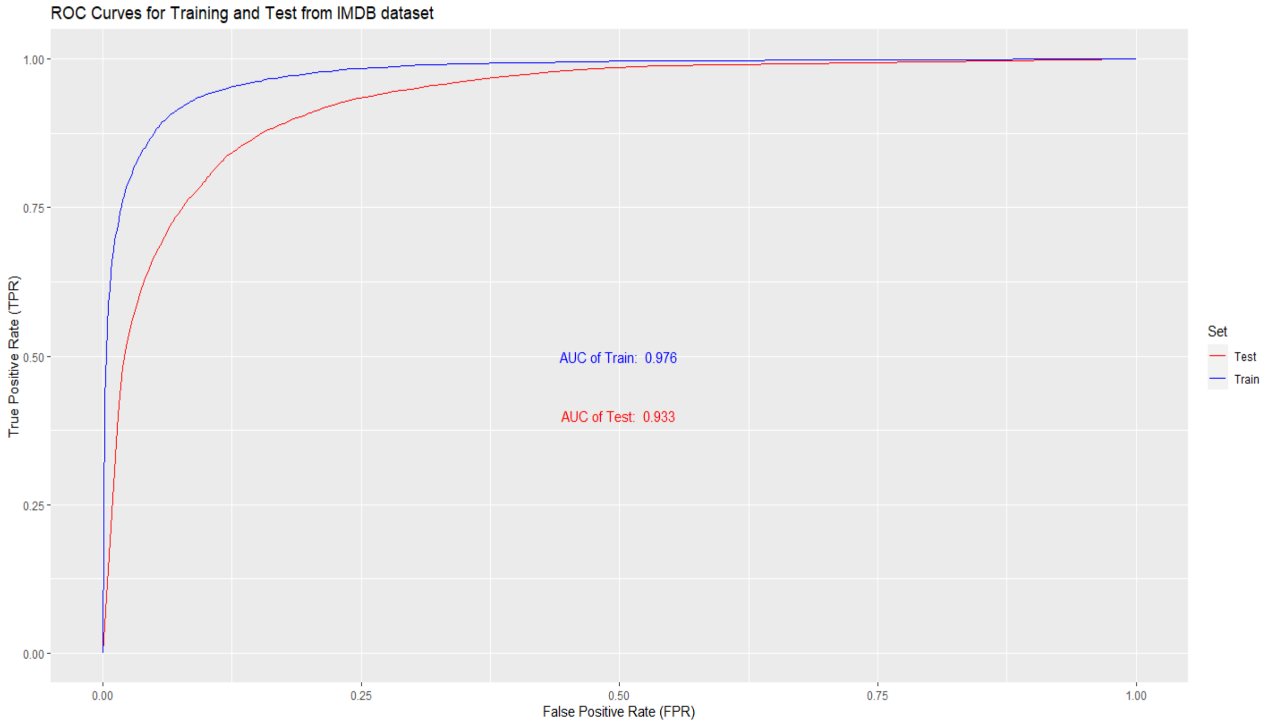
In the training set, for each observation, using logistic regression, calculate $\Pr[y = 1|X = x]$. For a sequence of thresholds $\theta = 0, 0.01, 0.02, 0.03, \cdots, 1$, calculate the the TPR and FPR, and using these plot the ROC curve and calculate the AUC. Note that to calculate the AUC you need the area under the ROC curve. Repeat the same for the test set. Plot the ROC for the train and the test on the same graph. Also in the graph report the train and test AUC. In other words, one figure should show the ROC of the train and test, and values of the AUC. Use color coding and make sure to label the horizontal and vertical axes. (8 points)

I calculated the FPR and TPR for each threshold value by first creating a dummy sequence for thresholds and then running a loop on that sequence for both training and test set.

Then I combined both the sets (training and test set) together to have all the FPRs and TPRs together for each threshold value. The values are printed in appendix section.

Lastly, I used the in-built method ‘colAUC’ from the library catools to calculate the AUC values and plotted the ROC curves for both training and test set for IMDB dataset as seen below:

AUC of Training set = 97.6% and **AUC of Test set = 93.3%**



# Appendix

(a), (b)

```
> # -- Part a and b
> cat(rev(tail(positive.Words,10)),sep = ", ")

> cat(head(negative.Words,10),sep = ", ")

# We can also get the result by changing the mw variable to 10 but
# since we have already done it in class simply calling the head and tail fun
ction will give us the result.
