Gospel of John

A textual analysis

Adam Wells

11/16/2020

Introduction

The Gospel of John is the last of the four gospels in the Christian Canon and an essential resource for understanding the development of sectarian religious communities in early Christianity. The tools of textual analytics can provide insight into the narrative structure and themes of this important text. To that end, the first part of this report will examine the most common words and two-word phrases in the gospel. The second part focuses on sentiment analysis of the gospel, examining the distribution of sentiments expressed and the development of positive and negative sentiments across the plot of the gospel. The third part uses topic modelling to explore latent themes in the text.

Exploratory Procedure

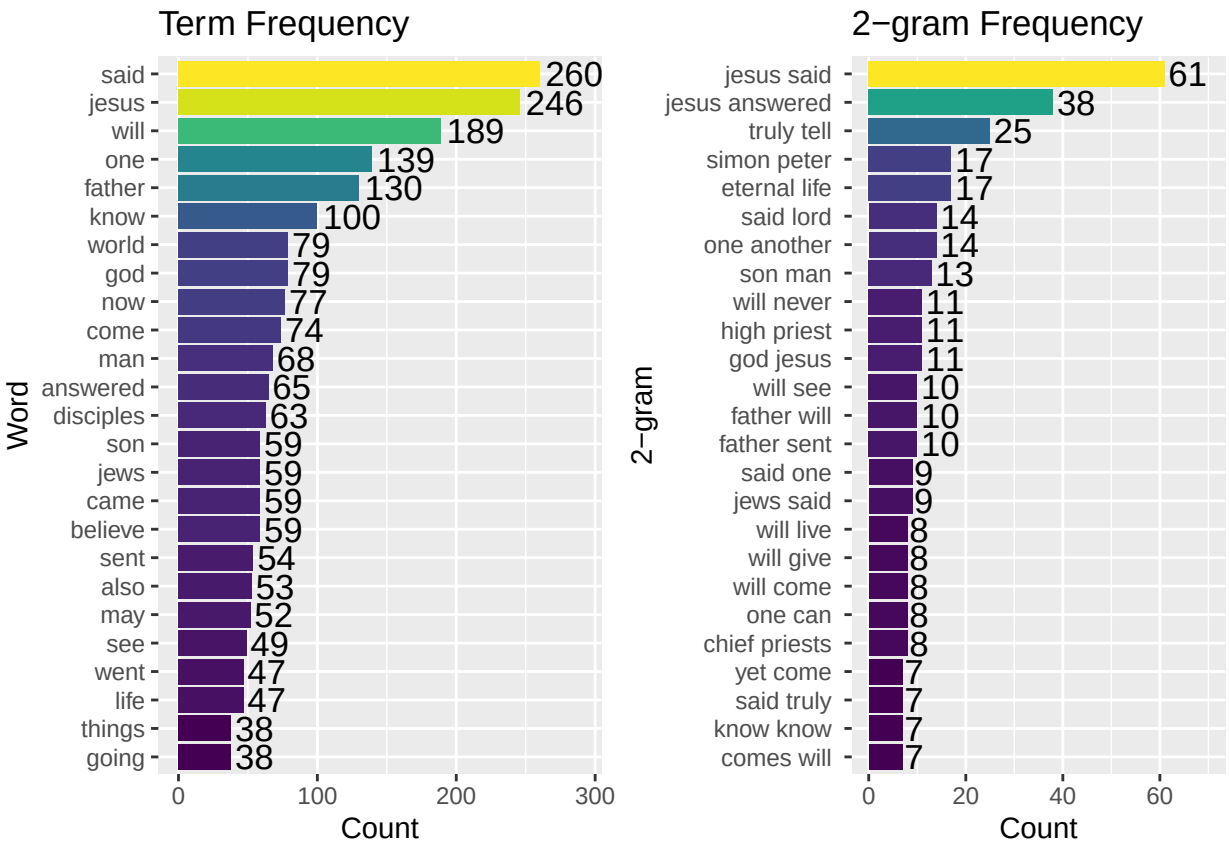
The New Revised Standard Version of the Gospel of John was scraped from www.biblegateway.com. Since each chapter was contained on a separate sub page, I created a loop to scrape data from all the relevant pages. I then excluded passage headings, which were added by the translators, and aggregated the text by chapter. To prepare the text for exploratory analysis, I removed all numbers, punctuation, non-English characters, extra white space, and stop words. I also converted all text to lowercase. I then examined term frequency and 2-gram frequency to get a sense of important words and phrases.

