Sentiment Analysis

Zander Bonnet

3/12/2025

Three hotels want to know what their customers like and dislike about their hotels, so that they can make marketing and renovation decisions based on their hotel reviews.

```
library(syuzhet)
berk <- read.csv('/Users/zanderbonnet/Desktop/GCU/DSC-570/Topic 3/DSC-570-RS-T3-HotelDatasets/DSC-570-R
bnb <- read.csv('/Users/zanderbonnet/Desktop/GCU/DSC-570/Topic 3/DSC-570-RS-T3-HotelDatasets/DSC-570-R-
inn <- read.csv('/Users/zanderbonnet/Desktop/GCU/DSC-570/Topic 3/DSC-570-RS-T3-HotelDatasets/DSC-570-R-
```

To accomplish this we need to split the reviews into sentences so that we can analyze the sentiment of the sentences within the reviews.

```
berk_sen <- get_sentences(berk$Reviews)
bnb_sen <- get_sentences(bnb$Reviews)
inn_sen <- get_sentences(inn$Reviews)
print(inn_sen[1:10])
```

```
##  [1] "Relaxing."
##  [2] "Nice and peaceful."
##  [3] "Room could use upgrades."
##  [4] "It is not walking distance to beach."
##  [5] "better to pay the extra to stay closer to beach and restaurants."
##  [6] "The hotel was great for our stay."
##  [7] "It was a couples trip for beach week."
##  [8] "The only things I didn't like was the cleanliness of the bathroom."
##  [9] "And as a women that's a big part of staying in the hotel"
## [10] "Not a bad stay , hotel could use a remodel."
```

Now that we have split the sentences, we can use the get_sentiment() function to get the sentence's sentiment.

```
(sent_berk <- get_sentiment(berk_sen))[1:10]
```

```
##  [1] 0.25 1.80 2.25 0.25 0.00 0.75 1.25 1.25 0.50 1.55
```

```
trim_sent_berk <- sent_berk[(sent_berk > 0) | (sent_berk < 0)]
(sent_bnb <- get_sentiment(bnb_sen))[1:10]
```

```
##  [1]  2.40  0.50  0.85  1.40  2.30 -0.75  0.80  0.50  0.30  1.60
```

```
trim_sent_bnb <- sent_bnb[(sent_bnb > 0) | (sent_bnb < 0)]
(sent_inn <- get_sentiment(inn_sen))[1:10]
```

```
##  [1]  0.00  1.25  0.00  0.00  1.10  0.50  0.00  1.00  0.25 -0.65
```

```
trim_sent_inn <- sent_inn[(sent_inn > 0) | (sent_inn < 0)]
```

From this we can see the most positive and negative within the first 10 reviews for each hotel.

```
## [1] "Berkley"
```

```
## [1] "Excellent customer service at the front desk, very polite and helpful, answered all our questio
## [2] "More"
```

```
## [1] "Bed&Breakfast"
```

```
## [1] "Breakfast was the best, and the best thing about the stay was the one and only host, great conv
## [2] "Sharing the bathroom, never a problem, and coming in and out with you wanted never a problem."
```

```
## [1] "Inn"
```

```
## [1] "Nice and peaceful."
## [2] "Not a bad stay , hotel could use a remodel."
```

Looking at the most positive and negative sentences from the first 10 sentences for each hotel we can see that the model is using traditionally positive words to determine that that the sentence is positive vs. negative. It do this by determining if a a world is positive (+1), negative (-1),or neutral (0) and then calculating the ratio of positive to negative values. This means that a sentence with a positive sentiment value would be considered a positive review and if it is a negative value it would have a negative sentiment. There is also the chance that the model results in a value of 0. This would mean that the sentence could be considered neutral. This could be a result of a review containing only neutral words, or if there is an equal amount of positive and negative words in the sentence. In this analysis I will be looking at neutral reviews because we want to see the true average review of the hotel. If I were to eliminate neutral reviews the extreme reviews would be able to skew the model, as it would eliminate a lot of our data.