```

(c)

```
# -- Part c
> review_index  <- which.min(abs(prob.train-0.5))
> review_index

> decoded_review <- sapply(train_data[[review_index]], function(index) {
+   word <- if (index >= 3) reverse_word_index[[as.character(index - 3)]]
+   if (!is.null(word)) word else "?"
+ })
> cat(decoded_review)
```

(d)

```
> # -- Part d
> # Most Positive
> review_index  <- which.min(abs(prob.train-1))
> review_index

> decoded_review <- sapply(train_data[[review_index]], function(index) {
+   word <- if (index >= 3) reverse_word_index[[as.character(index - 3)]]
+   if (!is.null(word)) word else "?"
+ })
> cat(decoded_review)


> # Most Negative
> review_index  <- which.min(abs(prob.train-0))
> review_index

> decoded_review <- sapply(train_data[[review_index]], function(index) {
+   word <- if (index >= 3) reverse_word_index[[as.character(index - 3)]]
+   if (!is.null(word)) word else "?"
+ })
> cat(decoded_review)
```

(e)

```
> # -- Part e
> df <- as.data.frame(X.train %*% beta.hat + beta0.hat)
> names(df)[1] <- "Score"
> df <- df %>% mutate(Type = ifelse(Score > 0, "Positive", "Negative"))
> library(ggplot2)
> ggplot(df, aes (Score, fill = Type)) +
+   geom_histogram(alpha = 0.5, aes (y = ..density..), position = "identity",
bins = 100) +
+   labs(title  = "Overlapping histogram of postive and negative reviews", x
= "Prob. Score = X*beta.hat + beta0.hat", y = "Density")
```

(f)

```
# -- Part f
> # To a create a dummy sequence of thresholds (theta) = 0, 0.1, 0.2, 0.3, 0.
4,....,1
> thrs_seq   <- c(seq(0,1, by = 0.01))
> FPR_train <- TPR_train <- FPR_test <- TPR_test <- rep(0, length(thrs_seq))
> prob.train                =         exp(X.train %*% beta.hat +  beta0.hat  )
/(1 + exp(X.train %*% beta.hat +  beta0.hat))
> prob.test                 =         exp(X.test %*% beta.hat +  beta0.hat  )/
(1 + exp(X.test %*% beta.hat +  beta0.hat))
> for (i in 1:length(thrs_seq)){
+   thrs                    = thrs_seq[i]
+   print(paste('For the threshold sequence:',sprintf("%.2f" , thrs_seq[i])))
+ 
+   # for training set
+   y.hat.train             =         ifelse(prob.train > thrs, 1, 0) #table(y
.hat.train, y.train)
+   FP.train                =         sum(y.train[y.hat.train==1] == 0) # fals
e positives = negatives in the data that were predicted as positive
+   TP.train                =         sum(y.hat.train[y.train==1] == 1) # true
positives = positives in the data that were predicted as positive
+   P.train                 =         sum(y.train==1) # total positives in the
data
+   N.train                 =         sum(y.train==0) # total negatives in the
data
+   FPR.train               =         FP.train/N.train # false positive rate =
type 1 error = 1 - specificity
+   TPR.train               =         TP.train/P.train # true positive rate =
1 - type 2 error = sensitivity = power
+   typeI.err.train         =         FPR.train
+   typeII.err.train        =         1 - TPR.train
+   FPR_train[i]            =         typeI.err.train
+   TPR_train[i]            =         1 - typeII.err.train
+   print(paste('FPR for training set is',FPR_train[i]))
+   print(paste('TPR for training set is',TPR_train[i]))
+ 
+   # for test set
+   y.hat.test              =         ifelse(prob.test > thrs,1,0) #table(y.ha
t.test, y.test)
+   FP.test                 =         sum(y.test[y.hat.test==1] == 0) # false
positives = negatives in the data that were predicted as positive