To prepare the text for sentiment analysis, I removed non-English characters and extra white space, but I left the sentence structure intact. I used the NRC sentiment dictionary, which provides measures of various sentiments (e.g., trust, anger, and fear), to get a sense of the variety of sentiments expressed in the fourth gospel. I also used the Bing sentiment dictionary to explore the development of positive and negative sentiments across the plot of the Gospel of John.

Finally, I used a topic modeling procedure to examine groups (or topics) occurring in the text. To prepare the text for topic modeling, I removed all numbers, punctuation, non-English characters, extra white space, and stop words. I converted the text to lowercase and stemmed every term. I then used the "stm" package to determine the most suitable number of topics and to construct the topic models. Finally, I examined the most representative chapters for each topic to get a sense of thematically linked sections of the gospel.

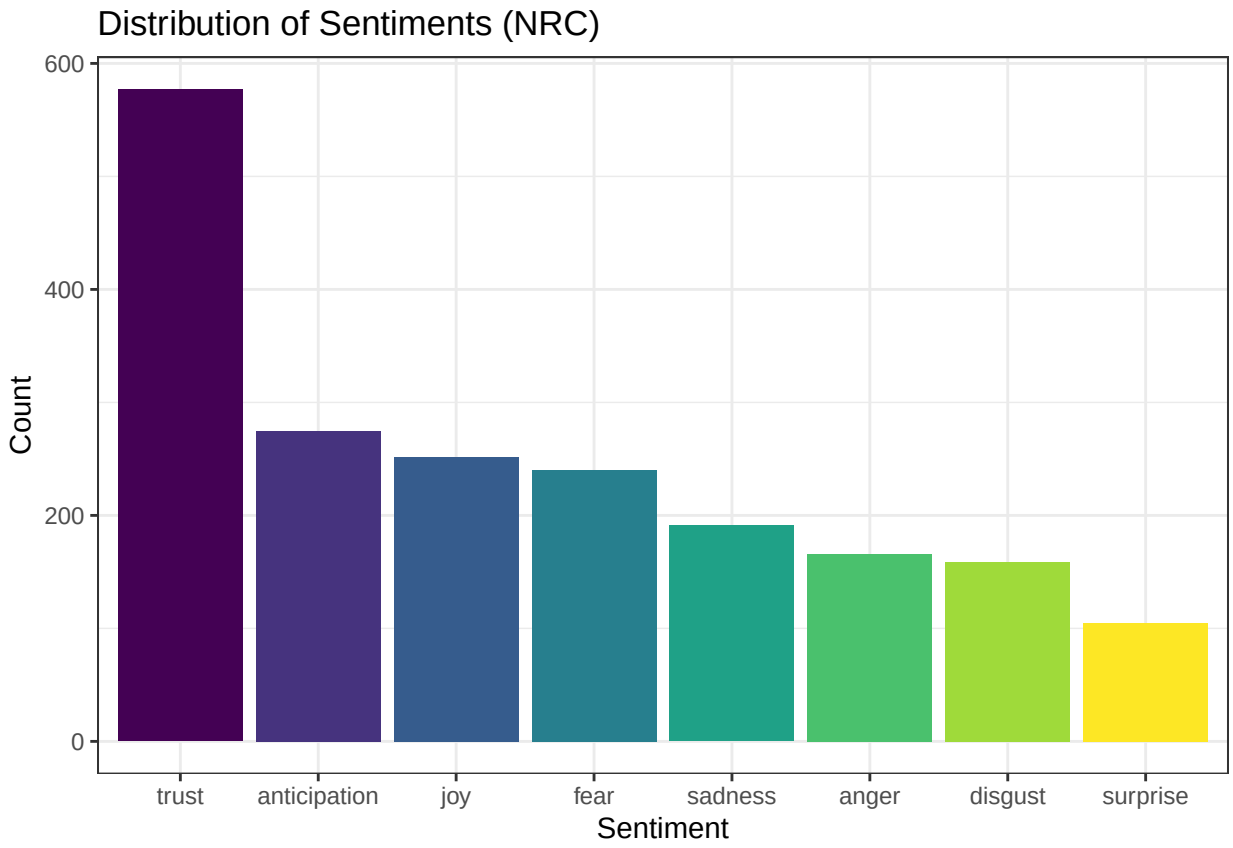
Descriptive Analysis

The following plots show the most frequent individual terms and term pairs (2-grams) in the Gospel of John. Many of the major themes of the gospel are evident here: eternal life, Jesus' identity as the Son of Man, an emphasis on Jewish leadership (high priests and chief priests), and the use of the term "Jews" to characterize opposition to Jesus. Additionally, the gospel routinely uses words having to do with knowledge, sight, and discipleship: know, see, believe, disciples, Simon Peter.

