Project 1

1. Data Preparation

```
library("xts")
```

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

```
## Loading required package: zoo
```

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

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# install.packages("dplyr")
library('dplyr') # data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:xts':
##
##     first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# library('readr') # input/output
# library('data.table') # data manipulation
# install.packages('tibble')
library('tibble') # data wrangling
```

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

```
# library('tidyr') # data wrangling
# library('stringr') # string manipulation
# library('forcats') # factor manipulation


# install.packages('tidytext')
library('tidytext')
```

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

```
# install.packages("magrittr")
library('magrittr')
# install.packages("tidyr")
library("tidyr")
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```
# install.packages("ggplot2")
library("ggplot2")

# install.packages("plotly")
library("plotly")
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
# install.packages("pamr")
library("pamr")
```

```
## Loading required package: cluster

## Loading required package: survival
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
source("../lib/multiplot.R")

# install.packages("gridExtra")
library("gridExtra")
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# We want to treat each column as characters, not factors, except for the author column.
spooky <- read.csv('../data/spooky.csv', colClasses = 'character')
spooky <- as.tibble(spooky)
spooky$author <- as.factor(spooky$author)
summary(spooky)
```

```
##       id                text           author
##  Length:19579       Length:19579       EAP:7900
##  Class :character   Class :character   HPL:5635
##  Mode  :character   Mode  :character   MWS:6044
```

2. Pronoun Occurence

```
# First use tidytext function to drop the punctuations and tokenize our file.
# We use bigrams because we want the word following the pronouns "he" and "she".
pronouns <- c("he","she")
data <- spooky %>%
  unnest_tokens(word, text, token = "ngrams", n = 2, to_lower = TRUE, drop = TRUE)


data_counts <- data %>%
  count(word, sort = TRUE) %>%
  separate(word, c("word1","word2"), sep = " ", remove = TRUE) %>%
  filter(word1 %in% pronouns) %>%
  filter(word2 != "he") %>%
  count(word1, word2, wt = n, sort = TRUE) %>%
  rename(counts = "nn")

all_po <- data.frame(
  count(data_counts, word1)
)
all_po <- all_po %>%
  rename(counts = "n",
         pronouns = "word1")

png("../figs/po1.png")
ggplot(all_po, aes(x="", y=counts, fill=pronouns))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y", start = 0) +
  ggtitle("pronoun occurence in total")

dev.off()
```

```
## pdf
##   2
```

```
data_counts_by_author <- data %>%
  count(author, word, sort = TRUE) %>%
  separate(word, c("word1","word2"), sep = " ", remove = TRUE) %>%
  filter(word1 %in% pronouns) %>%
  filter(word2 != "he") %>%
  rename(counts = "n")

MWS_po <- data_counts_by_author %>%
  filter(author == "MWS") %>%
  count(word1) %>%
  rename(pronoun = "word1", counts = "n")
HPL_po <- data_counts_by_author %>%
  filter(author == "HPL") %>%
  count(word1) %>%
  rename(pronoun = "word1", counts = "n")
EAP_po <- data_counts_by_author %>%
  filter(author == "EAP") %>%
  count(word1) %>%
  rename(pronoun = "word1", counts = "n")
```

```
p1 <- ggplot(MWS_po, aes(x="", y=counts, fill=pronoun))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y", start = 0) +
  ggtitle("MWS pronoun occurence")
p2 <- ggplot(HPL_po, aes(x="", y=counts, fill=pronoun))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y", start = 0) +
  ggtitle("HPL pronoun occurence")
p3 <- ggplot(EAP_po, aes(x="", y=counts, fill=pronoun))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y", start = 0) +
  ggtitle("EAP pronoun occurence")

layout <- matrix(c(1,2,3), 3,1, byrow = TRUE)
png("../figs/po2.png")
multiplot(p1, p2, p3, layout = layout)
```

```
## Loading required package: grid
```

```
dev.off()
```

```
## pdf
##   2
```

3. Gender Actions

```
word_ratios <- data_counts %>%
    group_by(word2) %>%
    filter(sum(counts) > 10) %>%
    ungroup() %>%
    spread(word1, counts, fill = 0) %>%
    mutate_if(is.numeric, funs((. + 1) / sum(. + 1))) %>%
    mutate(logratio = log2(she / he)) %>%
    arrange(desc(logratio))

# Equally likely:
equal <- word_ratios %>%
  mutate(abslogratio = abs(logratio)) %>%
  arrange(abslogratio)
words <- equal$word2
freqs <- seq(69,1)

png("../figs/eq.png")
wordcloud(words, freqs, max.words = 30, vfont = c("sans serif","plain"), color = c("purple4", "red4", "
dev.off()
```

```
## pdf
##   2
```

```
# Large difference:

png("../figs/verbs1.png")
word_ratios %>%
    mutate(abslogratio = abs(logratio)) %>%
```