+   TP.test                 =        sum(y.hat.test[y.test==1] == 1) # true p
ositives = positives in the data that were predicted as positive
+   P.test                  =        sum(y.test==1) # total positives in the
data
+   N.test                  =        sum(y.test==0) # total negatives in the
data
+   TN.test                 =        sum(y.hat.test[y.test==0] == 0)# negativ
es in the data that were predicted as negatives
+   FPR.test                =        FP.test/N.test # false positive rate = t
ype 1 error = 1 - specificity
+   TPR.test                =        TP.test/P.test # true positive rate = 1
- type 2 error = sensitivity = recall
+   typeI.err.test          =        FPR.test
+   typeII.err.test         =        1 - TPR.test
+   FPR_test[i]             =        typeI.err.test
+   TPR_test[i]             =        1 - typeII.err.test
+   print(paste('FPR for test set is',FPR_test[i]))
+   print(paste('TPR for test set is',TPR_test[i]))
+ }
[1] "For the threshold sequence: 0.00"
[1] "FPR for training set is 1"
[1] "TPR for training set is 1"
[1] "FPR for test set is 1"
[1] "TPR for test set is 1"
[1] "For the threshold sequence: 0.01"
[1] "FPR for training set is 0.5204"
[1] "TPR for training set is 0.99728"
[1] "FPR for test set is 0.53752"
[1] "TPR for test set is 0.9884"
[1] "For the threshold sequence: 0.02"
[1] "FPR for training set is 0.4588"
[1] "TPR for training set is 0.99584"
[1] "FPR for test set is 0.47536"
[1] "TPR for test set is 0.98328"
[1] "For the threshold sequence: 0.03"
[1] "FPR for training set is 0.41664"
[1] "TPR for training set is 0.99472"
[1] "FPR for test set is 0.43608"
[1] "TPR for test set is 0.97848"
[1] "For the threshold sequence: 0.04"
[1] "FPR for training set is 0.38616"
[1] "TPR for training set is 0.99424"
[1] "FPR for test set is 0.40872"
[1] "TPR for test set is 0.97432"
[1] "For the threshold sequence: 0.05"
[1] "FPR for training set is 0.3612"
[1] "TPR for training set is 0.99304"
[1] "FPR for test set is 0.38784"
[1] "TPR for test set is 0.9704"
[1] "For the threshold sequence: 0.06"
[1] "FPR for training set is 0.34216"
[1] "TPR for training set is 0.99176"
[1] "FPR for test set is 0.37032"
[1] "TPR for test set is 0.9672"
[1] "For the threshold sequence: 0.07"
[1] "FPR for training set is 0.32464"
[1] "TPR for training set is 0.9908"
```

```
[1] "FPR for test set is 0.356"
[1] "TPR for test set is 0.96344"
[1] "For the threshold sequence: 0.08"
[1] "FPR for training set is 0.30888"
[1] "TPR for training set is 0.98984"
[1] "FPR for test set is 0.3416"
[1] "TPR for test set is 0.9604"
[1] "For the threshold sequence: 0.09"
[1] "FPR for training set is 0.29432"
[1] "TPR for training set is 0.98856"
[1] "FPR for test set is 0.33008"
[1] "TPR for test set is 0.95736"
[1] "For the threshold sequence: 0.10"
[1] "FPR for training set is 0.28136"
[1] "TPR for training set is 0.98808"
[1] "FPR for test set is 0.31864"
[1] "TPR for test set is 0.95576"
[1] "For the threshold sequence: 0.11"
[1] "FPR for training set is 0.27064"
[1] "TPR for training set is 0.98648"
[1] "FPR for test set is 0.30824"
[1] "TPR for test set is 0.95272"
[1] "For the threshold sequence: 0.12"
[1] "FPR for training set is 0.2608"
[1] "TPR for training set is 0.98536"
[1] "FPR for test set is 0.2992"
[1] "TPR for test set is 0.94992"
[1] "For the threshold sequence: 0.13"
