Automating Reviews in Medicine

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.0     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Warning: package 'forcats' was built under R version 3.4.3
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caTools)
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.4.3
```

```
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.4
```

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.4.3
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.3
```

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

The medical literature is enormous. Pubmed, a database of medical publications maintained by the U.S. National Library of Medicine, has indexed over 23 million medical publications. Further, the rate of medical publication has increased over time, and now there are nearly 1 million new publications in the field each year, or more than one per minute.

The large size and fast-changing nature of the medical literature has increased the need for reviews, which search databases like Pubmed for papers on a particular topic and then report results from the papers found. While such reviews are often performed manually, with multiple people reviewing each search result, this is tedious and time consuming. In this problem, we will see how text analytics can be used to automate the process of information retrieval.

The dataset consists of the titles (variable title) and abstracts (variable abstract) of papers retrieved in a Pubmed search. Each search result is labeled with whether the paper is a clinical trial testing a drug therapy for cancer (variable trial). These labels were obtained by two people reviewing each search result and accessing the actual paper if necessary, as part of a literature review of clinical trials testing drug therapies for advanced and metastatic breast cancer.

1.1) Loading the Data

```
trials = read.csv('clinical_trial.csv', stringsAsFactors = FALSE)
```

Load clinical_trial.csv into a data frame called trials (remembering to add the argument stringsAsFactors=FALSE), and investigate the data frame with summary() and str().

IMPORTANT NOTE: Some students have been getting errors like "invalid multibyte string" when performing certain parts of this homework question. If this is happening to you, use the argument fileEncoding="latin1" when reading in the file with read.csv. This should cause those errors to go away.

We can use R's string functions to learn more about the titles and abstracts of the located papers. The nchar() function counts the number of characters in a piece of text. Using the nchar() function on the variables in the data frame, answer the following questions:

How many characters are there in the longest abstract? (Longest here is defined as the abstract with the largest number of characters.)

```
summary(trials)
```

```
##     title              abstract            trial
##  Length:1860        Length:1860        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000
##                                        Mean   :0.4392
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
```

```
str(trials)
```

```
## 'data.frame':    1860 obs. of  3 variables:
##  $ title   : chr  "Treatment of Hodgkin's disease and other cancers with 1,3-bis(2-chloroethyl)-1-ni
##  $ abstract: chr  "" "Twenty-eight cases of malignancies of different kinds were studied to assess T
##  $ trial   : int  1 0 1 1 1 0 1 0 0 0 ...
```

```
max(nchar(trials$abstract))
```

```
## [1] 3708
```

1.2) Loading the Data

How many search results provided no abstract? (HINT: A search result provided no abstract if the number of characters in the abstract field is zero.)

```
summary(nchar(trials$abstract) == 0)
```

```
##    Mode   FALSE    TRUE
## logical    1748     112
```

```
sum(nchar(trials$abstract) == 0)
```

```
## [1] 112
```

1.3) Loading the Data

Find the observation with the minimum number of characters in the title (the variable "title") out of all of the observations in this dataset. What is the text of the title of this article? Include capitalization and punctuation in your response, but don't include the quotes.

```
which.min(nchar(trials$title))
```

```
## [1] 1258
```

```
trials[1258, 'title']
```

```
## [1] "A decade of letrozole: FACE."
```

Explanation

To identify which title is the shortest, we can use

which.min(nchar(trials$title))

From this, we know the 1258th title is the shortest. We can access this title with trials$title[1258].

2.1) Preparing the Corpus

Because we have both title and abstract information for trials, we need to build two corpora instead of one. Name them **corpusTitle** and **corpusAbstract**.

Following the commands from lecture, perform the following tasks (you might need to load the "tm" package first if it isn't already loaded). Make sure to perform them in this order.

1) Convert the title variable to corpusTitle and the abstract variable to corpusAbstract.

```
corpus_title = VCorpus(VectorSource(trials$title))
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time(), tz = "GMT"): unknown timezone
## 'zone/tz/2018i.1.0/zoneinfo/America/Chicago'
```

```
corpus_abstract = VCorpus(VectorSource(trials$abstract))
```

2) Convert corpusTitle and corpusAbstract to lowercase.

```
corpus_title = tm_map(corpus_title, content_transformer(tolower))
corpus_abstract = tm_map(corpus_abstract, content_transformer(tolower))
```

3) Remove the punctuation in corpusTitle and corpusAbstract.

```
corpus_title = tm_map(corpus_title, removePunctuation)
corpus_abstract = tm_map(corpus_abstract, removePunctuation)
```

4) Remove the English language stop words from corpusTitle and corpusAbstract.

```
corpus_title = tm_map(corpus_title, removeWords, stopwords('english'))
corpus_abstract = tm_map(corpus_abstract, removeWords, stopwords('english'))
```

5) Stem the words in corpusTitle and corpusAbstract (each stemming might take a few minutes).

```
corpus_title = tm_map(corpus_title, stemDocument)
corpus_abstract = tm_map(corpus_abstract, stemDocument)
```

6) Build a document term matrix called dtmTitle from corpusTitle and dtmAbstract from corpusAbstract.

```
dtm_title = DocumentTermMatrix(corpus_title)
dtm_abstract = DocumentTermMatrix(corpus_abstract)
```

7) Limit dtmTitle and dtmAbstract to terms with sparseness of at most 95% (aka terms that appear in at least 5% of documents).

```
dtm_title = removeSparseTerms(dtm_title, 0.95)
dtm_abstract = removeSparseTerms(dtm_abstract, 0.95)
```

8) Convert dtmTitle and dtmAbstract to data frames (keep the names dtmTitle and dtmAbstract).

```
dtm_title = as.data.frame(as.matrix(dtm_title))
dtm_abstract = as.data.frame(as.matrix(dtm_abstract))
```

If the code length(stopwords("english")) does not return 174 for you, then please run the line of code in this file, which will store the standard stop words in a variable called sw. When removing stop words, use tm_map(corpusTitle, removeWords, sw) and tm_map(corpusAbstract, removeWords, sw) instead of tm_map(corpusTitle, removeWords, stopwords("english")) and tm_map(corpusAbstract, removeWords, stopwords("english")).

