Independent Study

Elisa Cangialosi

4/28/2019

In this report I perform text analysis of the publicly available review data posted on Booking for the major turistic italian cities: Milan, Rome, Florence Venice and Verona.

Below are some of the key findings.

The dataset comprises 13 columns, which are the following:

```
colnames(en_reviewResponse)
```

```
##  [1] "X"                "hotelid"          "reviewer_name"
##  [4] "reviewer_country" "review_title"     "review_date"
##  [7] "stay_date"        "review_score"     "fullText"
## [10] "response"         "helpfulvote"      "city"
## [13] "reviewID"
```

There are 87633 english reviews in the dataset.

Update stay date to a date format

```
toremoveStay = c("Stayed in ")

gsub(paste0(toremoveStay,collapse = "|"),"", en_reviewResponse$stay_date)
en_reviewResponse$stay_date <- gsub(paste0(toremoveStay,collapse = "|"),"", en_reviewResponse$stay_date)

library(lubridate)
en_reviewResponse$stay_date <- as.Date(paste('01', en_reviewResponse$stay_date), format='%d %b %Y')
```

Update review date to a date format

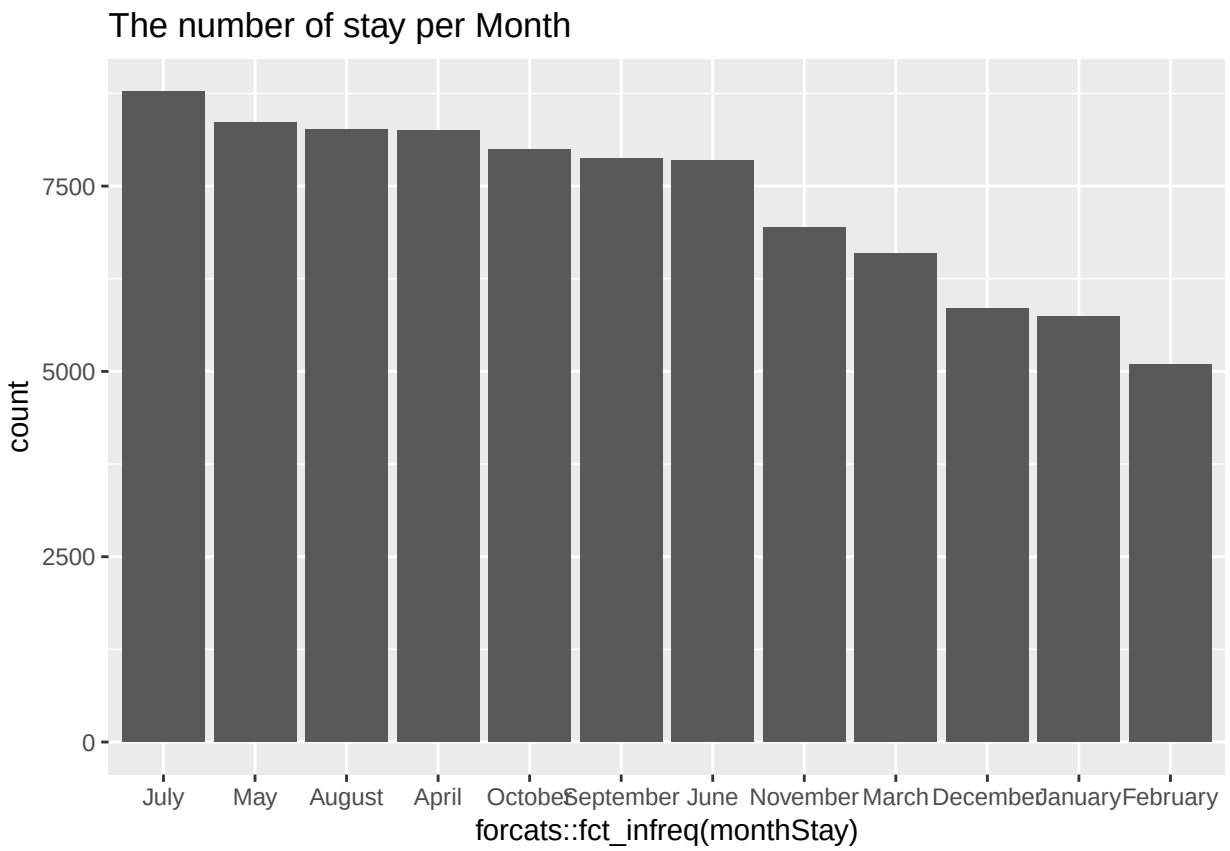
```
toremove = c("Reviewed: ")

gsub(paste0(toremove,collapse = "|"),"", en_reviewResponse$review_date)
en_reviewResponse$review_date <- gsub(paste0(toremove,collapse = "|"),"", en_reviewResponse$review_date)

en_reviewResponse$review_date <- AsDate(en_reviewResponse$review_date)
```