[1] "FPR for training set is 0.2516"
[1] "TPR for training set is 0.98472"
[1] "FPR for test set is 0.28944"
[1] "TPR for test set is 0.94808"
[1] "For the threshold sequence: 0.14"
[1] "FPR for training set is 0.24056"
[1] "TPR for training set is 0.98368"
[1] "FPR for test set is 0.28112"
[1] "TPR for test set is 0.94552"
[1] "For the threshold sequence: 0.15"
[1] "FPR for training set is 0.23256"
[1] "TPR for training set is 0.98216"
[1] "FPR for test set is 0.27464"
[1] "TPR for test set is 0.9432"
[1] "For the threshold sequence: 0.16"
[1] "FPR for training set is 0.22496"
[1] "TPR for training set is 0.98096"
[1] "FPR for test set is 0.26664"
[1] "TPR for test set is 0.94064"
[1] "For the threshold sequence: 0.17"
[1] "FPR for training set is 0.21888"
[1] "TPR for training set is 0.97936"
[1] "FPR for test set is 0.26088"
[1] "TPR for test set is 0.93824"
[1] "For the threshold sequence: 0.18"
[1] "FPR for training set is 0.21136"
[1] "TPR for training set is 0.97856"
[1] "FPR for test set is 0.254"
[1] "TPR for test set is 0.93656"
```

```
[1] "For the threshold sequence: 0.19"
[1] "FPR for training set is 0.20496"
[1] "TPR for training set is 0.9776"
[1] "FPR for test set is 0.24816"
[1] "TPR for test set is 0.93512"
[1] "For the threshold sequence: 0.20"
[1] "FPR for training set is 0.19936"
[1] "TPR for training set is 0.97608"
[1] "FPR for test set is 0.2424"
[1] "TPR for test set is 0.93296"
[1] "For the threshold sequence: 0.21"
[1] "FPR for training set is 0.19312"
[1] "TPR for training set is 0.9748"
[1] "FPR for test set is 0.2368"
[1] "TPR for test set is 0.93016"
[1] "For the threshold sequence: 0.22"
[1] "FPR for training set is 0.18736"
[1] "TPR for training set is 0.9732"
[1] "FPR for test set is 0.23248"
[1] "TPR for test set is 0.928"
[1] "For the threshold sequence: 0.23"
[1] "FPR for training set is 0.18216"
[1] "TPR for training set is 0.97216"
[1] "FPR for test set is 0.22864"
[1] "TPR for test set is 0.92608"
[1] "For the threshold sequence: 0.24"
[1] "FPR for training set is 0.17656"
[1] "TPR for training set is 0.97096"
[1] "FPR for test set is 0.22352"
[1] "TPR for test set is 0.92352"
[1] "For the threshold sequence: 0.25"
[1] "FPR for training set is 0.172"
[1] "TPR for training set is 0.96936"
[1] "FPR for test set is 0.21968"
[1] "TPR for test set is 0.92184"
[1] "For the threshold sequence: 0.26"
[1] "FPR for training set is 0.16728"
[1] "TPR for training set is 0.968"
[1] "FPR for test set is 0.2164"
[1] "TPR for test set is 0.91976"
[1] "For the threshold sequence: 0.27"
[1] "FPR for training set is 0.16208"
[1] "TPR for training set is 0.9668"
[1] "FPR for test set is 0.21272"
[1] "TPR for test set is 0.91744"
[1] "For the threshold sequence: 0.28"
[1] "FPR for training set is 0.15728"
[1] "TPR for training set is 0.96504"
[1] "FPR for test set is 0.20928"
[1] "TPR for test set is 0.91496"
[1] "For the threshold sequence: 0.29"
[1] "FPR for training set is 0.1528"
[1] "TPR for training set is 0.96304"
[1] "FPR for test set is 0.20528"
[1] "TPR for test set is 0.91224"
[1] "For the threshold sequence: 0.30"
[1] "FPR for training set is 0.14864"
```

```
[1] "TPR for training set is 0.96176"
[1] "FPR for test set is 0.20096"
[1] "TPR for test set is 0.90976"
[1] "For the threshold sequence: 0.31"
[1] "FPR for training set is 0.1452"
[1] "TPR for training set is 0.96072"
[1] "FPR for test set is 0.19792"