We can now look at the measures of the central tendency of the three hotels to gain insight into their overall reviews.

```
summary(sent_berk)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.0000  0.0000  0.2500  0.4689  0.8500  4.1500
```

The Berkley Hostel has a mean of .47, so it has positive reviews on average, but the median is .25 so there is a right skew to the data. This means that there are potentially outliers on the positive side dragging up the mean. We can also see that since Q1 is 0 we can be sure that at least 75% of the data has a vlaue of 0 or higher. So most reviews have a positive or neutral sentiment for this hotel.

```
summary(sent_bnb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.2500  0.0000  0.5000  0.6461  1.2500  5.2000
```

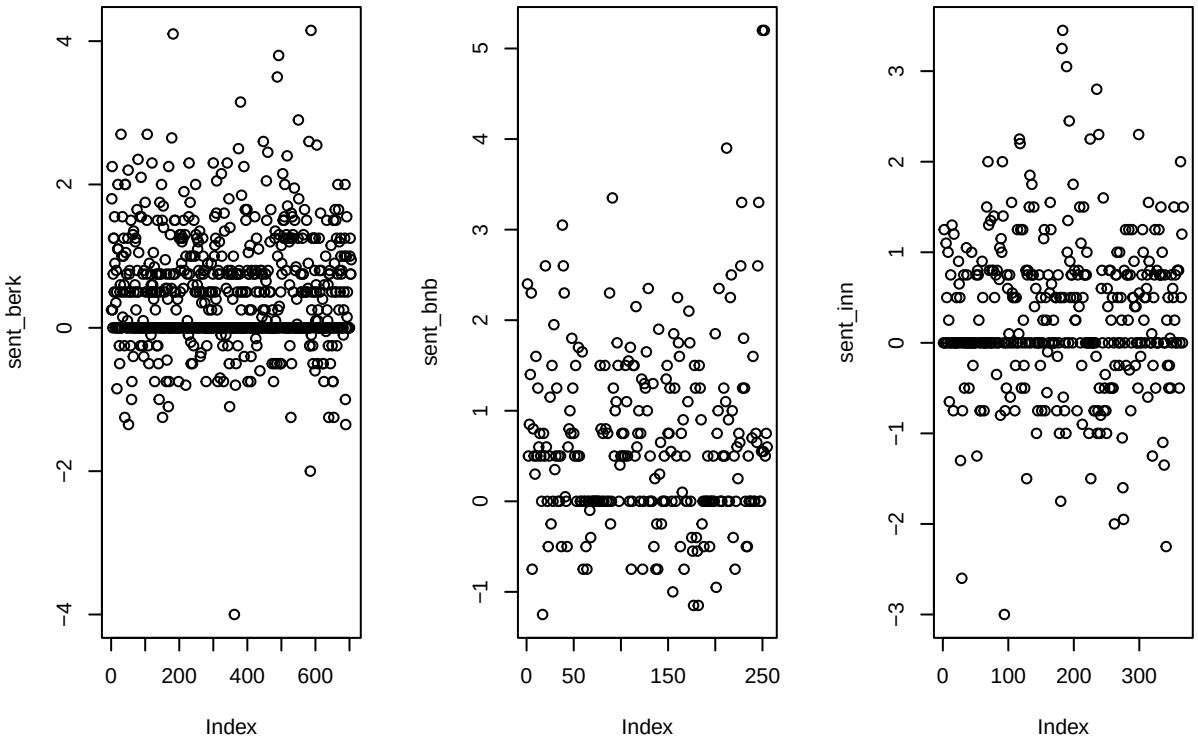
The bed & breakfast has similar features to the data as the Berkley Hostel. The spread between the mean and median is smaller hinting at there being less skew within the data. This model appears to have an extereme outlier with the max value being 5.2. If we were to use the IQR outlier method we would get an outlier threshold of (-1.235,2.515) so 5.2 would be much larger than this threshold. This hotel does have the most positive mean and median as well.

```
summary(sent_inn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.0000  0.0000  0.1000  0.2888  0.7500  3.4500
```

