Analyzing Amazon Product Reviews

## Load the required libraries

#Clean Rstudio environment   
rm(list=ls())

## Read in the data

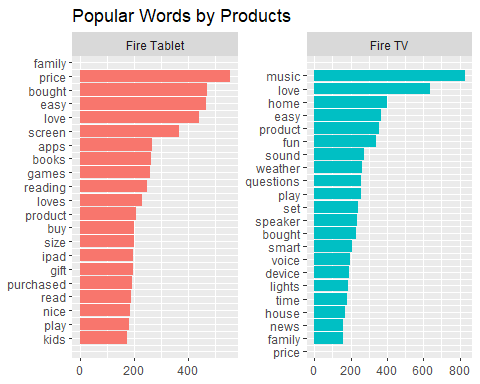
## Modify the dataset

## Tokenize and preprocess text

## Word Frequesncy

## Selecting by n

## Warning: `show.legend` must be a logical vector.



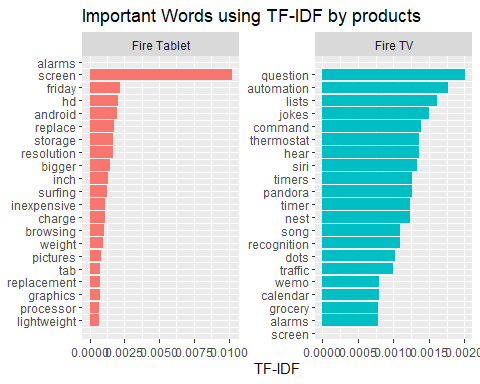
From the graphic, we can see that price is a concern for the Fire Tablet, but it is not in the top words for the Fire TV. For both devices, ease of use (“easy”) is a frequently mentioned word. Functionality is also discussed for both the Fire Tablet (“apps”, “books”, “games”, “reading”) and the Fire TV (“music”, “weather”, “speaker”, “news”).

## Tf-idf

## # A tibble: 6 x 6  
## name word n tf idf tf\_idf  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 Fire TV music 828 0.0320 0 0  
## 2 Fire TV love 636 0.0246 0 0  
## 3 Fire Tablet price 556 0.0224 0 0  
## 4 Fire Tablet bought 470 0.0189 0 0  
## 5 Fire Tablet easy 467 0.0188 0 0  
## 6 Fire Tablet love 440 0.0177 0 0

Based on this information, the most popular are frequently used in other reviews such that they have an idf of 0.

## Warning: `show.legend` must be a logical vector.



From this graphic, we can see the unique words for the Fire Tablet reviews pertain to issues with the product (“screen”, “replace”). The unique words for the Fire TV reviews include more niche uses for the product (“jokes”, “thermostat”, “timers”).

## Sentiment analysis

Implement sentiment analysis using the inner join function and different lexicons by performing an inner\_join() on the get\_sentiments() function.

# Get the lexicons  
bing<-get\_sentiments("bing")  
nrc<-get\_sentiments("nrc")  
afinn<-get\_sentiments("afinn")  
  
# Convert AFINN to negative and positive  
afinn\_neg\_pos <- afinn %>%  
 mutate( sentiment = ifelse( value >= 0, "positive",  
 ifelse( value < 0,  
 "negative", value)))  
afinn\_neg\_pos <-afinn\_neg\_pos %>%  
 select(word, sentiment)  
  
# Look at Match Ratios  
# Combine lexicons  
sentiments <- bind\_rows(list(bing=bing,nrc=nrc,afinn=afinn\_neg\_pos),.id = "lexicon")  
sentiments <- sentiments %>%   
 group\_by(lexicon) %>%  
 mutate(words\_in\_lexicon = n\_distinct(word)) %>%  
 ungroup()  
  
tidy %>%   
 mutate(words\_in\_reviews = n\_distinct(word)) %>%  
 inner\_join(sentiments) %>%  
 group\_by(lexicon,words\_in\_reviews, words\_in\_lexicon) %>%  
 summarise(lex\_match\_words = n\_distinct(word)) %>%  
 ungroup() %>%  
 mutate(total\_match\_words = sum(lex\_match\_words), #Not used but good to have  
 match\_ratio = lex\_match\_words / words\_in\_reviews) %>%  
 select(lexicon, lex\_match\_words, words\_in\_reviews, match\_ratio)