[1] "TPR for test set is 0.90744"
[1] "For the threshold sequence: 0.32"
[1] "FPR for training set is 0.1404"
[1] "TPR for training set is 0.95904"
[1] "FPR for test set is 0.19456"
[1] "TPR for test set is 0.90496"
[1] "For the threshold sequence: 0.33"
[1] "FPR for training set is 0.136"
[1] "TPR for training set is 0.9576"
[1] "FPR for test set is 0.19104"
[1] "TPR for test set is 0.9028"
[1] "For the threshold sequence: 0.34"
[1] "FPR for training set is 0.1316"
[1] "TPR for training set is 0.95568"
[1] "FPR for test set is 0.18784"
[1] "TPR for test set is 0.9012"
[1] "For the threshold sequence: 0.35"
[1] "FPR for training set is 0.12696"
[1] "TPR for training set is 0.95376"
[1] "FPR for test set is 0.1844"
[1] "TPR for test set is 0.8988"
[1] "For the threshold sequence: 0.36"
[1] "FPR for training set is 0.12352"
[1] "TPR for training set is 0.95256"
[1] "FPR for test set is 0.18072"
[1] "TPR for test set is 0.89632"
[1] "For the threshold sequence: 0.37"
[1] "FPR for training set is 0.12064"
[1] "TPR for training set is 0.95024"
[1] "FPR for test set is 0.17768"
[1] "TPR for test set is 0.89328"
[1] "For the threshold sequence: 0.38"
[1] "FPR for training set is 0.11776"
[1] "TPR for training set is 0.94856"
[1] "FPR for test set is 0.17504"
[1] "TPR for test set is 0.89152"
[1] "For the threshold sequence: 0.39"
[1] "FPR for training set is 0.11368"
[1] "TPR for training set is 0.94752"
[1] "FPR for test set is 0.17136"
[1] "TPR for test set is 0.88952"
[1] "For the threshold sequence: 0.40"
[1] "FPR for training set is 0.11088"
[1] "TPR for training set is 0.94664"
[1] "FPR for test set is 0.16864"
[1] "TPR for test set is 0.88768"
[1] "For the threshold sequence: 0.41"
[1] "FPR for training set is 0.10752"
[1] "TPR for training set is 0.94448"
[1] "FPR for test set is 0.16656"
```

```
[1] "TPR for test set is 0.88544"
[1] "For the threshold sequence: 0.42"
[1] "FPR for training set is 0.10384"
[1] "TPR for training set is 0.94288"
[1] "FPR for test set is 0.16376"
[1] "TPR for test set is 0.88256"
[1] "For the threshold sequence: 0.43"
[1] "FPR for training set is 0.10032"
[1] "TPR for training set is 0.94064"
[1] "FPR for test set is 0.1604"
[1] "TPR for test set is 0.88088"
[1] "For the threshold sequence: 0.44"
[1] "FPR for training set is 0.09736"
[1] "TPR for training set is 0.93832"
[1] "FPR for test set is 0.15752"
[1] "TPR for test set is 0.87936"
[1] "For the threshold sequence: 0.45"
[1] "FPR for training set is 0.09384"
[1] "TPR for training set is 0.93584"
[1] "FPR for test set is 0.15488"
[1] "TPR for test set is 0.87648"
[1] "For the threshold sequence: 0.46"
[1] "FPR for training set is 0.09064"
[1] "TPR for training set is 0.93368"
[1] "FPR for test set is 0.1528"
[1] "TPR for test set is 0.87416"
[1] "For the threshold sequence: 0.47"
[1] "FPR for training set is 0.08848"
[1] "TPR for training set is 0.93104"
[1] "FPR for test set is 0.15048"
[1] "TPR for test set is 0.87168"
[1] "For the threshold sequence: 0.48"
[1] "FPR for training set is 0.08568"
[1] "TPR for training set is 0.92936"
[1] "FPR for test set is 0.14832"
[1] "TPR for test set is 0.86904"
[1] "For the threshold sequence: 0.49"
[1] "FPR for training set is 0.08384"
[1] "TPR for training set is 0.92696"
[1] "FPR for test set is 0.14664"
[1] "TPR for test set is 0.8668"
[1] "For the threshold sequence: 0.50"
[1] "FPR for training set is 0.08128"