```
    group_by(logratio < 0) %>%
    top_n(15, abslogratio) %>%
    ungroup() %>%
    mutate(word = reorder(word2, logratio)) %>%
    ggplot(aes(word, logratio, color = logratio < 0)) +
    geom_segment(aes(x = word, xend = word,
                     y = 0, yend = logratio),
                 size = 1.1, alpha = 0.6) +
    geom_point(size = 3.5) +
    coord_flip() +
    labs(x = NULL,
         y = "Relative appearance after 'she' compared to 'he'",
         title = "Words paired with 'he' and 'she'",
         subtitle = "Women throw, sleep, and turn while men####") +
    scale_color_discrete(name = "", labels = c("More 'she'", "More 'he'")) +
    scale_y_continuous(breaks = seq(-3, 3),
                       labels = c("0.125x", "0.25x", "0.5x",
                                  "Same", "2x", "4x", "8x"))
dev.off()
```

```
## pdf
##   2
```

```
# Separately:
# EAP

EAP_word_ratios <- data_counts_by_author %>%
  filter(author == "EAP") %>%
  group_by(word2) %>%
  filter(sum(counts) > 5) %>%
  ungroup() %>%
  spread(word1, counts, fill = 0) %>%
  mutate_if(is.numeric, funs((. +1)/ sum(. +1))) %>%
  mutate(logratio = log2(she / he)) %>%
  arrange(desc(logratio)) %>%
  mutate(abslogratio = abs(logratio)) %>%
  group_by(logratio < 0) %>%
  top_n(15,abslogratio) %>%
  ungroup() %>%
  mutate(word = reorder(word2, logratio))

jpeg("../figs/EAPverbs.jpeg", quality = 100)
ggplot(EAP_word_ratios, aes(word, logratio, color = logratio < 0)) +
  geom_segment(aes(x=word, xend = word,
                   y = 0, yend = logratio),
               size = 1.1, alpha = 0.6) +
  geom_point(size = 3.5) +
  coord_flip() +
  labs(x = NULL,
       y = "Relative appearance after 'she' compared to 'he' in EAP's novels",
       title = "Words paired with 'he' and 'she'") +
  scale_color_discrete(name = "", labels = c("More 'she'", "More 'he'")) +
  scale_y_continuous(breaks = seq(-3, 3),
                     labels = c("0.125x", "0.25x", "0.5x",
```

```
                                  "Same", "2x", "4x", "8x"))
dev.off()
```

```
## pdf
##   2
```

```
# HPL

HPL_word_ratios <- data_counts_by_author %>%
  filter(author == "HPL") %>%
  group_by(word2) %>%
  filter(sum(counts) > 5) %>%
  ungroup() %>%
  spread(word1, counts, fill = 0) %>%
  mutate_if(is.numeric, funs((. +1)/ sum(. +1))) %>%
  mutate(logratio = log2(she / he)) %>%
  arrange(desc(logratio)) %>%
  mutate(abslogratio = abs(logratio)) %>%
  group_by(logratio < 0) %>%
  top_n(15,abslogratio) %>%
  ungroup() %>%
  mutate(word = reorder(word2, logratio))

jpeg("../figs/HPLverbs.jpeg", quality = 100)
ggplot(HPL_word_ratios, aes(word, logratio, color = logratio < 0)) +
  geom_segment(aes(x=word, xend = word,
                   y = 0, yend = logratio),
               size = 1.1, alpha = 0.6) +
  geom_point(size = 3.5) +
  coord_flip() +
  labs(x = NULL,
       y = "Relative appearance after 'she' compared to 'he' in HPL's novels",
       title = "Words paired with 'he' and 'she'") +
  scale_color_discrete(name = "", labels = c("More 'she'", "More 'he'")) +
  scale_y_continuous(breaks = seq(-3, 3),
                     labels = c("0.125x", "0.25x", "0.5x",
                                "Same", "2x", "4x", "8x"))
dev.off()
```

```
## pdf
##   2
```

```
# MWS

MWS_word_ratios <- data_counts_by_author %>%
  filter(author == "MWS") %>%
  group_by(word2) %>%
  filter(sum(counts) > 5) %>%
  ungroup() %>%
  spread(word1, counts, fill = 0) %>%
  mutate_if(is.numeric, funs((. +1)/ sum(. +1))) %>%
  mutate(logratio = log2(she / he)) %>%
  arrange(desc(logratio)) %>%
  mutate(abslogratio = abs(logratio)) %>%
  group_by(logratio < 0) %>%
```

```
  top_n(15,abslogratio) %>%
  ungroup() %>%
  mutate(word = reorder(word2, logratio))

jpeg("../figs/MWSverbs.jpeg", quality = 100)
ggplot(MWS_word_ratios, aes(word, logratio, color = logratio < 0)) +
  geom_segment(aes(x=word, xend = word,
                   y = 0, yend = logratio),
               size = 1.1, alpha = 0.6) +
  geom_point(size = 3.5) +
  coord_flip() +
  labs(x = NULL,
       y = "Relative appearance after 'she' compared to 'he' in MWS's novels",
       title = "Words paired with 'he' and 'she'") +
  scale_color_discrete(name = "", labels = c("More 'she'", "More 'he'")) +
  scale_y_continuous(breaks = seq(-3, 3),
                     labels = c("0.125x", "0.25x", "0.5x",
                                "Same", "2x", "4x", "8x"))
dev.off()
```