## Joining, by = "word"

## `summarise()` has grouped output by 'lexicon', 'words\_in\_reviews'. You can override using the `.groups` argument.

## # A tibble: 3 x 4  
## lexicon lex\_match\_words words\_in\_reviews match\_ratio  
## <chr> <int> <int> <dbl>  
## 1 afinn 511 5333 0.0958  
## 2 bing 842 5333 0.158   
## 3 nrc 973 5333 0.182

# NRC has the highest match ratio  
  
tidy\_bing <- tidy %>%   
 inner\_join(bing)

## Joining, by = "word"

tidy\_nrc <- tidy %>%   
 inner\_join(nrc)

## Joining, by = "word"

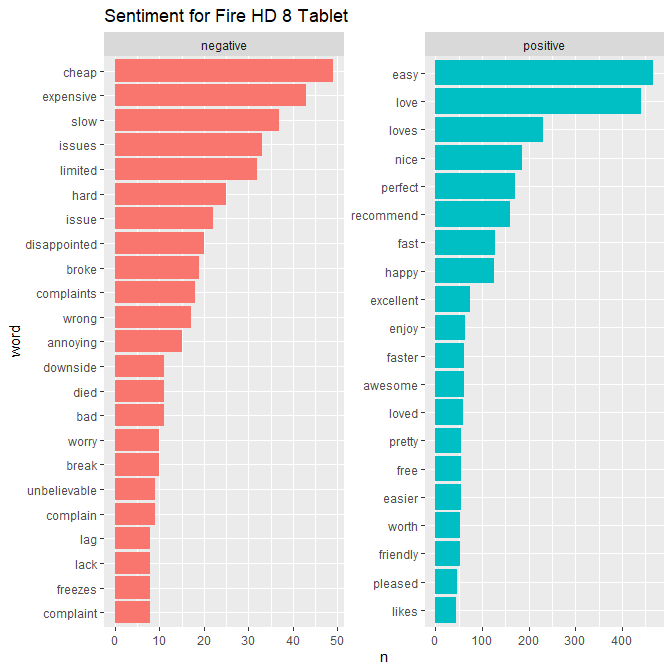
tidy\_afinn <- tidy %>%   
 inner\_join(afinn)

## Joining, by = "word"

## positive and negative words for each product

It’s important to understand which words specifically are driving sentiment scores, and since we are using tidy data principles, it’s not too difficult to check.

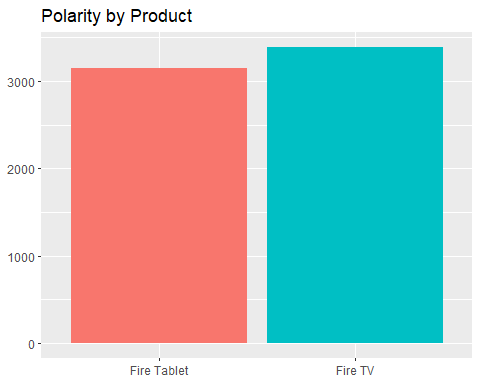
Product: Fire HD 8 Tablet



This graphic shows that the top negative words are contradictory descriptions of price (“cheap” & “expensive”). Other common negative words deal with issues with the product such as “slow”, “broke”, “lag”, and “freezes”. The common positive words include praise for the product (“love”, “nice”, “recommend”, “perfect”) and features (“fast”, “easy”).

## Polarity

We can break down the analysis using the Bing lexicon.



Looking at the polarity, both products have an overwhelmingly positive sentiment.

## Bigrams

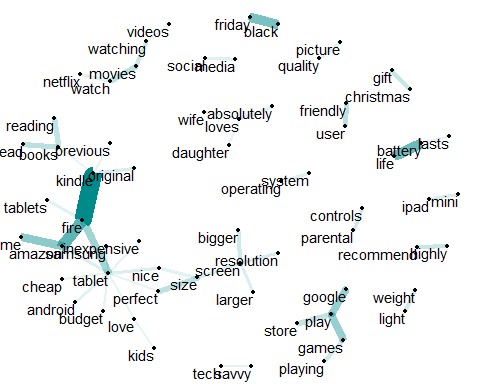
Product: Fire HD 8 Tablet

Count the most common bigrams.

