Module 5 Homework

### Load the libraries

library(gutenbergr)  
library(topicmodels)  
library(tidyverse)  
library(tidytext)

### Download and prep the data

pride\_raw <- gutenberg\_download(1342)

## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest

## Using mirror http://aleph.gutenberg.org

#We can create a new column chapter that keeps track of which of the   
#chapters each line of text comes from. We can detect in that text column  
#when it says "Chapter".  
pride<-pride\_raw %>%  
 mutate(chapter=ifelse(str\_detect(text,"Chapter"),text,NA))%>%  
 #fill down the new column "chapter" until it gets to a new value   
 fill(chapter)%>%  
 #remove the NA that comes at the end of the last chapter.  
 filter(chapter != "NA") %>%  
 #set the factor levels by the order you find them in the text  
 mutate(chapter = factor(chapter, levels = unique(chapter)))  
#Now the chapters are in order   
#Transform the text data set to the tidy text dataframe, filter out the   
#stop words, and count the most common words.  
tidy\_pride <- pride %>%  
#create a new column to track which line every word is coming from  
 mutate(line = row\_number()) %>%  
#unnest the text column into the word column   
 unnest\_tokens(word, text) %>%  
#filter out the stop words   
anti\_join(stop\_words)

## Joining, by = "word"

#count the words and identify the most common words   
tidy\_pride%>%  
 count(word,sort = TRUE)

## # A tibble: 6,029 x 2  
## word n  
## <chr> <int>  
## 1 elizabeth 597  
## 2 darcy 373  
## 3 bennet 294  
## 4 miss 283  
## 5 jane 263  
## 6 bingley 257  
## 7 time 203  
## 8 lady 183  
## 9 sister 179  
## 10 wickham 162  
## # ... with 6,019 more rows

#notice the number of times we have specific names repeted in our dataset.  
Names <- c("elizabeth", "bennet", "miss", "lady", "dear", "darcy", "jane",   
"bingley", "catherine", "wickham", "kitty", "lydia", "collins", "gardiner",   
"charlotte", "lucas", "lizzy", "colonel", "forster", "hurst", "sir",   
"william", "eliza", "darcy's", "reynaolds", "fitzwilliam")  
#remove the outlier words from the dataset  
tidy\_pride <- pride %>%  
 mutate(line = row\_number()) %>%  
 unnest\_tokens(word, text) %>%  
 anti\_join(stop\_words)%>%  
 filter(!word %in% Names)

## Joining, by = "word"

tidy\_pride

## # A tibble: 33,779 x 4  
## gutenberg\_id chapter line word   
## <int> <fct> <int> <chr>   
## 1 1342 " Chapter 1" 1 chapter  
## 2 1342 " Chapter 1" 1 1   
## 3 1342 " Chapter 2" 3 chapter  
## 4 1342 " Chapter 2" 3 2   
## 5 1342 " Chapter 3" 5 chapter  
## 6 1342 " Chapter 3" 5 3   
## 7 1342 " Chapter 4" 7 chapter  
## 8 1342 " Chapter 4" 7 4   
## 9 1342 " Chapter 5" 9 chapter  
## 10 1342 " Chapter 5" 9 5   
## # ... with 33,769 more rows

### Implement Topic Modeling (LDA)

Cast the tidied dataset to a dtm:

library(topicmodels)  
pride\_dtm<-tidy\_pride %>%  
 count(chapter, word) %>%  
 cast\_dtm(chapter, word, n)

topic\_model<-LDA(pride\_dtm, k=5, control = list(seed = 1234))  
topic\_model

## A LDA\_VEM topic model with 5 topics.

## 3B - Display top 15 words of each topic

terms(topic\_model, k=15)

## Topic 1 Topic 2 Topic 3 Topic 4 Topic 5   
## [1,] "time" "chapter" "letter" "sister" "sister"   
## [2,] "sister" "day" "time" "family" "time"   
## [3,] "house" "time" "hope" "replied" "netherfield"  
## [4,] "replied" "ladies" "father" "day" "love"   
## [5,] "walk" "sisters" "family" "manner" "family"   
## [6,] "feelings" "evening" "marriage" "pleasure" "day"   
## [7,] "chapter" "daughters" "longbourn" "time" "hope"   
## [8,] "looked" "girls" "sister" "heard" "mother"   
## [9,] "moment" "visit" "cried" "feelings" "longbourn"   
## [10,] "ladies" "till" "day" "friend" "happy"   
## [11,] "pleasure" "friend" "married" "morning" "friend"   
## [12,] "manner" "party" "mother" "world" "subject"   
## [13,] "family" "hear" "till" "character" "letter"   
## [14,] "subject" "family" "happy" "pride" "happiness"   
## [15,] "heard" "gentlemen" "aunt" "told" "father"