How many terms remain in dtmTitle after removing sparse terms (aka how many columns does it have)?

```
dim(dtm_title)
```

```
## [1] 1860   31
```

```
dim(dtm_abstract)
```

```
## [1] 1860  335
```

Explanation

Below we provide the code for corpusTitle; only minor modifications are needed to build corpusAbstract.

corpusTitle = VCorpus(VectorSource(trials$title))

corpusTitle = tm_map(corpusTitle, content_transformer(tolower))

corpusTitle = tm_map(corpusTitle, removePunctuation)

corpusTitle = tm_map(corpusTitle, removeWords, stopwords("english"))

corpusTitle = tm_map(corpusTitle, stemDocument)

dtmTitle = DocumentTermMatrix(corpusTitle)

dtmTitle = removeSparseTerms(dtmTitle, 0.95)

dtmTitle = as.data.frame(as.matrix(dtmTitle))

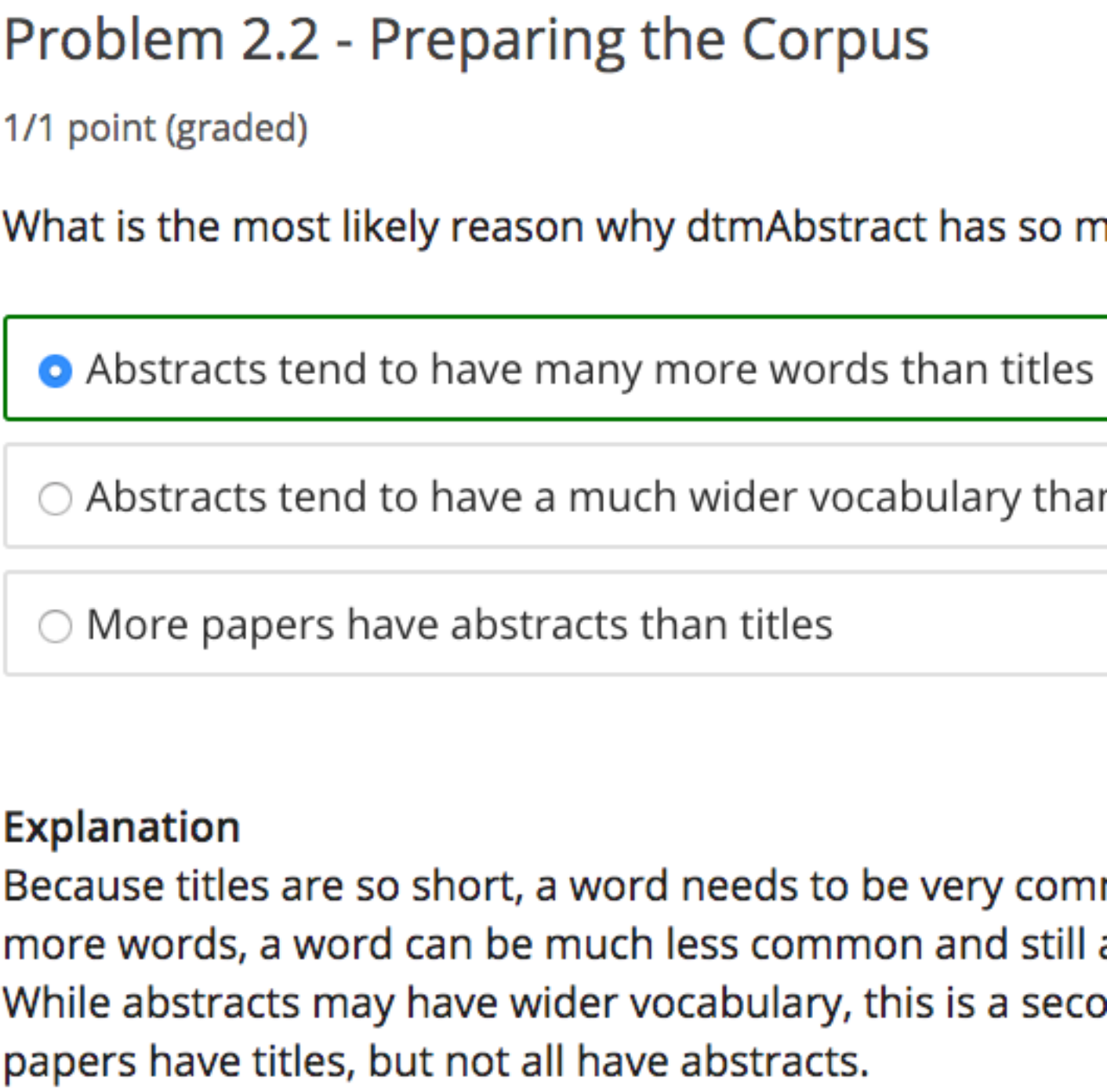
Explanation

These can be read from str(dtmTitle) and str(dtmAbstract). Other than str(), the dim() or ncol() functions could have been used. If you used fileEncoding="latin1" when reading in the datafile, you'll have a few extra terms in dtmAbstract, but you should get the answer correct.

2.2) Preparing the Corpus

What is the most likely reason why dtmAbstract has so many more terms than dtmTitle?

```
knitr::include_graphics('2.2.png')
```