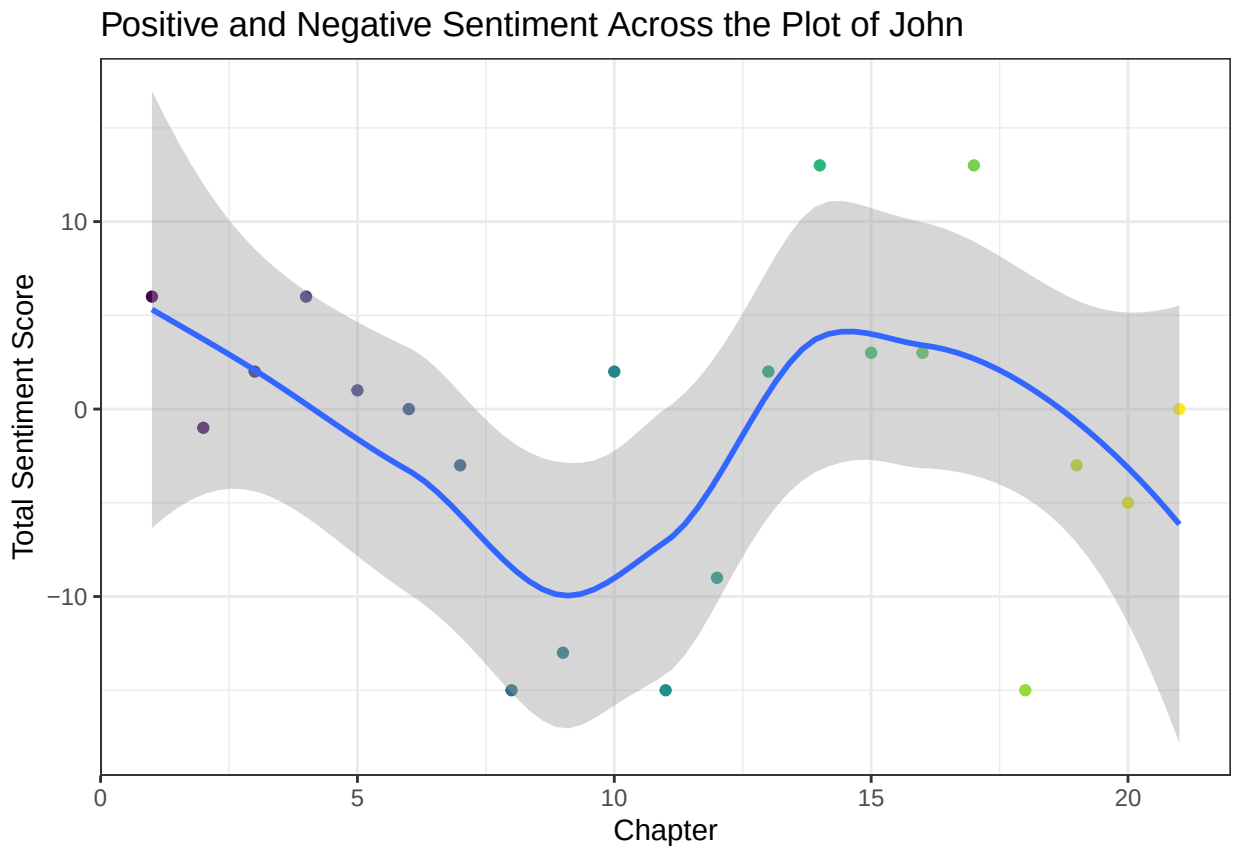


Sentiment Analysis

Sentiment analysis examines the type and frequency of sentiments expressed in a corpus. The first plot below shows the distribution of sentiments expressed in the Gospel of John. The most common sentiments are trust, anticipation, joy, and fear. The second plot below shows the development of positive and negative sentiments across each chapter of the Gospel of John. The gospel starts out positively and then quickly turns negative as the tension between Jesus and Jewish authorities builds. Sentiment begins to trend in a positive direction in chapter 10, which recounts the parable of the "good shepherd." The positive trend continues through chapter 15. Chapters 18, 19, and 20 focus on Jesus' arrest, trial, and execution. Accordingly, there is a decidedly negative turn in sentiment toward the end of the text.