## 3C - Interpret the results

I chose k=5 for the number of topics. Topic 1 - The noticeable words are “walk”, “feelings”, “pleasure”, and “house”. Sounds like bunch of people hanging out just having a subdued time. Topic 2: “Evening”, “visit”, “Party”, “gentlemen”. Sounds like a hip and happening time is afoot! Topic 3: “Father” , “marriage”, “married”, “cried”, “longbourn”, “hope”. After all of that partying in Topic 2, sounds like wedding bells with the mysterious Longbourn. Topic 4: “Replied”, “manner”, “pleasure”, “feelings”, “pride”, “world”, “character”. Sounds like the heavy, dramatic happenings are occuring. Topic 5: “Netherfield”, “love”, “hope”, “longbourn”, “happiness”, “happy”, “friend”. This is the upbeat theme. Everything must be beautiful if Netherfield and Longbourn are involved.

#### Word-topic probabilities (The beta matrix):

Let’s start with the tidy verb to calculate Beta. Beta is the probability assignments of words to topics. In other words, Beta is the probablity that a word contributes to a topic.

#install.packages("reshape2")  
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.0.5

##   
## Attaching package: 'reshape2'

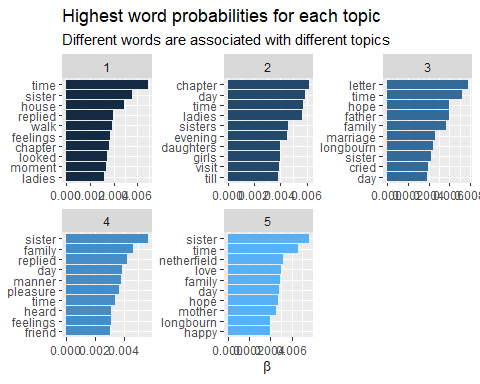
## The following object is masked from 'package:tidyr':  
##   
## smiths

topics <- tidy(topic\_model, matrix = "beta")  
topics

## # A tibble: 30,025 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 1 4.34e-95  
## 2 2 1 2.52e-12  
## 3 3 1 1.71e-96  
## 4 4 1 2.97e- 4  
## 5 5 1 1.71e-96  
## 6 1 chapter 3.61e- 3  
## 7 2 chapter 6.18e- 3  
## 8 3 chapter 3.09e- 3  
## 9 4 chapter 2.60e- 3  
## 10 5 chapter 3.00e- 3  
## # ... with 30,015 more rows

topics%>%  
 #let's group\_by each topic  
 group\_by(topic)%>%  
 #take the top 10 words in each topic  
 top\_n(10)%>%  
 ungroup%>%  
 mutate(term = reorder\_within(term, beta, topic)) %>%  
 ggplot(aes(term, beta, fill = topic)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 coord\_flip() +  
 labs(x = NULL, y = expression(beta),  
 title = "Highest word probabilities for each topic",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta

 ## 4C Interpret the results Topic 1 looks like a bunch of women Just hanging out and chilling. Walking around the house. Good stuff. Topic 2: There are evening visits that all the women are interested in. Topic 3: Father and Marriage are involved. Dad gets a shout-out, probably to pay for the wedding, which is probably why “cried” is listed. “letter” is for the invitation. Topic 4: I don’t see a lot of excitement and mayhem in this topic. Topic 5: Netherfield and Longbourn are part of the family, making everyone happy. Love is involved.

#### Document-topic probability (The gamma matrix):

LDA also models each document as a mixture of topics. With the matrix = “gamma” argument in tidy(), you can examine the per-document-per-topic probabilities, called γ (“gamma”). gamma is the probablity that a document contributes to a topic.

td\_gamma <- tidy(topic\_model, matrix = "gamma", document\_names = rownames(pride\_dtm))  
td\_gamma

## # A tibble: 610 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 " Chapter 1" 1 0.0151  
## 2 " Chapter 2" 1 0.0151  
## 3 " Chapter 3" 1 0.0151  
## 4 " Chapter 4" 1 0.939   
## 5 " Chapter 5" 1 0.939   
## 6 " Chapter 6" 1 0.0151  
## 7 " Chapter 7" 1 0.0151  
## 8 " Chapter 8" 1 0.939   
## 9 " Chapter 9" 1 0.939   
## 10 " Chapter 10" 1 0.0151  
## # ... with 600 more rows