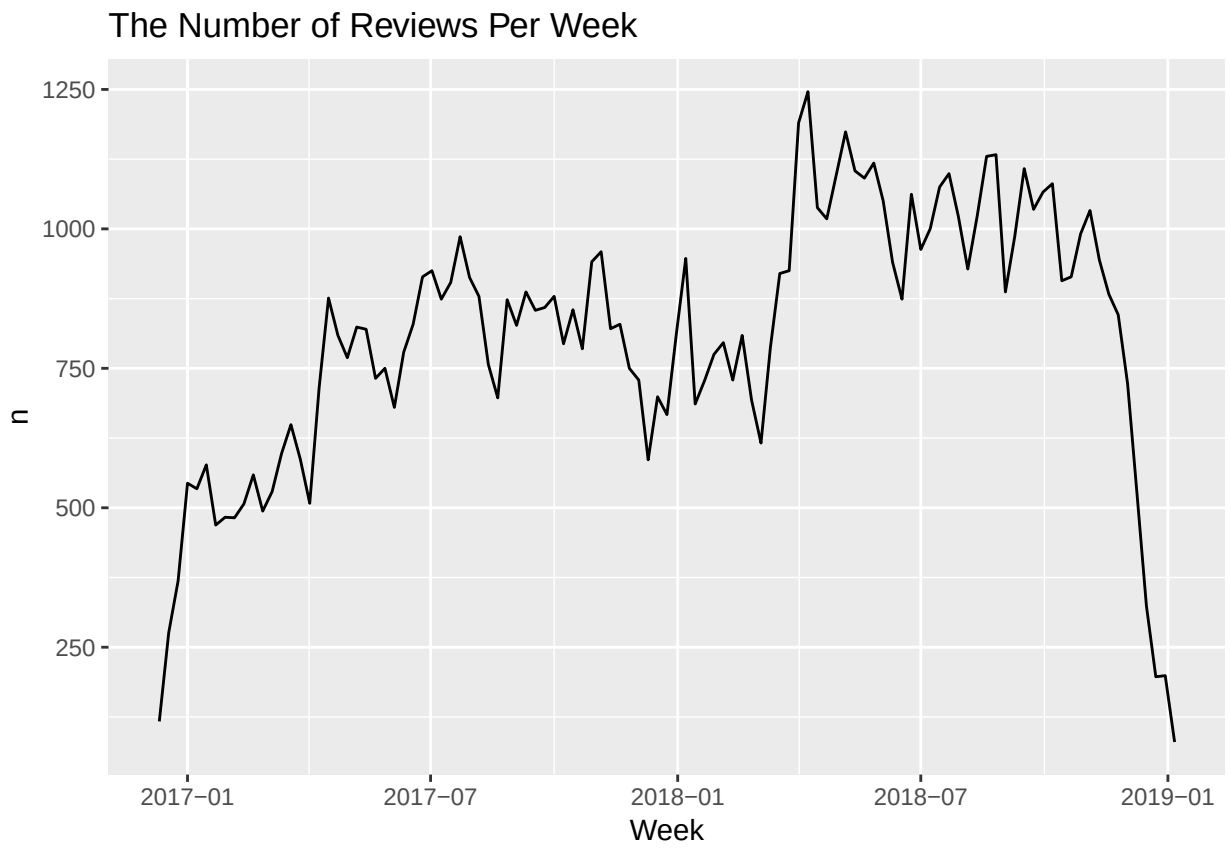
```
en_reviewResponse$monthStay <- as.character(format(en_reviewResponse$stay_date,'%B'))

en_reviewResponse[!is.na(en_reviewResponse$monthStay),] %>%
  ggplot(aes(x = forcats::fct_infreq(monthStay))) +
  geom_bar() +
  ggtitle('The number of stay per Month')
```



The highest number of weekly reviews was received within the half of 2018. The hotels received almost 1250 reviews in that week.