2.3) Preparing the Corpus

What is the most frequent word stem across all the abstracts? Hint: you can use colSums() to compute the frequency of a word across all the abstracts.

```
sort(colSums(dtm_abstract), decreasing = TRUE)
```

```
##         patient          breast          cancer       treatment
```

```
##            8381            3859            3726            2894
##           group    chemotherapi          respons         signific
##            2668            2344            2051            2043
##           studi          surviv          receiv       tamoxifen
##            1965            1927            1908            1632
##           month         therapi          random          result
##            1575            1564            1520            1485
##           women           trial          compar          effect
##            1484            1417            1359            1340
##            year            rate             day          median
##            1335            1253            1245            1180
##          differ           adjuv            dose           tumor
##            1176            1162            1123            1122
##            mgm2            week           toxic             use
##            1093            1074            1065            1053
##             arm            cycl          overal          diseas
##            1038             962             962             950
##          clinic          combin           evalu           treat
##             944             926             926             893
##          method             two            time         conclus
##             892             889             881             842
##         regimen         respect        metastat           level
##             807             758             755             743
##         increas         primari          observ          follow
##             729             718             700             675
##          assess            risk cyclophosphamid         complet
##             668             635             632             628
##            plus        progress         control          associ
##             622             622             621             604
##        efficaci    postmenopaus         analysi             cmf
##             591             590             587             586
##           grade        receptor             one          versus
##             580             573             570             570
##           three          improv          advanc          factor
##             564             562             556             552
##         benefit          status          either          includ
##             551             538             532             529
##            show            bone       docetaxel           posit
##             516             514             514             511
##           activ        followup           everi     doxorubicin
##             509             494             487             486
##           phase            node         placebo            alon
##             481             477             475             472
##          recurr         similar          assign          purpos
##             465             438             435             434
##           chang             age          hormon            oral
##             431             429             428             422
##         estrogen           first           addit          higher
##             421             421             420             415
##             may           daili          measur           event
##             413             412             411             409
##         surgeri            data          object          reduc
##             407             405             400             400
##      background      paclitaxel           total         statist
```

```
##             397             397             397             384
##            high           given           toler            four
##             378             374             373             369
##         predict           among            also      diseasefre
##             369             365             364             364
##            cell          report         express        previous
##             359             357             356             355
##         determin        metastas            less           relat
##             352             352             351             351
##         decreas          interv           durat           ratio
##             350             349             344             344
##         perform         baselin      epirubicin           howev
##             342             340             339             339
##          outcom            drug            well           earli
##             335             332             328             325
##         administ          tumour           serum            her2
##             322             320             315             314
##           occur             dfs            mean         without
##             312             310             310             306
##           prior        standard     premenopaus          hazard
##             305             305             303             301
##          reduct           incid           local          number
##             301             300             300             296
##        investig         partial        neoadjuv        axillari
##             295             295             293             292
##           sever           stage           cours            test
##             288             286             283             282
##         continu         greater             mbc           initi
##             281             279             276             275
##         suggest           indic        endocrin             iii
##             274             269             266             266
##       methotrex          safeti        randomis          failur
##             265             265             264             262
##             six         develop           negat          advers
##             261             259             258             256
##            valu        patholog          relaps           score
##             256             254             254             254
##       carcinoma        demonstr            0001           lymph
##             251             251             249             249
##            rang          achiev    radiotherapi        prognost
##             248             245             244             242
##          confid           agent          nausea        prospect
##             241             240             239             239
##           found           infus           lower         whether
##             238             237             236             235
##     neutropenia            case             100             end
##             234             233             225             221
##           enrol          design       administr             per
##             221             219             218             218
##         present           death    fluorouracil         schedul
##             218             215             215             215
##        endpoint   5fluorouracil          growth    anthracyclin
##             213             208             208             207
##          correl           point         consist         respond
```

```
##             203             202             200             200
##      nodeposit            seen           tissu            elig
##             199             199             197             196
##             low           hundr          longer            oper
##             196             195             193             193
##          system        intraven        subgroup        although
##             193             192             192             191
##             can          common             set          examin
##             191             191             191             190
##          marker         qualiti        function          better
##             189             189             188             186
##             aim            site         support          accord
##             185             183             183             182
##        firstlin       inhibitor          involv           model
##             182             182             180             180
##         confirm            life          analys         conduct
##             178             178             177             177
##            find           least          postop            size
##             177             177             177             177
##        pretreat           vomit            five          wherea
##             175             174             173             173
##           sampl          within        aromatas             new
##             172             172             171             171
##          period             500          requir        sequenti
##             170             169             168             168
##            side       experienc      mastectomi          appear
##             168             167             165             164
##             001            popul        superior          profil
##             162             162             161             158
##  progressionfre          remain       secondari         potenti
##             158             158             158             156
##          provid             due       multivari           stabl
##             155             154             154             154
##        frequent            term   receptorposit            evid
##             153             153             152             150
##         distant     doubleblind        independ           singl
##             149             149             149             149
##          detect        identifi          obtain           human
##             148             148             147             144
##        particip         prevent           estim          import
##             144             143             139             138
##          second          consid            main           start
##             138             131             131             131
##         possibl           andor            need        histolog
##             130             128             128             127
##           limit         general       multicent            type
##             127             126             126             126
##         prolong         proport             005          analyz
##             125             125             124             124
##            base           eight          inform          select
##             124             124             124             124
##           defin           major         regress    characterist
##             123             122             120             119
##           enter       hematolog           shown      comparison
```

```
##             117             117             117             116
##           caus           trend     progesteron           avail
##             115             115             114             108
##           larg           seven         regard
##             108             108             105
```

Explanation

We can compute the column sums and then identify the most common one with:

csAbstract = colSums(dtmAbstract)

which.max(csAbstract)