[1] "TPR for training set is 0.92496"
[1] "FPR for test set is 0.1448"
[1] "TPR for test set is 0.86408"
[1] "For the threshold sequence: 0.51"
[1] "FPR for training set is 0.0792"
[1] "TPR for training set is 0.92224"
[1] "FPR for test set is 0.14208"
[1] "TPR for test set is 0.86112"
[1] "For the threshold sequence: 0.52"
[1] "FPR for training set is 0.07728"
[1] "TPR for training set is 0.91952"
[1] "FPR for test set is 0.13888"
[1] "TPR for test set is 0.85848"
[1] "For the threshold sequence: 0.53"
```

```
[1] "FPR for training set is 0.07496"
[1] "TPR for training set is 0.9172"
[1] "FPR for test set is 0.13624"
[1] "TPR for test set is 0.856"
[1] "For the threshold sequence: 0.54"
[1] "FPR for training set is 0.07176"
[1] "TPR for training set is 0.914"
[1] "FPR for test set is 0.13392"
[1] "TPR for test set is 0.85328"
[1] "For the threshold sequence: 0.55"
[1] "FPR for training set is 0.06968"
[1] "TPR for training set is 0.9112"
[1] "FPR for test set is 0.13096"
[1] "TPR for test set is 0.85016"
[1] "For the threshold sequence: 0.56"
[1] "FPR for training set is 0.0672"
[1] "TPR for training set is 0.90824"
[1] "FPR for test set is 0.12904"
[1] "TPR for test set is 0.84712"
[1] "For the threshold sequence: 0.57"
[1] "FPR for training set is 0.06488"
[1] "TPR for training set is 0.90592"
[1] "FPR for test set is 0.12664"
[1] "TPR for test set is 0.84448"
[1] "For the threshold sequence: 0.58"
[1] "FPR for training set is 0.06336"
[1] "TPR for training set is 0.90368"
[1] "FPR for test set is 0.12456"
[1] "TPR for test set is 0.84208"
[1] "For the threshold sequence: 0.59"
[1] "FPR for training set is 0.06168"
[1] "TPR for training set is 0.90008"
[1] "FPR for test set is 0.12208"
[1] "TPR for test set is 0.83936"
[1] "For the threshold sequence: 0.60"
[1] "FPR for training set is 0.05952"
[1] "TPR for training set is 0.89688"
[1] "FPR for test set is 0.11984"
[1] "TPR for test set is 0.83648"
[1] "For the threshold sequence: 0.61"
[1] "FPR for training set is 0.05736"
[1] "TPR for training set is 0.89384"
[1] "FPR for test set is 0.11816"
[1] "TPR for test set is 0.8332"
[1] "For the threshold sequence: 0.62"
[1] "FPR for training set is 0.05576"
[1] "TPR for training set is 0.88952"
[1] "FPR for test set is 0.11664"
[1] "TPR for test set is 0.82936"
[1] "For the threshold sequence: 0.63"
[1] "FPR for training set is 0.05432"
[1] "TPR for training set is 0.886"
[1] "FPR for test set is 0.1144"
[1] "TPR for test set is 0.82632"
[1] "For the threshold sequence: 0.64"
[1] "FPR for training set is 0.05256"
[1] "TPR for training set is 0.8828"
```

```
[1] "FPR for test set is 0.11264"
[1] "TPR for test set is 0.82368"
[1] "For the threshold sequence: 0.65"
[1] "FPR for training set is 0.05128"
[1] "TPR for training set is 0.87952"
[1] "FPR for test set is 0.11104"
[1] "TPR for test set is 0.81944"
[1] "For the threshold sequence: 0.66"
[1] "FPR for training set is 0.04912"
[1] "TPR for training set is 0.87512"
[1] "FPR for test set is 0.10848"
[1] "TPR for test set is 0.8152"
[1] "For the threshold sequence: 0.67"
[1] "FPR for training set is 0.04808"
[1] "TPR for training set is 0.8704"
[1] "FPR for test set is 0.10656"
[1] "TPR for test set is 0.81224"
[1] "For the threshold sequence: 0.68"
[1] "FPR for training set is 0.04608"
[1] "TPR for training set is 0.86616"
[1] "FPR for test set is 0.1048"
[1] "TPR for test set is 0.80792"