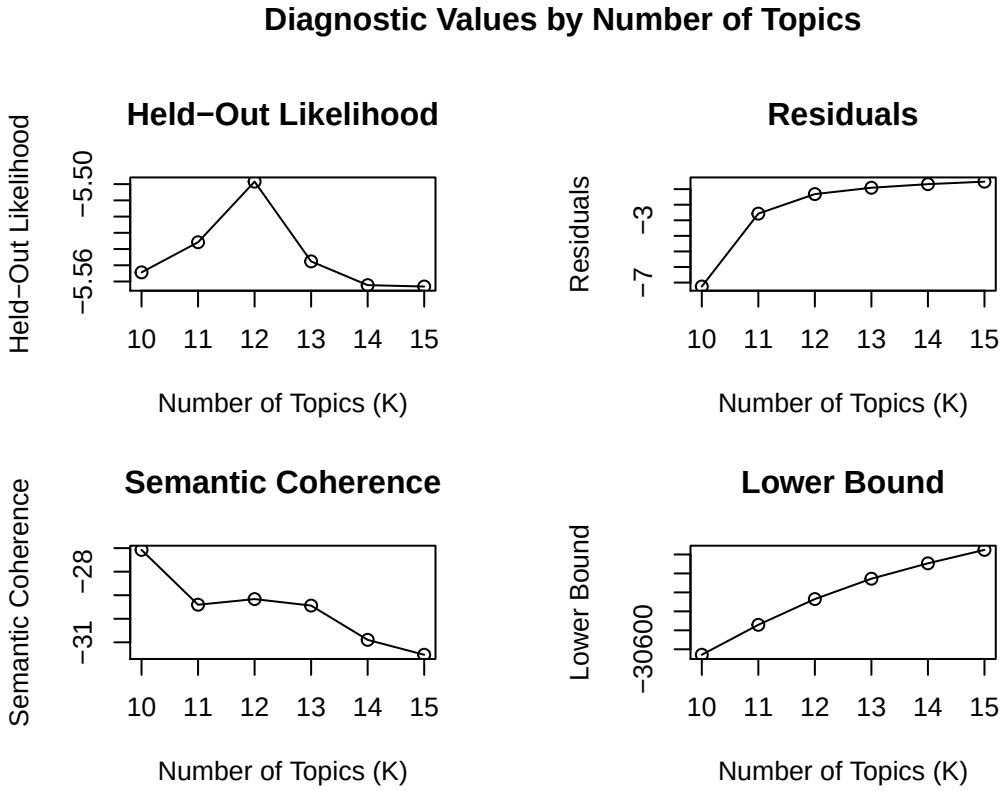

Distribution of Sentiments (NRC)
600
400
200
0
Count
trust
anticipation
joy
fear
sadness
anger
disgust
surprise
Sentiment

\


Positive and Negative Sentiment Across the Plot of John
10
0
−10
Total Sentiment Score
0
5
10
15
20
Chapter

Topic Model

Topic modeling is an unsupervised machine learning technique that attempts to discover latent themes or structures in a text by detecting groups of words that best characterize thematically-linked sections of a corpus. The first task in constructing a topic model is to determine the appropriate number of topics. The plots below show the held-out log likelihood, residuals, semantic coherence, and lower bound (an internal measure of fit) for a range of topic numbers (K). A model with 12 topics maximizes held-out log likelihood, while also producing low residuals, and decent semantic coherence and lower bound values.



The following chart summarizes the selected model. "Most Probable Words" refers to the words with the highest probability of appearing in a given topic. "FREX" (i.e., FRequency and EXclusivity) is a measure similar to TF-IDF; it balances word frequency and topic exclusivity. FREX therefore gives a sense of the words that define a particular topic over-and-against other topics. Finally, I have linked the topics to relevant chapters in John by quantifying the proportion of words assigned to particular topics within a given chapter. The results have interesting interpretive ramifications. For example, topic 5 suggests a thematic link between chapter 2, where Jesus turns water into wine at a wedding, and chapter 18, where Jesus is put on trial and sentenced to death. The wedding scene in chapter 2 is often read theologically as a