HarryPotterSA

Ghazalpreet Kaur

October 15, 2018

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

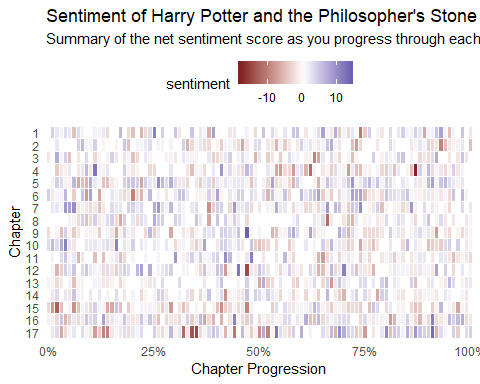
## Including Plots

You can also embed plots, for example:

## starting httpd help server ... done

## # A tibble: 6,598 x 1  
## sentence   
## <chr>   
## 1 the boy who lived mr. and mrs.   
## 2 dursley, of number four, privet drive, were proud to say that they were~  
## 3 they were the last people you'd expect to be involved in anything stran~  
## 4 mr.   
## 5 dursley was the director of a firm called grunnings, which made drills.   
## 6 he was a big, beefy man with hardly any neck, although he did have a ve~  
## 7 mrs.   
## 8 dursley was thin and blonde and had nearly twice the usual amount of ne~  
## 9 the dursleys had a small son called dudley and in their opinion there w~  
## 10 the dursleys had everything they wanted, but they also had a secret, an~  
## # ... with 6,588 more rows

## Joining, by = "word"



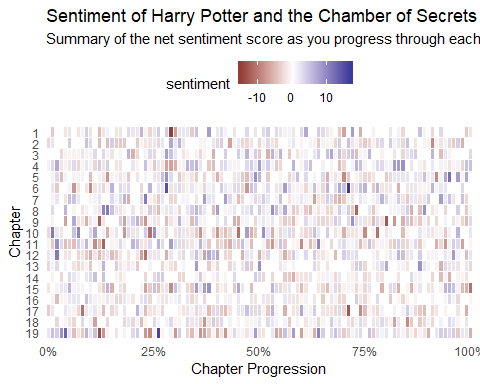
tibble(text= chamber\_of\_secrets)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 6,908 x 1  
## sentence   
## <chr>   
## 1 the worst birthday not for the first time, an argument had broken out ~  
## 2 mr.   
## 3 vernon dursley had been woken in the early hours of the morning by a lo~  
## 4 "\"third time this week!\""   
## 5 he roared across the table.   
## 6 "\"if you can't control that owl, it'll have to go!\""   
## 7 harry tried, yet again, to explain.   
## 8 "\"she's bored,\" he said."   
## 9 "\"she's used to flying around outside."   
## 10 "if i could just let her out at night -\" \"do i look stupid?\""   
## # ... with 6,898 more rows

chamber\_of\_secrets\_book<- tibble(chapter= 1: length(chamber\_of\_secrets),  
 text=chamber\_of\_secrets)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
View(chamber\_of\_secrets\_book)  
?get\_sentiments  
  
book\_2 <- chamber\_of\_secrets\_book %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num1 = 1:n(),  
 index = round(sentence\_num1/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter,index)%>%  
 summarize(sentiment=sum(score, na.rm=TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

#visualilzation  
ggplot(book\_2, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill=sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand=c(0,0))+  
 scale\_y\_discrete(expand = c(0,0))+  
 labs(x="Chapter Progression", y="Chapter")+  
 ggtitle("Sentiment of Harry Potter and the Chamber of Secrets", subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