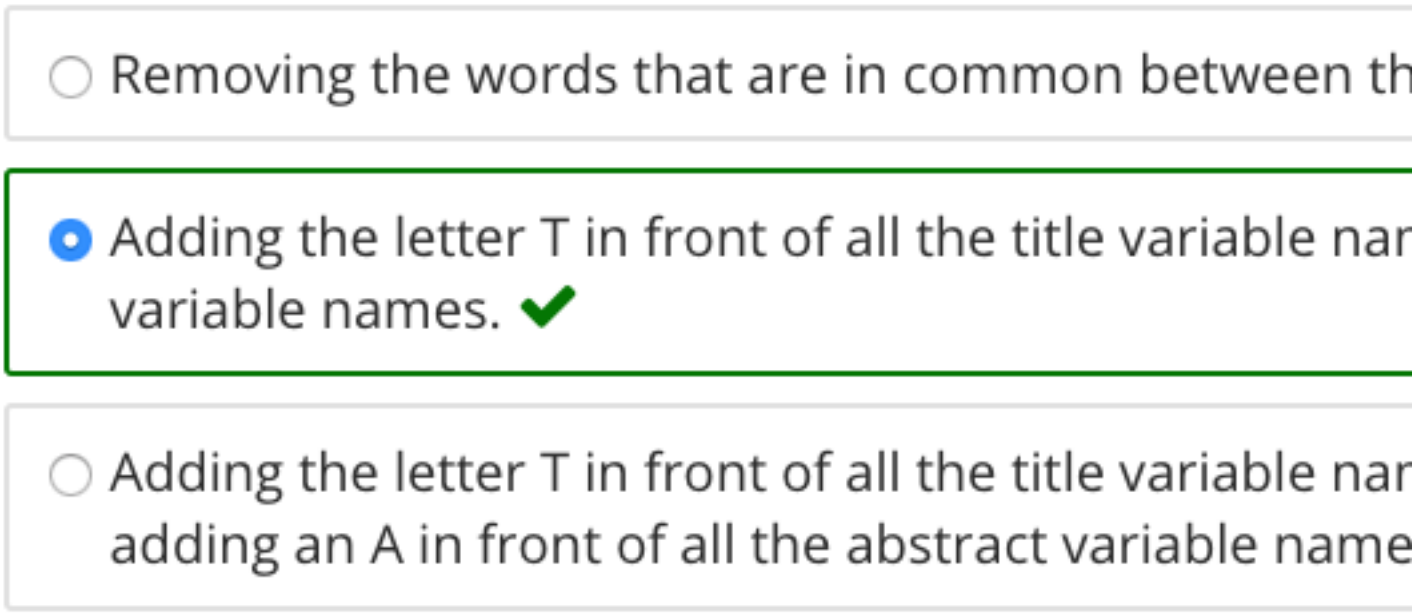
3.1) Building a model

We want to combine dtmTitle and dtmAbstract into a single data frame to make predictions. However, some of the variables in these data frames have the same names. To fix this issue, run the following commands:

```
colnames(dtm_title) = paste0("T", colnames(dtm_title))
colnames(dtm_abstract) = paste0("A", colnames(dtm_abstract))
```

What was the effect of these functions?

```
include_graphics('3.1.png')
```



3.2) Building a Model

Using cbind(), combine dtmTitle and dtmAbstract into a single data frame called dtm:

```
dtm = cbind(dtm_title, dtm_abstract)
```

As we did in class, add the dependent variable "trial" to dtm, copying it from the original data frame called trials. How many columns are in this combined data frame?

```
dtm$trial = trials$trial

dim(dtm)
```

```
## [1] 1860  367
```

Explanation

The combination can be accomplished with:

dtm = cbind(dtmTitle, dtmAbstract)

dtm$trial = trials$trial

The number of variables in the combined data frame can be read from str(dtm) or ncol(dtm). If you used fileEncoding="latin1" when reading in the file, you should have 5 extra variables (but the answer should be graded as correct).

3.3) Building a Model

Now that we have prepared our data frame, it's time to split it into a training and testing set and to build regression models. Set the random seed to 144 and use the sample.split function from the caTools package to split dtm into data frames named "train" and "test", putting 70% of the data in the training set.

```
set.seed(144)

spl = sample.split(dtm$trial, SplitRatio = 0.7)

train = subset(dtm, spl == TRUE)

test = subset(dtm, spl == FALSE)
```

What is the accuracy of the baseline model on the training set? (Remember that the baseline model predicts the most frequent outcome in the training set for all observations.)

```
table(train$trial)
```

```
##
##   0   1
## 730 572
```

```
730 / (730 + 572)
```

```
## [1] 0.5606759
```

Explanation

Just as in any binary classification problem, the naive baseline always predicts the most common class. From table(train$trial), we see 730 training set results were not trials, and 572 were trials. Therefore, the naive baseline always predicts a result is not a trial, yielding accuracy of 730/(730+572).

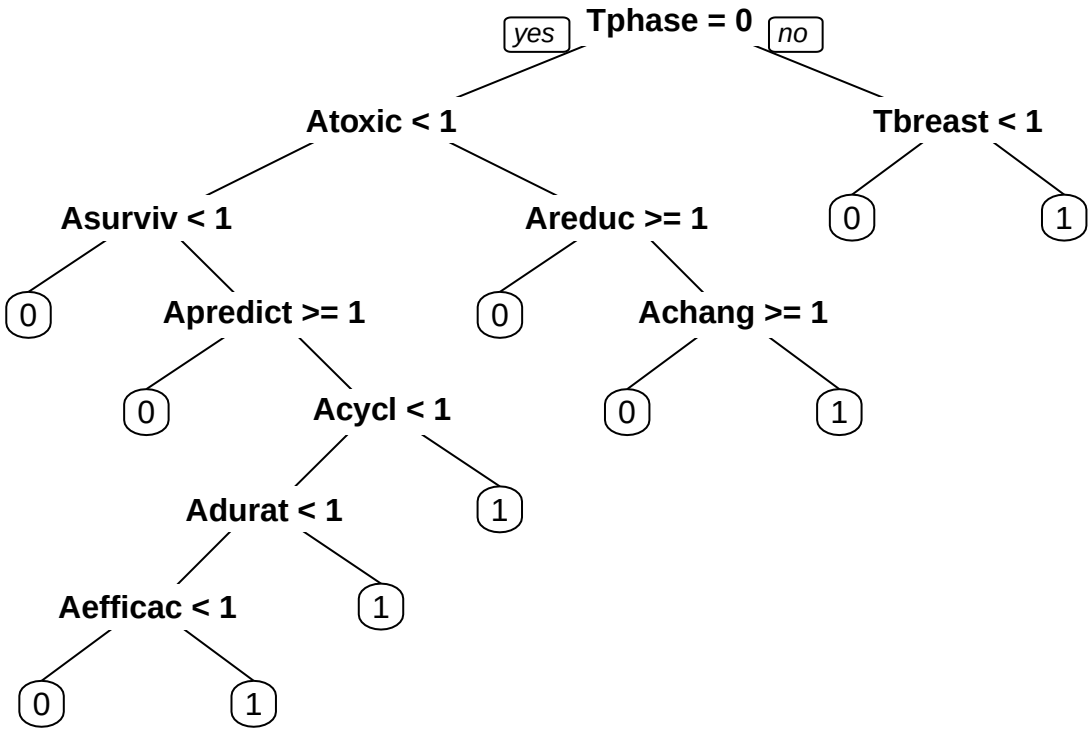
3.4) Building a Model

Build a CART model called trialCART, using all the independent variables in the training set to train the model, and then plot the CART model. Just use the default parameters to build the model (don't add

a minbucket or cp value). Remember to add the method="class" argument, since this is a classification problem.

```
trial_cart = rpart(trial ~ ., data = train, method = 'class')

prp(trial_cart)
```



What is the name of the first variable the model split on?