[1] "For the threshold sequence: 0.69"
[1] "FPR for training set is 0.0444"
[1] "TPR for training set is 0.86216"
[1] "FPR for test set is 0.1032"
[1] "TPR for test set is 0.804"
[1] "For the threshold sequence: 0.70"
[1] "FPR for training set is 0.04264"
[1] "TPR for training set is 0.85672"
[1] "FPR for test set is 0.10088"
[1] "TPR for test set is 0.79992"
[1] "For the threshold sequence: 0.71"
[1] "FPR for training set is 0.04152"
[1] "TPR for training set is 0.8524"
[1] "FPR for test set is 0.09952"
[1] "TPR for test set is 0.796"
[1] "For the threshold sequence: 0.72"
[1] "FPR for training set is 0.03936"
[1] "TPR for training set is 0.84856"
[1] "FPR for test set is 0.09704"
[1] "TPR for test set is 0.79152"
[1] "For the threshold sequence: 0.73"
[1] "FPR for training set is 0.03768"
[1] "TPR for training set is 0.84232"
[1] "FPR for test set is 0.09544"
[1] "TPR for test set is 0.7872"
[1] "For the threshold sequence: 0.74"
[1] "FPR for training set is 0.03592"
[1] "TPR for training set is 0.83664"
[1] "FPR for test set is 0.09312"
[1] "TPR for test set is 0.78296"
[1] "For the threshold sequence: 0.75"
[1] "FPR for training set is 0.03432"
[1] "TPR for training set is 0.8316"
[1] "FPR for test set is 0.09056"
[1] "TPR for test set is 0.77792"
```

```
[1] "For the threshold sequence: 0.76"
[1] "FPR for training set is 0.03304"
[1] "TPR for training set is 0.82704"
[1] "FPR for test set is 0.08808"
[1] "TPR for test set is 0.77352"
[1] "For the threshold sequence: 0.77"
[1] "FPR for training set is 0.03128"
[1] "TPR for training set is 0.82176"
[1] "FPR for test set is 0.0856"
[1] "TPR for test set is 0.7696"
[1] "For the threshold sequence: 0.78"
[1] "FPR for training set is 0.02952"
[1] "TPR for training set is 0.81472"
[1] "FPR for test set is 0.08312"
[1] "TPR for test set is 0.76512"
[1] "For the threshold sequence: 0.79"
[1] "FPR for training set is 0.02832"
[1] "TPR for training set is 0.80824"
[1] "FPR for test set is 0.08072"
[1] "TPR for test set is 0.75952"
[1] "For the threshold sequence: 0.80"
[1] "FPR for training set is 0.02664"
[1] "TPR for training set is 0.8004"
[1] "FPR for test set is 0.0784"
[1] "TPR for test set is 0.75304"
[1] "For the threshold sequence: 0.81"
[1] "FPR for training set is 0.02496"
[1] "TPR for training set is 0.79344"
[1] "FPR for test set is 0.07616"
[1] "TPR for test set is 0.7472"
[1] "For the threshold sequence: 0.82"
[1] "FPR for training set is 0.02304"
[1] "TPR for training set is 0.78576"
[1] "FPR for test set is 0.07408"
[1] "TPR for test set is 0.74136"
[1] "For the threshold sequence: 0.83"
[1] "FPR for training set is 0.02216"
[1] "TPR for training set is 0.77744"
[1] "FPR for test set is 0.07072"
[1] "TPR for test set is 0.734"
[1] "For the threshold sequence: 0.84"
[1] "FPR for training set is 0.0208"
[1] "TPR for training set is 0.76976"
[1] "FPR for test set is 0.06776"
[1] "TPR for test set is 0.72608"
[1] "For the threshold sequence: 0.85"
[1] "FPR for training set is 0.01944"
[1] "TPR for training set is 0.762"
[1] "FPR for test set is 0.06512"
[1] "TPR for test set is 0.7192"
[1] "For the threshold sequence: 0.86"
[1] "FPR for training set is 0.018"
[1] "TPR for training set is 0.75168"
[1] "FPR for test set is 0.06248"
[1] "TPR for test set is 0.70944"
[1] "For the threshold sequence: 0.87"
[1] "FPR for training set is 0.01672"
```

```
[1] "TPR for training set is 0.74208"