The Ambassadors has the smallest mean and median showing that it might have the worst reviews of the bunch. It has similar features to the rest as it has the right skew and at least 75% of the data is nuetral or better.

```
par(mfrow = c(1,3))
plot(sent_berk)
plot(sent_bnb)
plot(sent_inn)
```

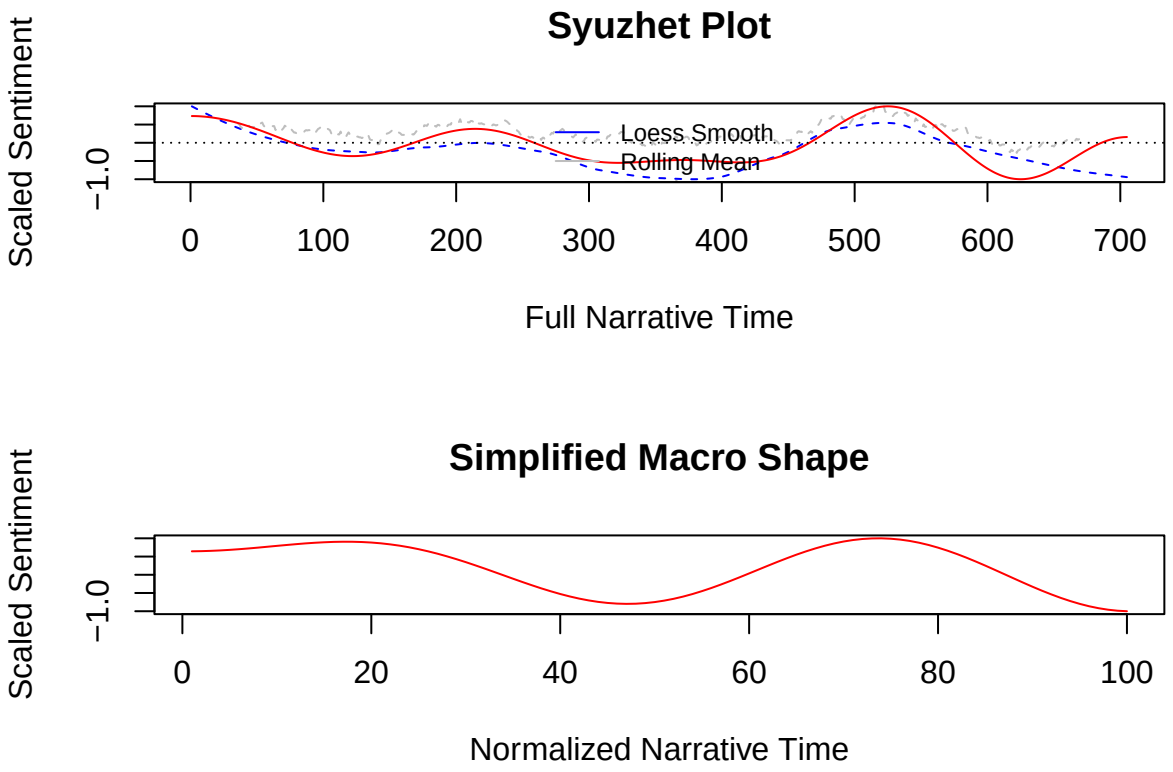


We can use these plots to visualize the spread of the data that we talked about with the measures of central

tendency. We see that the Berkeley hostel has a majority of its data 0 or higher and has a small proportion of less than 0 values. On the other hand the Ambassador Inn has a more evenly spread review distribution. This shows that they have a larger percentage of negative reviews.