## # A tibble: 4,119 x 3  
## word1 word2 n  
## <chr> <chr> <int>  
## 1 kindle fire 156  
## 2 amazon fire 97  
## 3 battery life 88  
## 4 black friday 79  
## 5 amazon prime 67  
## 6 google play 61  
## 7 play games 61  
## 8 fire tablet 54  
## 9 read books 39  
## 10 user friendly 37  
## # ... with 4,109 more rows

The top two bigrams reference the product (“kindle fire” & “amazon fire”). The other top bigrams include product features (“battery life”, “play games”, “read books”, & “user friendly”).

## IGRAPH dfd8b5a UN-- 64 51 --   
## + attr: name (v/c), n (e/n)  
## + edges from dfd8b5a (vertex names):  
## [1] kindle --fire amazon --fire battery --life   
## [4] black --friday amazon --prime google --play   
## [7] play --games fire --tablet read --books   
## [10] user --friendly watch --movies perfect --size   
## [13] play --store reading --books christmas--gift   
## [16] highly --recommend watching --movies playing --games   
## [19] bigger --screen parental --controls light --weight   
## [22] screen --resolution kindle --original ipad --mini   
## + ... omitted several edges



Based on this graphic, we can see the strong links around the product name (“kindle” to “fire”). This chain for the product name also links to the features of the tablet (“size”, “screen”, “resolution”) through positive qualifiers (“nice”, “perfect”). We can also see links around different activities like games (“google”, “play”), reading (“read”, “books”), and movies (“netflix”, “watch”, “videos”).

## Topic modeling

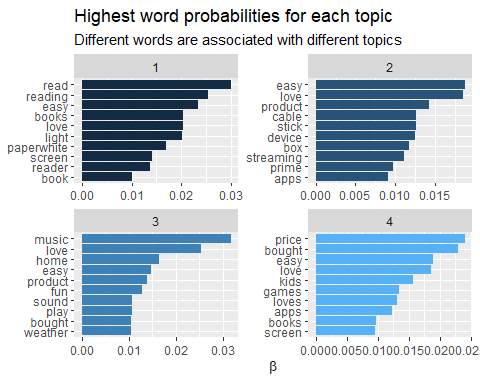
## Joining, by = "word"

## Warning: package 'topicmodels' was built under R version 4.1.3

## A LDA\_VEM topic model with 4 topics.

## Beta Matrix

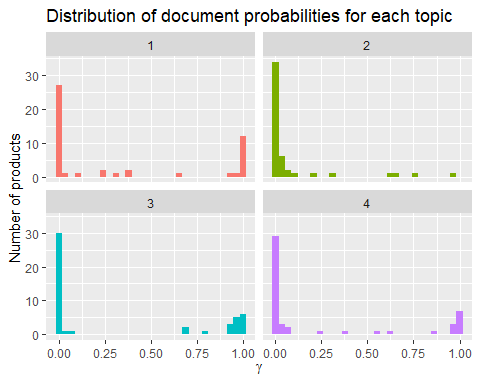
## Selecting by beta



Based on the graphic, the topics are grouped around reading (topic 1), TV (topic 2), kids (topic 4). Topic 3 is not as clearly defined with words like “music”, “home”, “sound”, and “weather”. Topic 3 could refer to a multi-functional product.

## Gamma Matrix

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



Topics 1,3, and 4 have products that are clearly associated with them. However, topic 2 only has a few products that have a high probability of being associated with it.

## Additional Question

What is the difference in sentiment between the best rated product (highest average rating) and the least recommended product (lowest average rating)?

# Determine products with the highest # of "Yes" recommendations, limit to products that have at least 100 reviews to limit impact of a single rating  
amazon %>%   
 group\_by(name) %>%  
 summarise(rating = mean(reviews.rating, na.rm = TRUE), count = n()) %>%   
 filter(count > 100) %>%   
 arrange(desc(rating))