[1] "FPR for test set is 0.06032"
[1] "TPR for test set is 0.70128"
[1] "For the threshold sequence: 0.88"
[1] "FPR for training set is 0.01576"
[1] "TPR for training set is 0.73088"
[1] "FPR for test set is 0.05712"
[1] "TPR for test set is 0.6912"
[1] "For the threshold sequence: 0.89"
[1] "FPR for training set is 0.01456"
[1] "TPR for training set is 0.71864"
[1] "FPR for test set is 0.054"
[1] "TPR for test set is 0.68136"
[1] "For the threshold sequence: 0.90"
[1] "FPR for training set is 0.01336"
[1] "TPR for training set is 0.70712"
[1] "FPR for test set is 0.05048"
[1] "TPR for test set is 0.67016"
[1] "For the threshold sequence: 0.91"
[1] "FPR for training set is 0.01184"
[1] "TPR for training set is 0.69384"
[1] "FPR for test set is 0.0472"
[1] "TPR for test set is 0.65704"
[1] "For the threshold sequence: 0.92"
[1] "FPR for training set is 0.01016"
[1] "TPR for training set is 0.67768"
[1] "FPR for test set is 0.04416"
[1] "TPR for test set is 0.64352"
[1] "For the threshold sequence: 0.93"
[1] "FPR for training set is 0.0092"
[1] "TPR for training set is 0.66024"
[1] "FPR for test set is 0.04016"
[1] "TPR for test set is 0.6272"
[1] "For the threshold sequence: 0.94"
[1] "FPR for training set is 0.008"
[1] "TPR for training set is 0.64"
[1] "FPR for test set is 0.0364"
[1] "TPR for test set is 0.60856"
[1] "For the threshold sequence: 0.95"
[1] "FPR for training set is 0.00696"
[1] "TPR for training set is 0.61552"
[1] "FPR for test set is 0.03328"
[1] "TPR for test set is 0.5868"
[1] "For the threshold sequence: 0.96"
[1] "FPR for training set is 0.0056"
[1] "TPR for training set is 0.58656"
[1] "FPR for test set is 0.02904"
[1] "TPR for test set is 0.562"
[1] "For the threshold sequence: 0.97"
[1] "FPR for training set is 0.00464"
[1] "TPR for training set is 0.54896"
[1] "FPR for test set is 0.0236"
[1] "TPR for test set is 0.5248"
[1] "For the threshold sequence: 0.98"
[1] "FPR for training set is 0.00376"
[1] "TPR for training set is 0.5036"
[1] "FPR for test set is 0.01896"
```

```
[1] "TPR for test set is 0.4804"
[1] "For the threshold sequence: 0.99"
[1] "FPR for training set is 0.00216"
[1] "TPR for training set is 0.42776"
[1] "FPR for test set is 0.01456"
[1] "TPR for test set is 0.40168"
[1] "For the threshold sequence: 1.00"
[1] "FPR for training set is 0"
[1] "TPR for training set is 0"
[1] "FPR for test set is 0"
[1] "TPR for test set is 0"

> train = data.frame(FPR = FPR_train, TPR = TPR_train, Set = 'Train', Thresho
ld = thrs_seq)
> test  = data.frame(FPR = FPR_test,  TPR = TPR_test,  Set = 'Test',  Thresho
ld = thrs_seq)
> df <- rbind(train, test)

> # Using colAUC method in catools library to get the AUC value
> library(caTools)

> # ROC for training set
> AUC_train = colAUC(prob.train, y.train, plotROC = F)[1]
> # ROC for test set
> AUC_test = colAUC(prob.test, y.test, plotROC = F)[1]

> # Plot the ROC curves
> ggplot(data = df, aes(x=FPR, y = TPR, col = Set))+
+   geom_line(show.legend = T)+
+   labs(title = 'ROC Curves for Training and Test from IMDB dataset', x = 'F
alse Positive Rate (FPR)', y = 'True Positive Rate (TPR)' ) +
+   annotate(geom="text",
+            x=c(0.5,0.5),
+            y=c(0.4,0.5),
+            label=c(paste('AUC of Test: ',round(AUC_test, 3)), paste('AUC of
Train: ',round(AUC_train, 3))),
+            color=c('red', 'blue'))+
+   scale_color_manual(values=c('red', 'blue'))
>
```