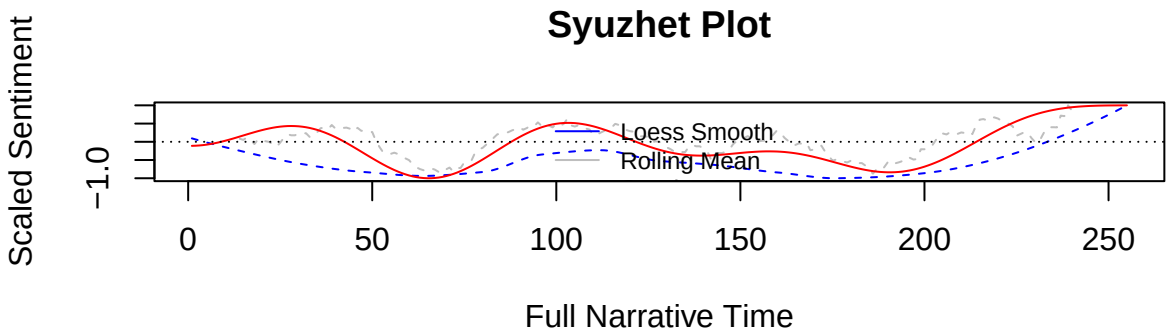
In order to gain insight of of the average reviews over time we can normalize the sentiment values and look at the trend over time. Furthermore, we can standardizes the range of time from 0 to 100, so that we can compare all of the hotels across the same time period. Even though they all have different lengths of sentiments.

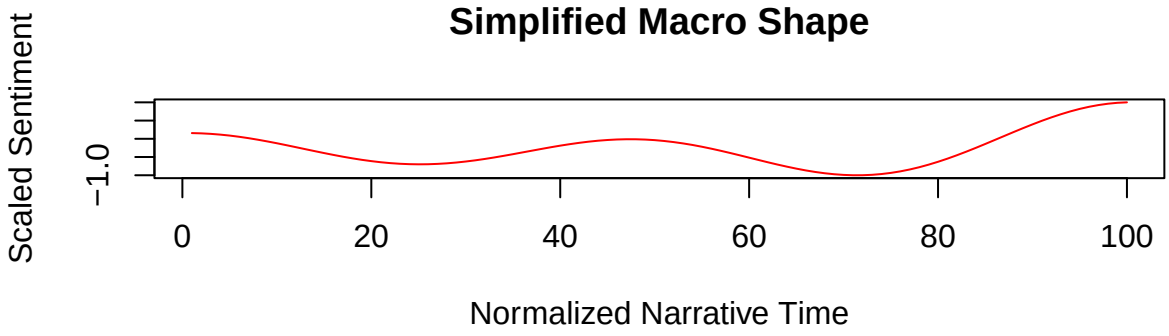
```
simple_plot(sent_berk)
```



In this hotel we can see that the reviews started very good in our sample, but the sentiments appear to be seasonal.