reference to Christian use of wine in Eucharist rituals, which are themselves symbolic references to Jesus' suffering, death, and resurrection. Topic 5 makes the thematic link between the wedding scene and Jesus' trial clear. Similarly, topic 1 implies a link between chapter 13, where Jesus washes the disciples' feet, and chapter 21, where Jesus appears to the disciples after the resurrection. Perhaps this topic has to do with the self-sacrificial nature of discipleship, or the redemption of Peter, who has a large role in both chapters.

Topic	Most Probable Words	FREX	John Chapters
1	jesu, said, discipl, lord, peter, love, simon	tomb, simon, lord, fish, feet, peter, piec	21, 13
2	jesu, said, come, will, bread, life, father	bread, eat, woman, worship, drink, sea, boat	4, 6
3	world, mai, given, word, sent, father, love	given, world, glorifi, mai, behalf, word, known	17, 15
4	father, will, sheep, abid, know, fruit, love	abid, fruit, sheep, bear, branch, command, hate	10, 15
5	jesu, said, jew, discipl, pilat, priest, ask	pilat, high, wine, priest, soldier, mother, king	18, 2
6	father, will, son, man, life, jesu, come	well, honor, accept, made, judgment, life, hear	5
7	jesu, said, will, come, believ, lazaru, went	lazaru, martha, mari, let, sister, crowd, dead	11, 12
8	god, born, can, son, come, believ, thing	born, can, must, baptiz, light, john, heaven	3
9	said, jesu, father, god, will, know, come	philip, prophet, light, nathanael, descend, john, judg	8, 1
10	man, said, know, sai, jesu, blind, come	blind, man, open, pharise, sai, teach, crowd	9, 7
11	will, father, said, come, world, know, love	littl, will, longer, father, heart, love, leav	14, 16
12	saw, said, came, believ, dai, discipl, god	saw, dai, first, came, awai, among, seen	20, 1