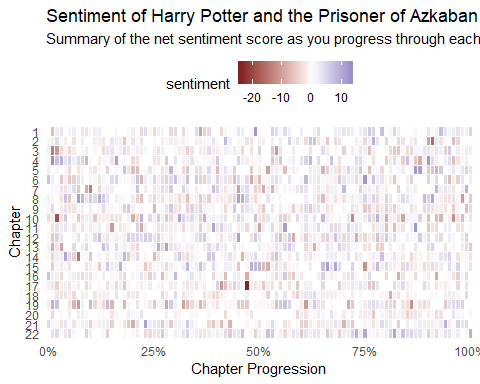
tibble(text= prisoner\_of\_azkaban)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 9,011 x 1  
## sentence   
## <chr>   
## 1 owl post harry potter was a highly unusual boy in many ways.   
## 2 for one thing, he hated the summer holidays more than any other time of~  
## 3 for another, he really wanted to do his homework but was forced to do i~  
## 4 and he also happened to be a wizard.   
## 5 it was nearly midnight, and he was lying on his stomach in bed, the bla~  
## 6 "harry moved the tip of his eagle-feather quill down the page, frowning~  
## 7 the quill paused at the top of a likely-looking paragraph.   
## 8 harry pushed his round glasses up the bridge of his nose, moved his fla~  
## 9 on the rare occasion that they did catch a real witch or wizard, burnin~  
## 10 the witch or wizard would perform a basic flame freezing charm and then~  
## # ... with 9,001 more rows

prisoner\_of\_azkaban\_book<- tibble(chapter= 1: length(prisoner\_of\_azkaban),  
 text=prisoner\_of\_azkaban)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
View(prisoner\_of\_azkaban\_book)  
?get\_sentiments  
  
book\_3 <- prisoner\_of\_azkaban\_book %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num1 = 1:n(),  
 index = round(sentence\_num1/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter,index)%>%  
 summarize(sentiment=sum(score, na.rm=TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

#visualilzation  
ggplot(book\_3, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill=sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand=c(0,0))+  
 scale\_y\_discrete(expand = c(0,0))+  
 labs(x="Chapter Progression", y="Chapter")+  
 ggtitle("Sentiment of Harry Potter and the Prisoner of Azkaban", subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



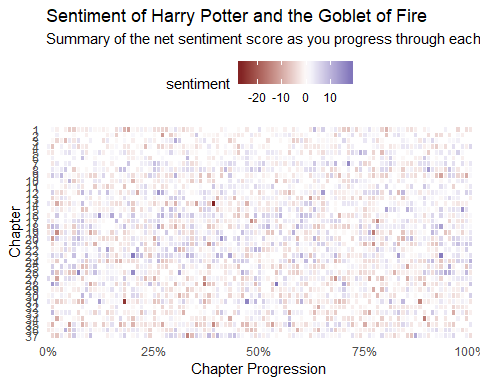
tibble(text= goblet\_of\_fire)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 15,467 x 1  
## sentence   
## <chr>   
## 1 "the villagers of little hangleron still called it \"the riddle house,\~  
## 2 it stood on a hill overlooking the village, some of its windows boarded~  
## 3 once a fine-looking manor, and easily the largest and grandest building~  
## 4 "the little hagletons all agreed that the old house was \"creepy.\""   
## 5 half a century ago, something strange and horrible had happened there, ~  
## 6 the story had been picked over so many times, and had been embroidered ~  
## 7 every version of the tale, however, started in the same place: fifty ye~  
## 8 the maid had run screaming down the hill into the village and roused as~  
## 9 "\"lying there with their eyes wide open!"   
## 10 cold as ice!   
## # ... with 15,457 more rows

goblet\_of\_fire\_book<- tibble(chapter= 1: length(goblet\_of\_fire),  
 text=goblet\_of\_fire)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
View(goblet\_of\_fire\_book)  
?get\_sentiments  
  
book\_4 <- goblet\_of\_fire\_book %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num1 = 1:n(),  
 index = round(sentence\_num1/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter,index)%>%  
 summarize(sentiment=sum(score, na.rm=TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

#visualilzation  
ggplot(book\_4, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill=sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand=c(0,0))+  
 scale\_y\_discrete(expand = c(0,0))+  
 labs(x="Chapter Progression", y="Chapter")+  
 ggtitle("Sentiment of Harry Potter and the Goblet of Fire", subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