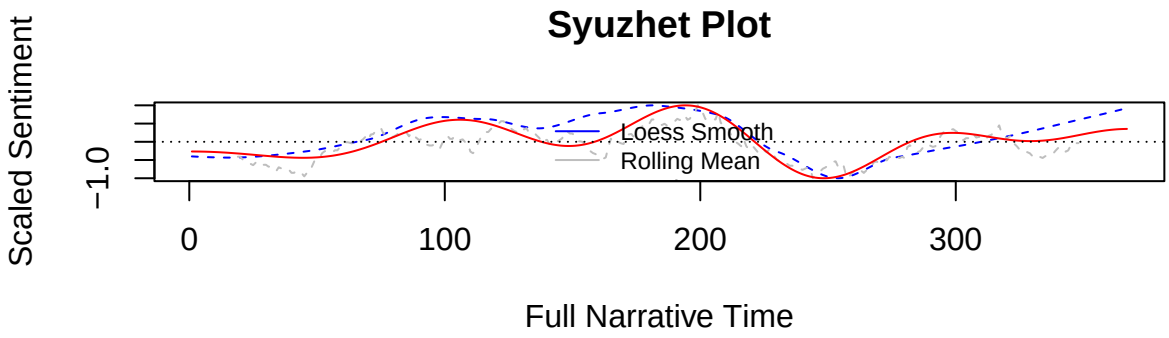
```
simple_plot(sent_bnb)
```

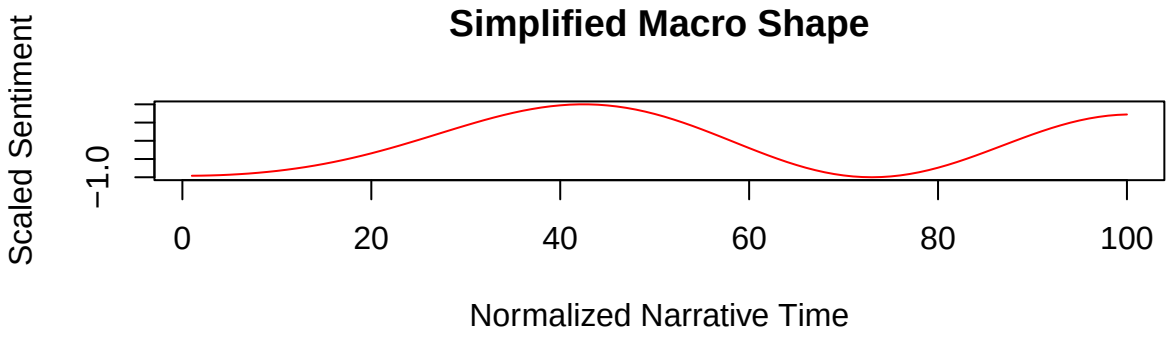




This hotel has had an improvement in its sentiment over the last portion of the data. It started out as fairly nuetral reviews, but somthing in the end has caused its sentiments to spike.

```
simple_plot(sent_inn)
```





This hotel had very negative sentiments to start the data, but was able to recover to roughly neutral average sentiment in the more recent reviews.

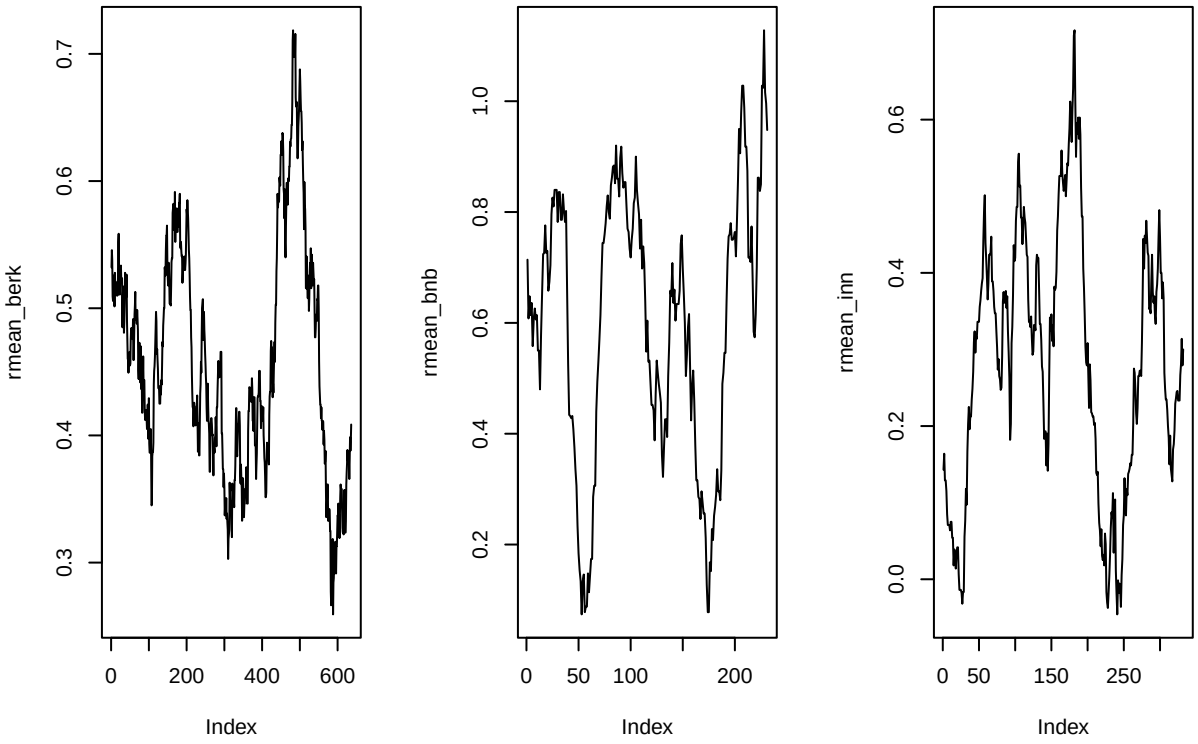
To get more information we want to be able to look at these plots side by side. To accomplish this we must first calculate the rolling mean of the sentiments to get the average sentiment at a given point.

```
library(zoo)
```