Conclusion

The preceding analysis yielded a number of important insights about the Gospel of John. First, many of the major themes of the Gospel are evident in word and n-gram frequency: knowledge, sight discipleship, eternal life, Jesus' identity as the Son of Man, an emphasis on Jewish leadership (high priests and chief priests), and the use of the term "Jews" to characterize opposition to Jesus. Second, sentiment analysis showed that the most common sentiment expressed in the gospel is trust. Additionally, by tracking positive and negative sentiments across the text, we get a better sense of the progression of the gospel's plot. Finally, topic modeling reveals important thematic connections between different parts of the gospel. For example, topic 5 suggests a thematic link between the wedding at Cana and Jesus' trial. In sum, this analysis serves as an excellent jumping off point for further interpretive work.

Appendix: Code

```
knitr::opts_chunk$set(echo = F,warning = F,message = F)
library(tidyverse)
library(xml2)
library(stringr)
library(rvest)
library(readxl)
library(purrr)
library(tm)
library(stm)
library(stringr)
library(textstem)
library(tidytext)
library(wordcloud)
library(syuzhet)
library(viridis)
library(ngram)
library(gridExtra)
library(kableExtra)
setwd("/Users/adamwells/Desktop/STAT5526/HW5_6")

John<-list()
for (i in c(1:21)){
John[[i]]<-read_html(paste0("https://www.biblegateway.com/passage/?search=John%20",i,"&version=NRSV"))
}

John_data<-data.frame()
for(i in c(1:21)){
J<-John[[i]]%>%
  html_nodes(xpath="//h3/following-sibling::p")%>%
  html_text()%>%
  trimws()%>%
  as.data.frame()%>%
  select(text=1)%>%
  mutate(text = strsplit(as.character(text), "[0-9]+")) %>%
  unnest(text)%>%
  filter(text!="")%>%
  mutate(chapter=i)

John_data<-bind_rows(John_data,J)
}

#Clean and Stem text
John_data_clean<-John_data%>%
  group_by(chapter)%>%
  summarize(text=tolower(gsub("\\b\\w{1}\\b"," ",removePunctuation(gsub("[^a-zA-Z]", " ",
                       (paste(unlist(text), collapse =" "))))), chapter=unique(chapter))

John_data_clean$text<-trimws(stripWhitespace(removeWords(John_data_clean$text,stopwords())))
John_data_clean_vec<-as.character(stem_strings(John_data_clean$text))
```

```r
#Term Frequency
John_tokenized<-str_split(John_data_clean$text,pattern=" ")%>%
  unlist

Unique<-unique(John_tokenized)
tf_John <- sapply(1:length(Unique), function(i){sum(John_tokenized == Unique[i])})
John1<-data.frame("Words"=Unique, "Count"=tf_John)%>%
  arrange(desc(Count))%>%
  mutate(Proportion=round(Count/sum(Count),5))

tf_John_plot<-ggplot(head(John1, 25), aes(x = reorder(Words, Count), y = Count)) +
  geom_bar(stat = "identity", aes(fill = Count)) +
  labs(title = "Term Frequency", x = "Word", y = "Count") +
  coord_flip() + theme(legend.position = "none") +
  scale_fill_viridis() +
  geom_text(stat = "identity", aes(label = Count), hjust = -0.1, size = 4.5) +
  expand_limits(y = 290)

#n-grams
John_ngram2 <- ngram(concatenate(John_data_clean$text), 2, sep = " ")%>%
  get.phrasetable()

ngram_plot<-ggplot(head(John_ngram2%>%arrange(desc(freq)), 25),