Explanation

This can be accomplished with:

trialCART = rpart(trial~., data=train, method="class")

prp(trialCART)

The first split checks whether or not Tphase is less than 0.5

3.5) Building a Model

Obtain the training set predictions for the model (do not yet predict on the test set). Extract the predicted probability of a result being a trial (recall that this involves not setting a type argument, and keeping only the second column of the predict output). What is the maximum predicted probability for any result?

```
pred = predict(trial_cart, train)

pred[1:10,]
```

```
##           0          1
## 1  0.8636364 0.13636364
## 2  0.8636364 0.13636364
## 3  0.1281139 0.87188612
## 5  0.2176871 0.78231293
```

```
## 6  0.9454545 0.05454545
## 7  0.2176871 0.78231293
## 10 0.1281139 0.87188612
## 12 0.7125000 0.28750000
## 13 0.1281139 0.87188612
## 14 0.7125000 0.28750000
```

```
pred_prob = pred[,2]

max(pred_prob)
```

```
## [1] 0.8718861
```

Explanation

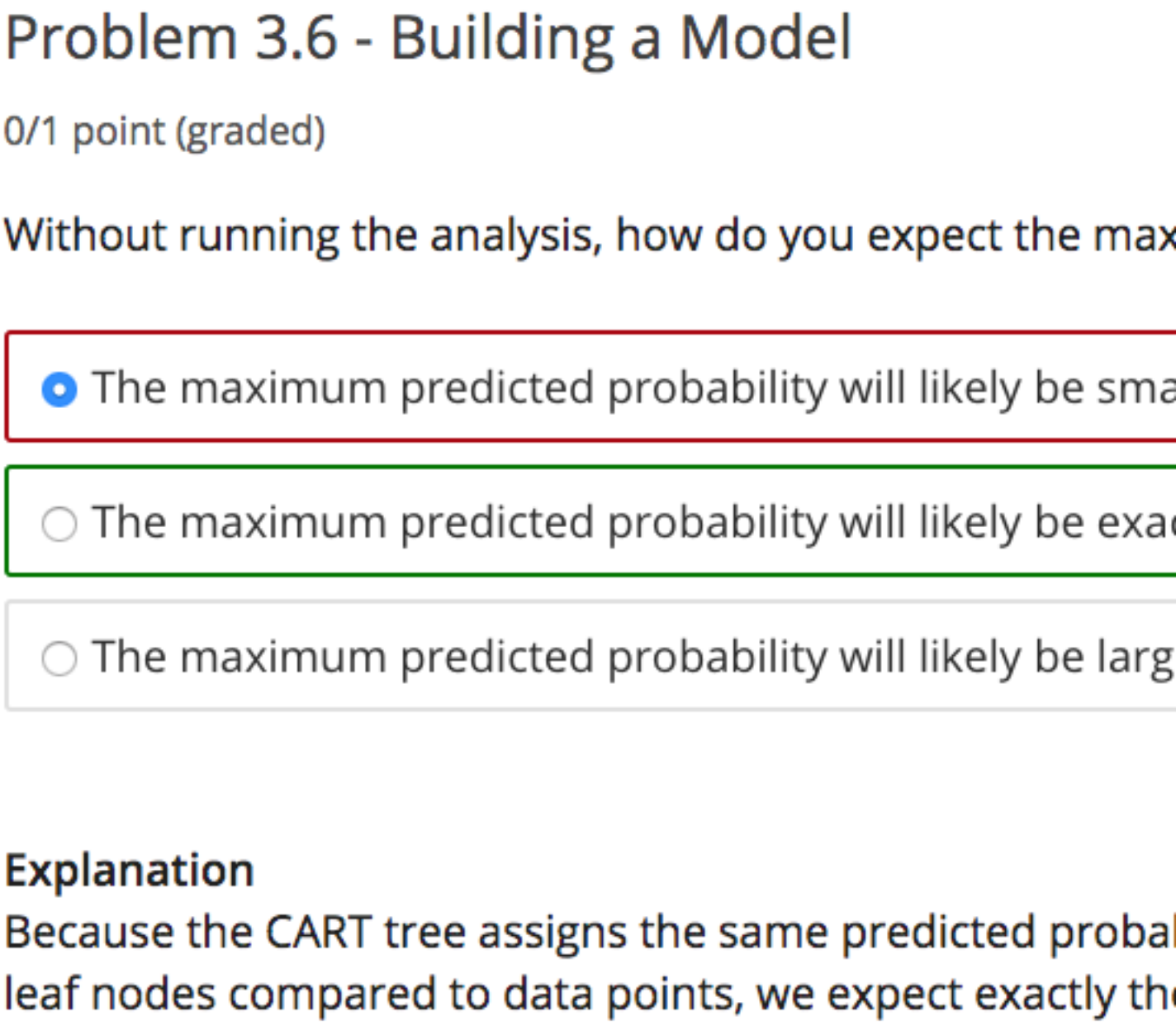
The training set predictions can be obtained and summarized with the following commands:

predTrain = predict(trialCart)[,2]

summary(predTrain)

3.6) Building a Model

```
include_graphics('3.6.png')
```



3.7) Building a Model

For these questions, use a threshold probability of 0.5 to predict that an observation is a clinical trial.