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

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

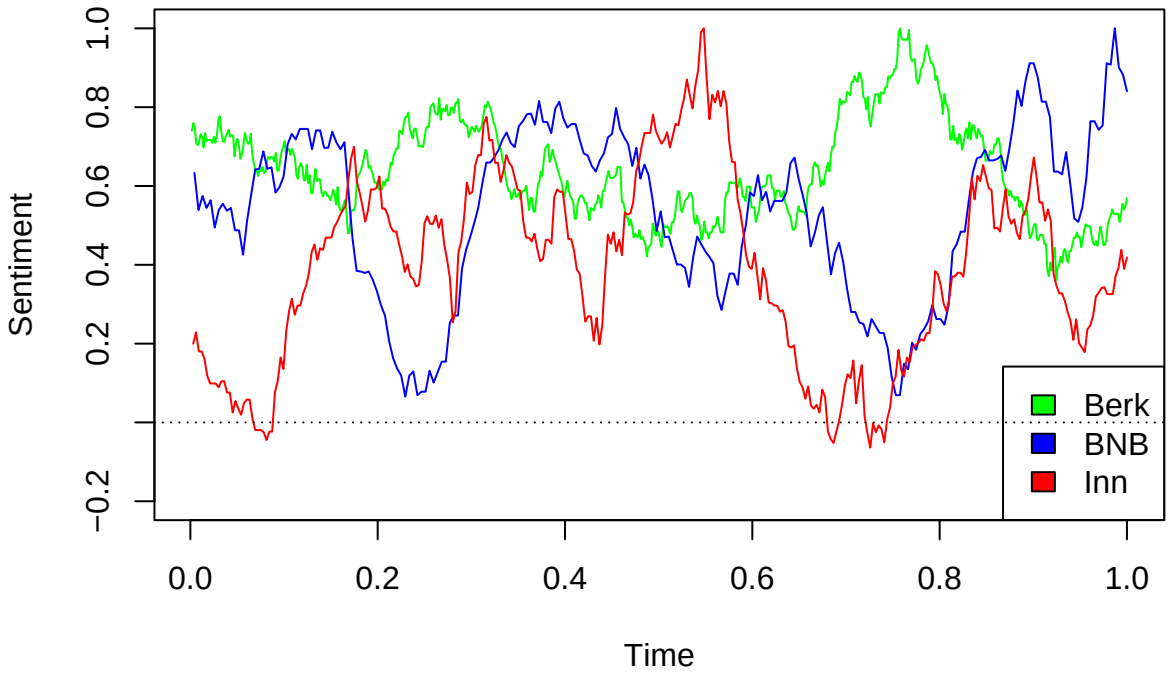
```
rmean_berk <- rollmean(sent_berk,length(sent_berk)*.1)
rmean_bnb <- rollmean(sent_bnb, length(sent_bnb)*.1)
rmean_inn <- rollmean(sent_inn, length(sent_inn)*.1)

par(mfrow = c(1,3))
plot(rmean_berk, type ='l')
plot(rmean_bnb, type ='l')
plot(rmean_inn, type ='l')
```



To keep the windows similar I chose to use a window of 10% of the total length of the input data. This will allow for a curve that shows some detail while also being smooth enough to interpret. The is still the problem that the input data does not share the same length so we cannot put them on the same plot. We can then use the resale function to scale the x values so we can plot all three plots on the same graph.

```
scale_berk <- rescale_x_2(rmean_berk)
scale_bnb <- rescale_x_2(rmean_bnb)
scale_inn <- rescale_x_2(rmean_inn)
```

```
plot(scale_berk, type = 'l', col = 'green', ylim = c(-.2,1), xlab = 'Time', ylab = 'Sentiment')
lines(scale_bnb, type = 'l', col = 'blue')
lines(scale_inn, type ='l', col = 'red')
abline(h = 0, lty = 3, col = 'black')
legend(x = 'bottomright', legend = c('Berk', 'BNB', 'Inn'),
       fill = c('green', 'blue', 'red'))
```