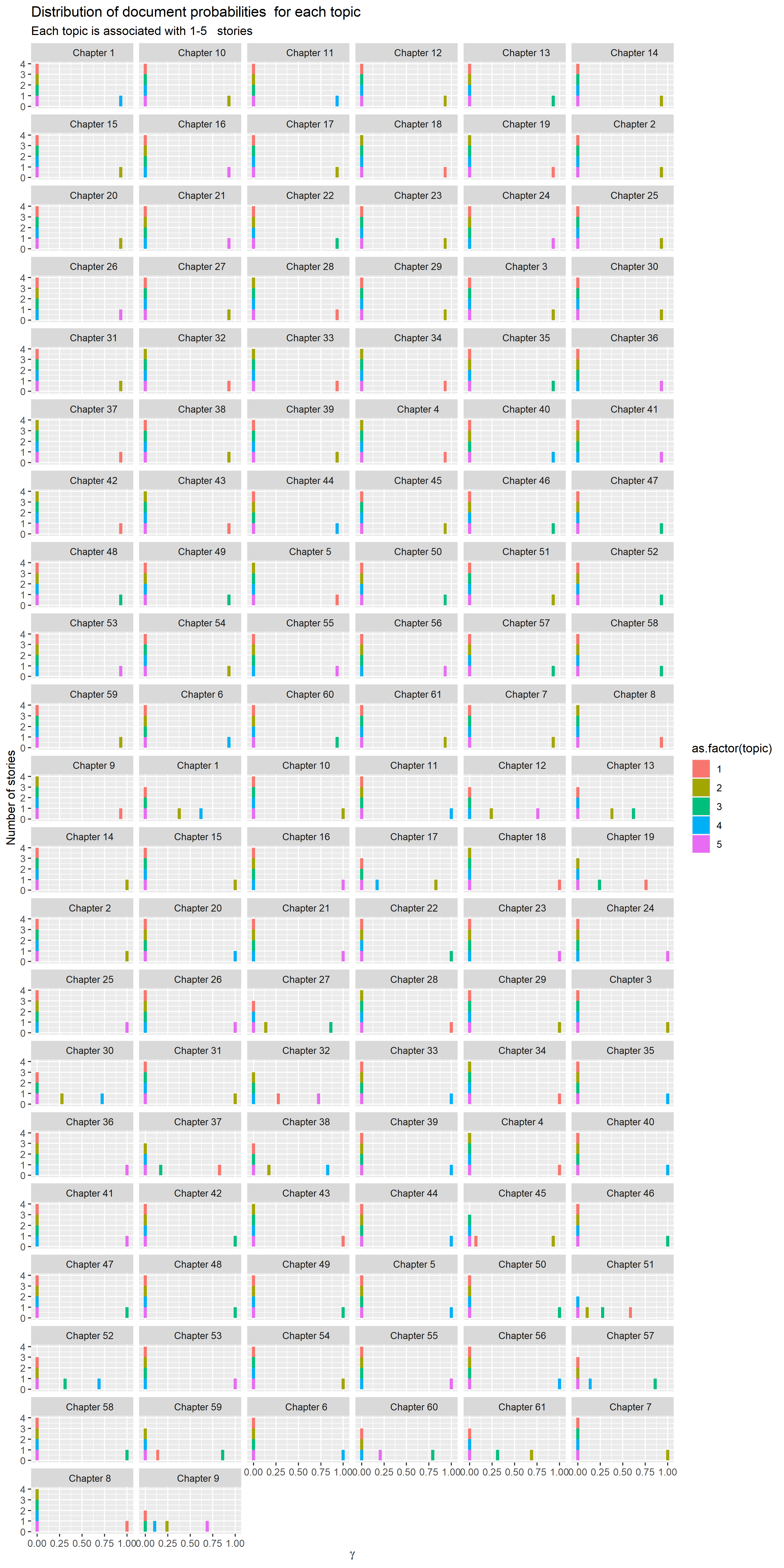
Each of these values is an estimated proportion of words from that document that are generated from that topic.For example, the model estimates that about 99% of the words in ADVENTURE V. were generated from topic 1.

We can plot the results. The y-axis is plotted on a log scale; γ runs from 0 to 1.

p1 <- ggplot(td\_gamma, aes(gamma, fill = as.factor(topic))) +  
 #a histogram of gamma  
 #we don't need to see the legend  
 geom\_histogram() +  
 #show the graphs in three columns  
 facet\_wrap(~ document, ncol = 6) +  
 labs(title = "Distribution of document probabilities for each topic",  
 subtitle = "Each topic is associated with 1-5 stories",  
 y = "Number of stories", x = expression(gamma))  
  
ggsave(p1, filename = "chaptertopics.png", height = 20, width = 10)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

knitr::include\_graphics("chaptertopics.png")



## 5C Interpret the results