## # A tibble: 16 x 3  
## name rating count  
## <chr> <dbl> <int>  
## 1 "Amazon Fire Hd 10 Tablet, Wi-Fi, 16 Gb, Special Offers - Silve~ 4.77 128  
## 2 "Amazon Kindle Paperwhite - eBook reader - 4 GB - 6 monochrome ~ 4.76 3176  
## 3 "Kindle Voyage E-reader, 6 High-Resolution Display (300 ppi) wi~ 4.74 580  
## 4 "Amazon Fire Hd 8 8in Tablet 16gb Black B018szt3bk 6th Gen (201~ 4.74 135  
## 5 "Amazon - Amazon Tap Portable Bluetooth and Wi-Fi Speaker - Bla~ 4.73 318  
## 6 <NA> 4.66 6760  
## 7 "Amazon Fire Tv,,,\r\nAmazon Fire Tv,,," 4.65 2527  
## 8 "Echo (White),,,\r\nEcho (White),,," 4.65 3309  
## 9 "All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Include~ 4.59 2814  
## 10 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers,~ 4.56 372  
## 11 "Brand New Amazon Kindle Fire 16gb 7 Ips Display Tablet Wifi 16~ 4.53 1038  
## 12 "Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-P~ 4.51 1685  
## 13 "Amazon 5W USB Official OEM Charger and Power Adapter for Fire ~ 4.46 208  
## 14 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers,~ 4.45 10966  
## 15 "All-New Kindle E-reader - Black, 6 Glare-Free Touchscreen Disp~ 4.43 212  
## 16 "All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 32 GB - Include~ 4.30 158

The Amazon Fire Hd 10 Tablet, Wi-Fi, 16 Gb, Special Offers - Silver Aluminum has the highest average rating, and the All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 32 GB - Includes Special Offers, Magenta has the lowest average rating.

## Pre-process & Filter Data

# Pre-process the data and filter to relevant 2 products  
name <- amazon$name  
products<-split(amazon,name)  
  
product1<-products[[2]]  
product2<-products[[11]]  
  
products <- bind\_rows(product1%>%   
 mutate(name = "All-New Fire HD 8 Tablet"),  
 product2%>%   
 mutate(name = "Fire HD 10 Tablet")  
 )   
  
# Remove contractions  
products$text <- sapply(products $text, fix.contractions)  
  
# Remove stop words, undesirable words, and words with 2 or fewer characters  
tidy\_ratings <-products %>%  
 unnest\_tokens("word",text)%>%  
 anti\_join(stop\_words)%>%  
 filter (!word %in% undesirable\_words) %>%  
 filter(nchar(word) > 2)

## Warning: Outer names are only allowed for unnamed scalar atomic inputs

## Joining, by = "word"

tidy\_ratings$word <- gsub("\\s+","", tidy\_ratings$word)  
tidy\_ratings$word <- gsub("[^a-zA-Z]","", tidy\_ratings$word)

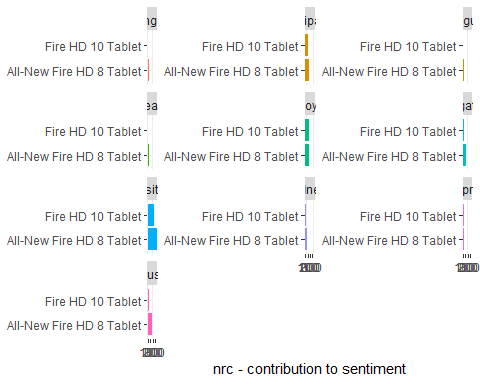
## Sentiment Analysis and Comparison

# Want to examine sentiments associated with each product, so we will use the NRC lexicon  
  
tidy\_ratings\_nrc <- tidy\_ratings %>%   
 inner\_join(get\_sentiments("nrc"))

## Joining, by = "word"

tidy\_ratings\_nrc %>%   
 count(sentiment,name,sort = TRUE)%>%   
 group\_by(sentiment)%>%  
 top\_n(5)%>%  
 ungroup()%>%  
 #mutate(word=reorder (word,n))%>%  
 ggplot(aes(name,n,fill=sentiment))+  
 geom\_col(show.legend = FALSE)+  
 facet\_wrap(~sentiment,ncol=3,scales = "free\_y")+  
 labs(y="nrc - contribution to sentiment",x=NULL)+  
 coord\_flip()

## Selecting by n



Based on the sentiment comparison between the two products, we can see that the Fire HD 10 Tablet (the higher rated product) had much fewer words with negative sentiments (anger, disgust, fear) when compared to the lower rated All-New Fire HD 8 Tablet. However, the All-New Fire HD 8 Tablet has more sentiment words than the Fire HD 10 Tablet across all sentiments. The differences in between the two products is evident in the gap in the number of words with each sentiment. The differences in the two products can be seen in the absolute difference between the number of sentiment words. The difference in negative words is much greater when compared to the difference in positive words.