From this plot we can see the rolling averages of the three hotels over the same period of time. We can now preform some comparative analysis on the hotels. For instance we can see the the BNB has had a surge in sentiment within reviews while the Berkeley has seen its average sentiment slip. We can also see that the Ambassador tends to have the lowest reviews of the bunch, but did have a large review spike in about the middle of the data set. This could lead us to make decisions like possibly renovating the Ambassador Inn or the Berkeley hotel, while upping the marketing budget on the BNB as it is currently doing very well.

To gather more information on the more general trend of the data we can utilize discrete cosine transformation. This method will eliminate noise and will weight the most important features higher than others. This will give us a smooth curve that represents the average sentiments at that time.

```
dct_berk <- rescale_x_2(get_dct_transform(sent_berk))
dct_bnb <- rescale_x_2(get_dct_transform(sent_bnb))
dct_inn <- rescale_x_2(get_dct_transform(sent_inn))
```

Check to ensure that all the hotel data sets are the same length of 100.

```
length(dct_berk$y)
```

```
## [1] 100
```

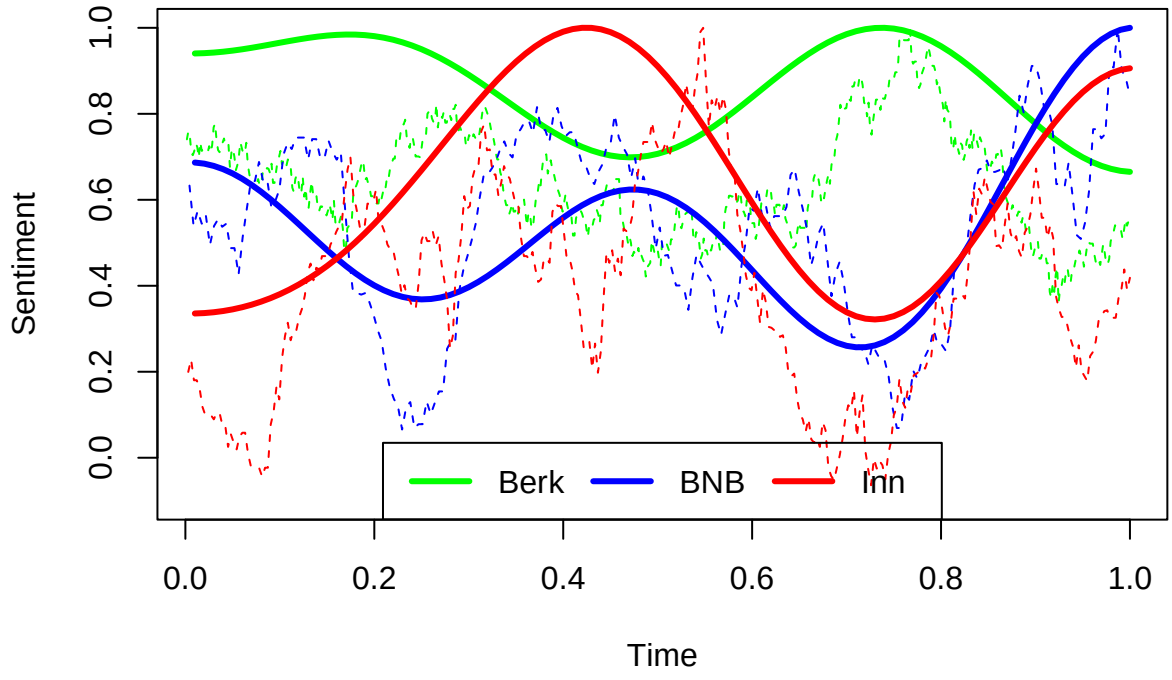
```
length(dct_bnb$y)
```

```
## [1] 100
```

```
length(dct_inn$y)
```

```
## [1] 100
```

```
plot(dct_berk, type = 'l',
     ylim = c(-.1,1)
     , col = 'green',
     ylab = 'Sentiment',
     xlab = 'Time',
     lwd = 3)
lines(scale_berk, type = 'l', col = 'green', lty = 2, lwd = 1)

lines(dct_bnb, type = 'l', col = 'blue', lwd = 3)
lines(scale_bnb, type = 'l', col = 'blue', lty = 2, lwd = 1)

lines(dct_inn, type = 'l', col = 'red', lwd =3)
lines(scale_inn, type = 'l', col = 'red', lty = 2, lwd = 1)

legend(x = 'bottom', legend = c('Berk', 'BNB', 'Inn'),
       col = c('green', 'blue', 'red'), ncol = 3,
       lty = 1, lwd = 3)
```