```
## pdf
##   2
```

4. NSC experiments

```
# Reformatting our data file to meet the requirements of the pamr package.

gender <- data_counts_by_author %>%
  spread(word1, counts, fill = 0)
temp_male <- gender %>%
  spread(word2, he, fill = 0) %>%
  group_by(author) %>%
  summarise_if(is.numeric,sum) %>%
  ungroup() %>%
  rename(gender = "she") %>%
  mutate(gender = "male")
temp_female <- gender %>%
  spread(word2, she, fill = 0) %>%
  group_by(author) %>%
  summarise_if(is.numeric,sum) %>%
  ungroup() %>%
  rename(gender = "he") %>%
  mutate(gender = "female")
gender <- rbind(temp_male, temp_female)
gender$author <- as.character(gender$author)
first_row <- c("","",rep("word", 1134) %>%
  paste0(1:1134))
second_row <- c("", "", colnames(gender)[-c(1,2)])
gender <- rbind(first_row,second_row, gender) %>%
  as.matrix()
colnames(gender) <- NULL
gender <- t(gender)
```

```r
# Draw a graph of the table to get an idea of how it looks like
temp <- tableGrob(head(gender))
grid.newpage()
png("../figs/data.png")
grid.draw(temp)
dev.off()
```

```
## pdf
##   2
```

```r
write.table(gender, "../data/pam.txt", sep = "\t", row.names = FALSE, col.names = FALSE)

# Input the data file into the package
gender.data <- pamr.from.excel("../data/pam.txt", ncols = 8, sample.labels = TRUE)
```

```
##
## Read in  1134 genes
## Read in  6 samples
## Read in  6 sample labels
##
## Make sure these figures are correct!!
```

```r
# train the model
gender.train <- pamr.train(gender.data)
```

```
## 1234567891011121314151617181920212223242526272829 30
```

```r
# cross-validate the model
gender.results <- pamr.cv(gender.train, gender.data)
```

```
## 12Fold 1 :1234567891011121314151617181920212223242526272829 30
## Fold 2 :1234567891011121314151617181920212223242526272829 30
## Fold 3 :1234567891011121314151617181920212223242526272829 30
```

```r
## Plot the cross-validated error curves
png("../figs/cv.png")
pamr.plotcv(gender.results)
dev.off()
```

```
## pdf
##   2
```

```r
## Compute the confusion matrix for a particular model (threshold=7.0)
pamr.confusion(gender.results, threshold=2.2)
```

```
##        female male Class Error rate
## female      3    0               0
## male        0    3               0
## Overall error rate= 0
```

```r
## Plot the cross-validated class probabilities by class
png("../figs/cvprob.png")
pamr.plotcvprob(gender.results, gender.data, threshold=2.2)
dev.off()
```

```
## pdf
##   2
```

```
## Plot the class centroids
# These are the words that matters
png("../figs/plotcen.png")
pamr.plotcen(gender.train, gender.data, threshold=2.2)
dev.off()
```

```
## pdf
##   2
```

```
## Make a gene plot of the most significant words
png("../figs/geneplot.png")
pamr.geneplot(gender.train, gender.data, threshold=3)
dev.off()
```

```
## pdf
##   2
```

```
ID <- data.frame(pamr.listgenes(gender.train, gender.data, threshold = 2.2))[,3]
```

```
##       id       female-score male-score
##  [1,] word875  -2.3712      2.3712
##  [2,] word1098 -1.5872      1.5872
##  [3,] word799  -1.1431      1.1431
##  [4,] word1121 -1.0057      1.0057
##  [5,] word496  -0.6157      0.6157
##  [6,] word1033 -0.5367      0.5367
##  [7,] word141  -0.4406      0.4406
##  [8,] word892  -0.4337      0.4337
##  [9,] word1081 -0.3855      0.3855
## [10,] word124  -0.3266      0.3266
## [11,] word264  -0.3266      0.3266
## [12,] word974  -0.3266      0.3266
## [13,] word1024 -0.3266      0.3266
## [14,] word1131 -0.3266      0.3266
## [15,] word282  -0.243       0.243
## [16,] word92   -0.2398      0.2398
## [17,] word1107 -0.2057      0.2057
## [18,] word323  -0.1993      0.1993
## [19,] word660  -0.1957      0.1957
## [20,] word676  -0.1676      0.1676
## [21,] word153  -0.1225      0.1225
## [22,] word475  -0.1221      0.1221
## [23,] word955  -0.0617      0.0617
## [24,] word879  -0.0103      0.0103
```

```
ID <- as.numeric(ID)
wordlist <- c("said","who","read","would","held","took","came","seemed","was",
              "yet","threatened","stretched","descended","bore","did","became",
              "wished","drew","must","not","ceased","had","spoke","sat")
png("../figs/cloud.png")
# Most useful words given by NSC
wordcloud(wordlist,ID)
dev.off()
```

```
## pdf
##   2
```