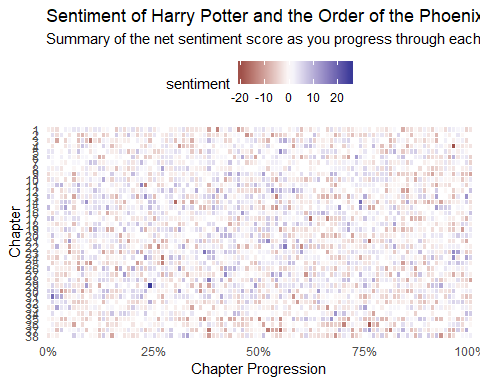
tibble(text= order\_of\_the\_phoenix)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 17,927 x 1  
## sentence   
## <chr>   
## 1 dudley demented the hottest day of the summer so far was drawing to ~  
## 2 cars that were usually gleaming stood dusty in their drives and lawns t~  
## 3 deprived of their usual car-washing and lawn-mowing pursuits, the inhab~  
## 4 the only person left outdoors was a teenage boy who was lying flat on h~  
## 5 he was a skinny, black-haired, bespectacled boy who had the pinched, sl~  
## 6 his jeans were torn and dirty, his t-shirt baggy and faded, and the sol~  
## 7 harry potters appearance did not endear him to the neighbours, who were~  
## 8 in fact, the only way he would be spotted was if his uncle vernon or au~  
## 9 on the whole, harry thought he was to be congratulated on his idea of h~  
## 10 he was not, perhaps, very comfortable lying on the hot, hard earth but,~  
## # ... with 17,917 more rows

order\_of\_the\_phoenix\_book<- tibble(chapter= 1: length(order\_of\_the\_phoenix),  
 text=order\_of\_the\_phoenix)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
View(order\_of\_the\_phoenix\_book)  
?get\_sentiments  
  
book\_5 <- order\_of\_the\_phoenix\_book %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num1 = 1:n(),  
 index = round(sentence\_num1/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter,index)%>%  
 summarize(sentiment=sum(score, na.rm=TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

#visualilzation  
ggplot(book\_5, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill=sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand=c(0,0))+  
 scale\_y\_discrete(expand = c(0,0))+  
 labs(x="Chapter Progression", y="Chapter")+  
 ggtitle("Sentiment of Harry Potter and the Order of the Phoenix", subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



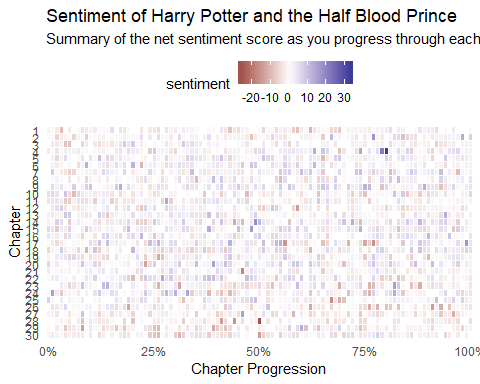
tibble(text= half\_blood\_prince)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 12,504 x 1  
## sentence   
## <chr>   
## 1 it was nearing midnight and the prime minister was sitting alone in his~  
## 2 he was waiting for a call from the president of a far distant country, ~  
## 3 the more he attempted to focus on the print on the page before him, the~  
## 4 this particular opponent had appeared on the news that very day, not on~  
## 5 the prime minister's pulse quickened at the very thought of these accus~  
## 6 how on earth was his government supposed to have stopped that bridge co~  
## 7 it was outrageous for anybody to suggest that they were not spending en~  
## 8 the bridge was fewer than ten years old, and the best experts were at a~  
## 9 and how dare anyone suggest that it was lack of policemen that had resu~  
## 10 or that the government should have somehow foreseen the freak hurricane~  
## # ... with 12,494 more rows

half\_blood\_prince\_book<- tibble(chapter= 1: length(half\_blood\_prince),  
 text=half\_blood\_prince)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
View(half\_blood\_prince\_book)  
?get\_sentiments  
  
book\_6 <- half\_blood\_prince\_book %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num1 = 1:n(),  
 index = round(sentence\_num1/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter,index)%>%  
 summarize(sentiment=sum(score, na.rm=TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