                   aes(x = reorder(ngrams, freq), y = freq)) +
  geom_bar(stat = "identity", aes(fill = freq)) +
  labs(title = "2-gram Frequency", x = "2-gram", y = "Count") +
  coord_flip() + theme(legend.position = "none") +
  scale_fill_viridis() +
  geom_text(stat = "identity", aes(label = freq), hjust = -0.1, size = 4.5) +
  expand_limits(y = 70)

grid.arrange(tf_John_plot,ngram_plot,ncol=2)

#Sentiment analysis using NRC and Bing
John_data_sent<-John_data$text%>%
  gsub("[^a-zA-Z]", " ",.)%>%
  stripWhitespace()%>%
  trimws()%>%
  get_nrc_sentiment()%>%
  bind_cols(Bing=get_sentiment(John_data$text,method="bing"),John_data)%>%
  rownames_to_column("Verse")%>%
  as.data.frame()

#Plot sentiment distribution
John_data_sent%>%
  select(c(2:9))%>%
  colSums()%>%
  as.data.frame()%>%
  setNames("count")%>%
  rownames_to_column("sentiment")%>%
  ggplot(aes(x=reorder(sentiment,-count),y=count,fill=reorder(sentiment,-count)))+
  geom_bar(stat="identity")+
```

```
  scale_fill_viridis(discrete=T)+
  labs(title="Distribution of Sentiments (NRC)",x="Sentiment",y="Count")+
  theme_bw()+
  theme(legend.position = "none")

#Plot sentiment across chapters
John_data_sent%>%
  group_by(chapter)%>%
  summarize(sent_chap=sum(Bing))%>%
  as.data.frame()%>%
  ggplot(aes(x=chapter,y=sent_chap))+
  geom_point(aes(color=as.factor(chapter)))+
  geom_smooth()+
  scale_color_viridis(discrete = T)+
  labs(title="Positive and Negative Sentiment Across the Plot of John",
       x="Chapter",
       y="Total Sentiment Score")+
  theme_bw()+
  theme(legend.position = "none")

#Create DTM
DTMJohn <- DocumentTermMatrix(VCorpus(VectorSource(John_data_clean_vec)))

#Prepare DTM for use with stm package
ingest<-readCorpus(DTMJohn,type="slam")
prep<-prepDocuments(ingest$documents,ingest$vocab)

#Determine K
search<-searchK(prep$documents,prep$vocab,K=c(10:15),proportion=0.2,heldout.seed = 1234)

#Best model w/ K=12
select<-selectModel(prep$documents,prep$vocab,K=12,max.em.its = 250,N=5,seed=0)
fit<-select$runout[[5]]
plot(search)

#Probability and FREX
topic_words<-labelTopics(fit)
tw<-cbind(topic_words$topicnums,topic_words$prob)%>%
  data.frame%>%
  within(., `Most Probable Words`<-paste(X2,X3,X4,X5,X6,X7,X8,sep=", "))%>%
  select(Topic=X1,`Most Probable Words`)

tw2<-cbind(topic_words$topicnums,topic_words$frex)%>%
  data.frame%>%
  within(., FREX<-paste(X2,X3,X4,X5,X6,X7,X8,sep=", "))%>%
  select(Topic=X1,FREX)

words_by_topic<-full_join(tw,tw2,by="Topic")

#Link topics to chapters
df<-(findThoughts(fit,texts=John_data_clean_vec,n=2,topics=c(1:12),thresh = 0.05))
docs<-df$index%>%
```

```
  reduce(rbind)
for (i in 1:nrow(docs)){
  docs[i,][duplicated(docs[i,])]<-""
}

colnames(docs)<-c("Chapter1","Chapter2")
rownames(docs)<-c(1:12)
docs<-as.data.frame(docs)%>%
  within(., `John Chapters`<-paste(Chapter1,Chapter2,sep=", "))%>%
  mutate(`John Chapters`=gsub(", $","",.$`John Chapters`))%>%
  mutate(`John Chapters`=gsub(", $","",.$`John Chapters`))%>%
  select(`John Chapters`)%>%
  rownames_to_column("Topic")

#Combine dfs
full_join(words_by_topic,docs,by="Topic")%>%
  kbl(booktabs=T)%>%
  kable_styling(latex_options = c("striped","scale_down"))
```