From this plot we can see that the DCT transformations show us a line similar to the simplified macro shape from the simple_plot() function. It shows the overall trend of the data, and is less impacted by the neutral inputs, because DCT would give them a smaller weight. Looking at this we can see that the Ambassador Inn and the BNB are leading the way with reviews and the Berkeley is slipping.

```r
paste('Berk and BNB Cor:',round(cor(dct_berk$y,dct_bnb$y),3))
```

```
## [1] "Berk and BNB Cor: -0.696"
```

```r
paste('Berk and Inn Cor:',round(cor(dct_berk$y,dct_inn$y),3))
```

```
## [1] "Berk and Inn Cor: -0.876"
```

```r
paste('BNB and Inn Cor:',round(cor(dct_bnb$y,dct_inn$y),3))
```

```
## [1] "BNB and Inn Cor: 0.453"
```

When we look at the correlation between the hotel sentiments we can see if there is any link between the hotels. There is a strong correlation between the Berkeley and the Ambassador Inn, but it is negative. This would mean that these two hotels have an inverse relationship between them. So when one of the hotel's ratings go up the others tend to go down. This is a similar case with the Berkeley and the BNB, but it is a smaller correlation. The BNB and Ambassador have a mild correlation of .45, but that is most likely due to their steep rise in sentiment over the last portion of the data set.

If I were managing these hotels I would look at this data and have a couple of ideas. First The Berkeley is seeing a slow decline in sentiment, so it might be time for a small updating of the hotel. It still performs very well, but as an owner, you want to preserve that success. For the BNB it has seen a very sharp rise in sentiment. We need to find out what caused that and how we can improve on that. Once we find out what caused this spike the owner should focus marketing heavily on this aspect. The Ambassador is seeing an increase in sentiment in reviews, but it still has been one of the lower-rated hotels in the last part of the data. It might be time for an update. Another thing we need to focus on is how all of these hotels are seasonal. Some perform better than others at different times within the data. We need to identify when these hotels perform best and focus marketing on that. Like if the Inn is seeing its best reviews in the summer and worst in the winter, and we know the Berkeley has the inverse from its negative correlation we should gear our marketing on that aspect as well.

There are a handful of ethical outcomes that we must be aware of in this analysis. We are basing an entire analysis on written hotel reviews and the computer is assigning the values of the sentiment based on pure verbiage. There are a couple of problems with this. One, written reviews tend to be skewed in the extreme direction as people tend to only leave reviews if they had a great time or a terrible time. This is a potential bias within the data set. For example, there might have been people who had an okay time at the hotel with a minor inconvenience, but they decided to not leave a review because that takes time and it wasn't that big of a deal. This might cause the reviews to be more skewed in the positive direction. Or the opposite could happen. There is no real way of telling without getting a review from every single person that stays there. The other issue is the computer does not consider the context when analyzing the text. Someone could have been sarcastic or used slang in the review, and this would have caused the computer to incorrectly identify the sentiment. These potential biases could create an error in the analysis that could cause someone to lose their job, hotels being closed, budgets being cut, etc. So we must ensure that we preface this with our analysis and do our best to mitigate any potential bais form being introduced. This is why we need to properly vet our data and ensure the reviews are a representative sample of the guests that stay at these hotels.