What is the training set accuracy of the CART model?

```
pred_train = predict(trial_cart, train, type = 'class')

table(train$trial, pred_train)
```

```
##    pred_train
##       0   1
##   0 631  99
##   1 131 441
```

```
(631 + 441) / (631 + 441 + 99 + 131)
```

```
## [1] 0.8233487
```

What is the training set sensitivity of the CART model?

```
table(train$trial, pred_train)
```

```
##    pred_train
##       0   1
##   0 631  99
##   1 131 441
```

```
# TP / (TP + FN)

441/ (441 + 131)
```

```
## [1] 0.770979
```

What is the training set specificity of the CART model?

```
table(train$trial, pred_train)
```

```
##    pred_train
##       0   1
##   0 631  99
##   1 131 441
```

```
# TN / (TN + FP)

631 / (631 + 99)
```

```
## [1] 0.8643836
```

Explanation

We can compare the predictions with threshold 0.5 to the true results in the training set with:

table(train$trial, predTrain >= 0.5)

From this, we read the following confusion matrix (rows are true outcome, columns are predicted outcomes):

FALSE TRUE

0 631 99

1 131 441

We conclude that the model has training set accuracy (631+441)/(631+441+99+131), sensitivity 441/(441+131) and specificity 631/(631+99).

4.1) Evaluating the model on the testing set

Evaluate the CART model on the testing set using the predict function and creating a vector of predicted probabilities predTest.

```
pred_test = predict(trial_cart, newdata = test, type = 'class')
```

What is the testing set accuracy, assuming a probability threshold of 0.5 for predicting that a result is a clinical trial?

```
table(test$trial, pred_test)
```

```
##    pred_test
##       0   1
##   0 261  52
##   1  83 162
```

```
(261 + 162) / (261 + 162 + 52 + 83)
```

```
## [1] 0.7580645
```

Explanation

The testing set predictions can be obtained and compared to the true outcomes with:

predTest = predict(trialCART, newdata=test)[,2]

table(test$trial, predTest >= 0.5)

This yields the following confusion matrix:

FALSE TRUE

0 261 52

1 83 162

From this, we read that the testing set accuracy is (261+162)/(261+162+83+52).

4.2) Evaluating the Model on the Testing Set

Using the ROCR package, what is the testing set AUC of the prediction model?

pred = predict(trial_cart, train)

pred[1:10,]

pred_prob = pred[,2]

max(pred_prob)

```
pred_test_2 = predict(trial_cart, newdata = test)

pred_test_2[1:5,]
```

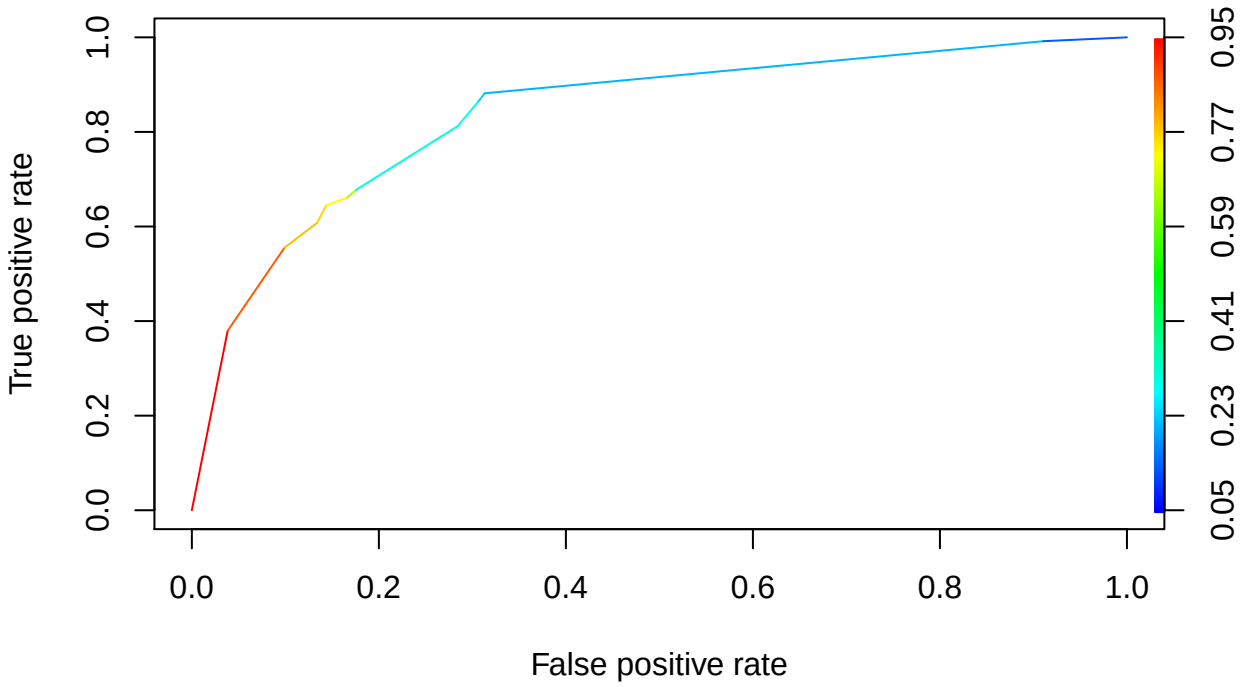
```
##            0         1
## 4  0.1281139 0.8718861
## 8  0.8636364 0.1363636
## 9  0.7125000 0.2875000
## 11 0.7125000 0.2875000
## 19 0.7125000 0.2875000
```

```
pred_prob_test = pred_test_2[,2]
```

```
pred_ROCR = prediction(pred_prob_test, test$trial)

perf_ROCR = performance(pred_ROCR, "tpr", "fpr")

plot(perf_ROCR, colorize=TRUE)
```




```
performance(pred_ROCR, "auc")@y.values
```

```
## [[1]]
## [1] 0.8371063
```

Explanation

The AUC can be determined using the following code:

library(ROCR)

pred = prediction(predTest, test$trial)

as.numeric(performance(pred, "auc")@y.values)

Part 5: decision-maker tradeoffs

The decision maker for this problem, a researcher performing a review of the medical literature, would use a model (like the CART one we built here) in the following workflow:

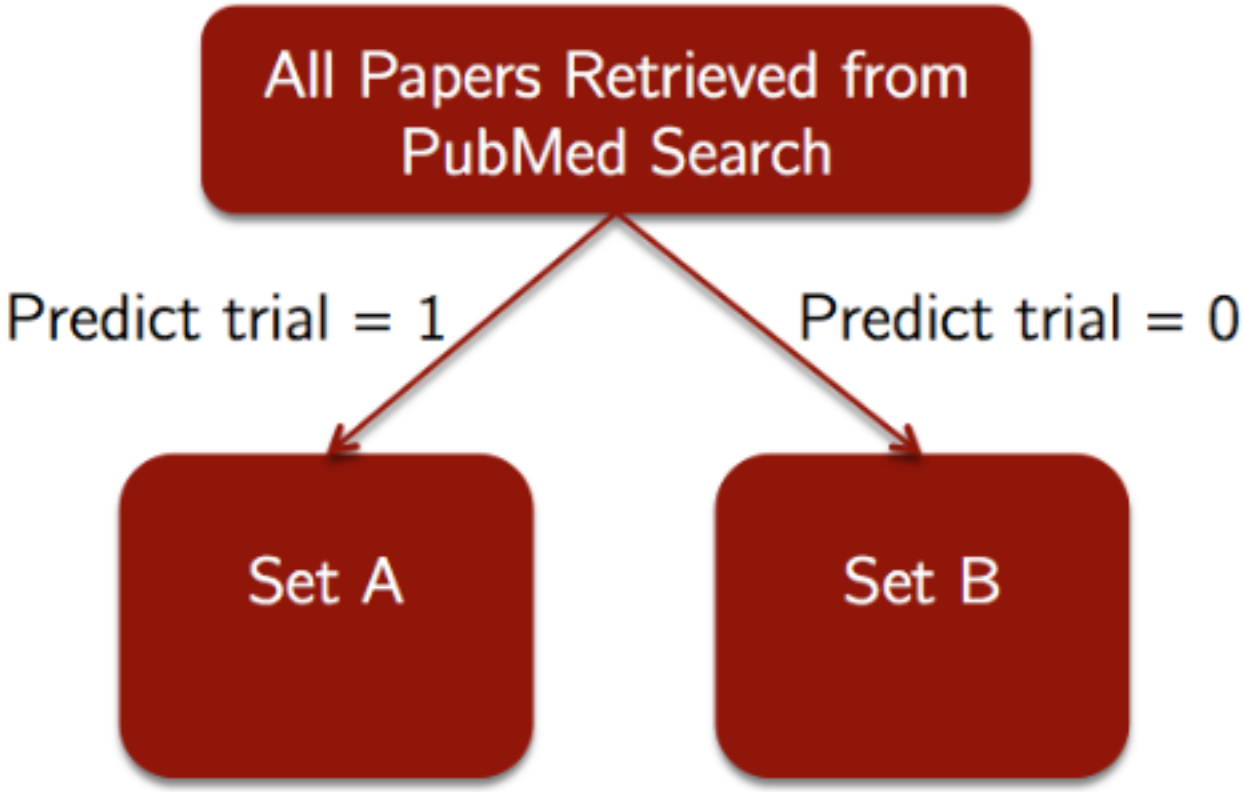
1) For all of the papers retreived in the PubMed Search, predict which papers are clinical trials using the model. This yields some initial Set A of papers predicted to be trials, and some Set B of papers predicted not to be trials. (See the figure below.)

2) Then, the decision maker manually reviews all papers in Set A, verifying that each paper meets the study's detailed inclusion criteria (for the purposes of this analysis, we assume this manual review is 100% accurate at identifying whether a paper in Set A is relevant to the study). This yields a more limited set of papers to be included in the study, which would ideally be all papers in the medical literature meeting the detailed inclusion criteria for the study.

3) Perform the study-specific analysis, using data extracted from the limited set of papers identified in step 2.

This process is shown in the figure below.