#visualilzation  
ggplot(book\_6, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill=sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand=c(0,0))+  
 scale\_y\_discrete(expand = c(0,0))+  
 labs(x="Chapter Progression", y="Chapter")+  
 ggtitle("Sentiment of Harry Potter and the Half Blood Prince", subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



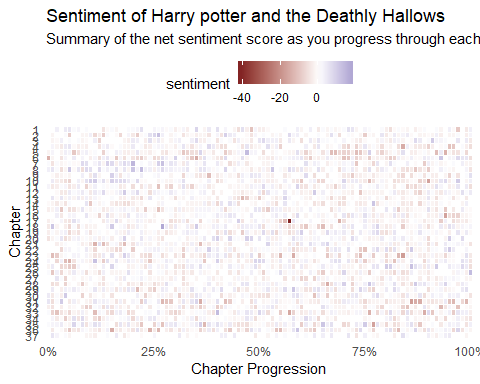
#tokenize text into sentences  
  
#sentiment Analysis with larger Units.  
#tokenize the sentence  
  
tibble(text=deathly\_hallows)%>%  
 unnest\_tokens(sentence, text, token = "sentences")

## # A tibble: 15,138 x 1  
## sentence   
## <chr>   
## 1 the two men appeared out of nowhere, a few yards apart in the narrow, m~  
## 2 "for a second they stood quite still, wands directed at each other's ch~  
## 3 "news?\""   
## 4 "asked the taller of the two.\""   
## 5 "the best,\" replied severus snape."   
## 6 the lane was bordered on the left by wild, low-growing brambles, on the~  
## 7 "the men's long cloaks flapped around their ankles as they marched.\""   
## 8 "thought i might be late,\" said yaxley, his blunt features sliding in ~  
## 9 "\"it was a little trickier than i expected."   
## 10 but i hope he will be satisfied.   
## # ... with 15,128 more rows

?tibble  
  
#lets go ahead and break up the philosopher's stone text by chapter and sentence  
  
ps\_sentences<- tibble(chapter=1: length(deathly\_hallows),  
 text=deathly\_hallows)%>%  
 unnest\_tokens(sentence, text, token = "sentences")  
  
#This will allow us to assess the net sentiment  
#by chapter and by sentence. First, we need to track the sentence   
#numbers and then I create an index that tracks the progress   
#through each chapter. I then unnest the sentences by words.   
#This gives us a tibble that has individual words by sentence   
#within each chapter.  
  
# Join "afinn" lexicon and compute the net sentiment score  
  
book\_sent <- ps\_sentences %>%  
 group\_by(chapter)%>%  
 mutate(sentence\_num= 1:n(),  
 index=round(sentence\_num/n(),2))%>%  
 unnest\_tokens(word, sentence)%>%  
 inner\_join(get\_sentiments("afinn"))%>%  
 group\_by(chapter, index)%>%  
 summarize(sentiment= sum(score, na.rm = TRUE))%>%  
 arrange(desc(sentiment))

## Joining, by = "word"

View(book\_sent)  
  
#Now, as before, I join the AFINN lexicon and compute the net sentiment score for each chapter.The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.  
  
#Visualize using ggplot  
  
ggplot(book\_sent, aes(index, factor(chapter, levels = sort(unique(chapter), decreasing = TRUE)), fill = sentiment))+  
 geom\_tile(color = "white")+  
 scale\_fill\_gradient2()+  
 scale\_x\_continuous(labels = scales::percent, expand = c(0,0))+  
 scale\_y\_discrete(expand=c(0,0))+  
 labs(x="Chapter Progression", y = "Chapter")+  
 ggtitle("Sentiment of Harry potter and the Deathly Hallows",  
 subtitle = "Summary of the net sentiment score as you progress through each chapter")+  
 theme\_minimal()+  
 theme(panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 legend.position = "top")



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.