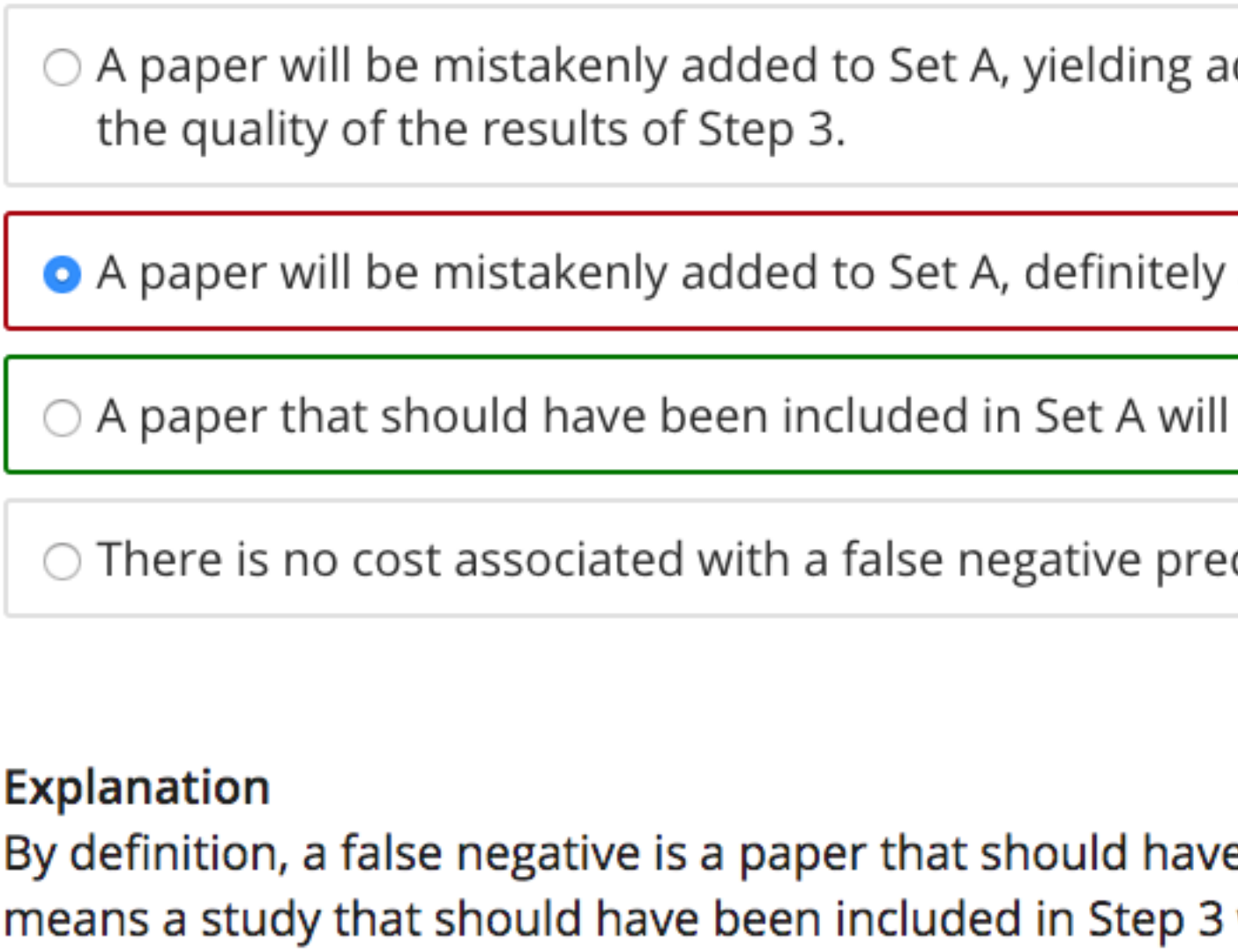
```
include_graphics('part5.png')
```



5.1) Decision-Maker Tradeoffs

What is the cost associated with the model in Step 1 making a false negative prediction?

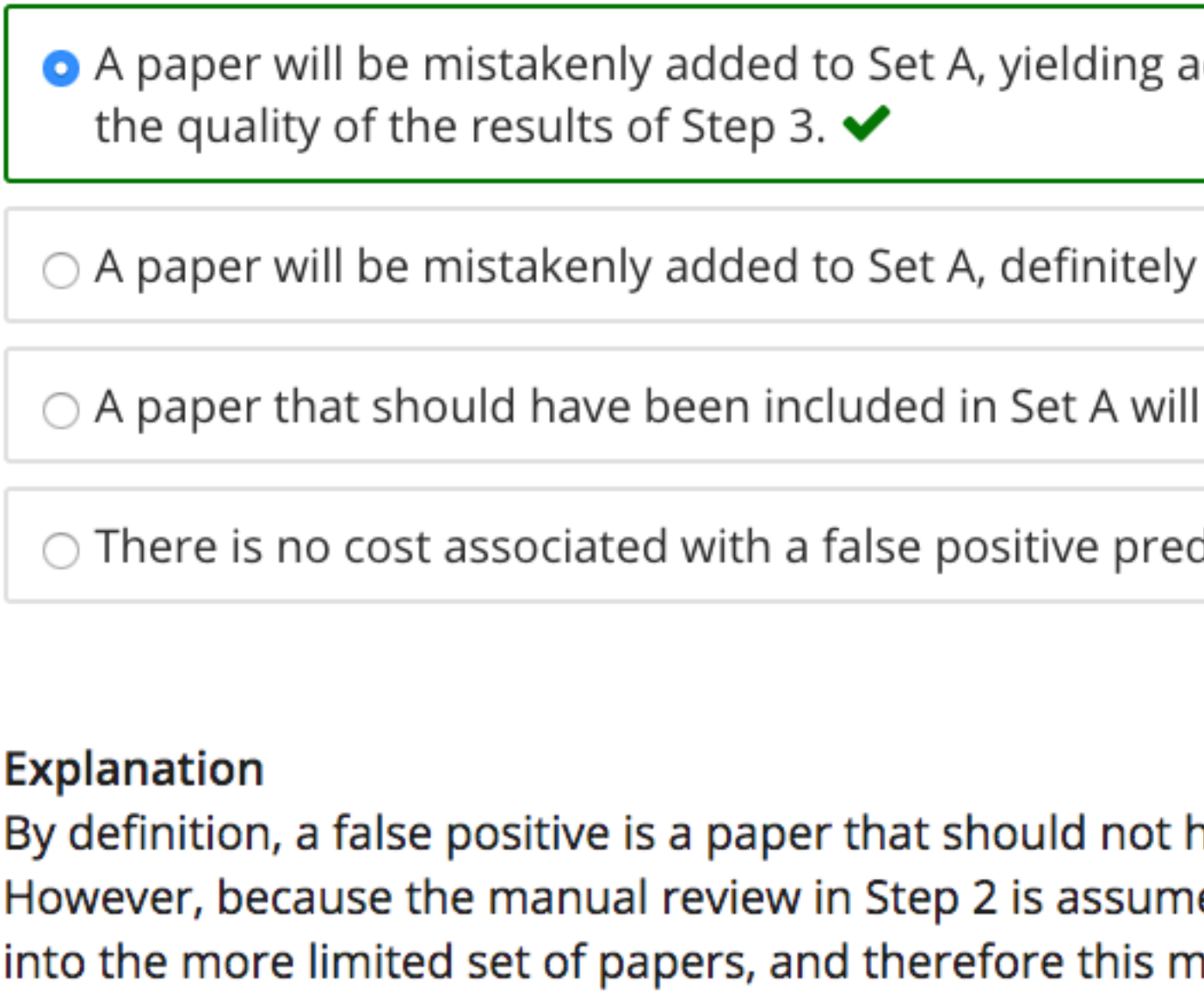
```
include_graphics('5.1.png')
```



5.2) Decision-Maker Tradeoffs

What is the cost associated with the model in Step 1 making a false positive prediction?

```
include_graphics('5.2.png')
```



5.3) Decision-Makder Tradeoffs

(Increasing the threshold leads to an **increase in specificity** and a **decrease in sensitivity**)

Decrease in sensitivity => decrease in true positive rate

Increase in specificity => increase in true negative rate

```
include_graphics('5.3.png')
```

- ○ A false positive is more costly than a false negative, greater than 0.5 for the machine learning model.
- ○ A false positive is more costly than a false negative, than 0.5 for the machine learning model.
- ◉ A false negative is more costly than a false positive, greater than 0.5 for the machine learning model. ✖
- ○ A false negative is more costly than a false positive, than 0.5 for the machine learning model. ✔

Explanation

A false negative might negatively affect the results of th nuisance (one additional paper that needs to be manua higher than the cost of a false positive, so much so that 1) and have two people manually review each search re cases where false negatives are more costly than false