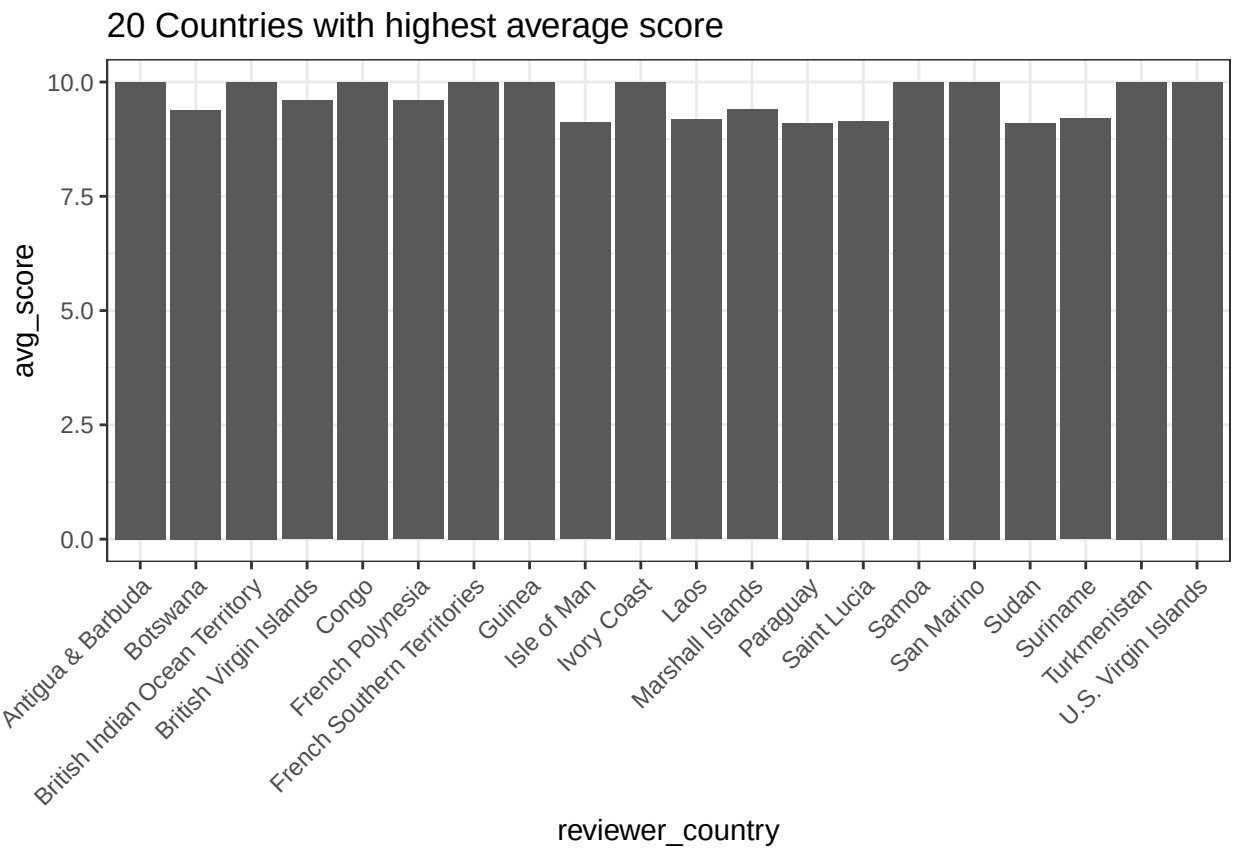
```
en_reviewResponse %>%
  count(Week = round_date(review_date, "week")) %>%
  ggplot(aes(Week, n)) +
  geom_line() +
  ggtitle('The Number of Reviews Per Week')
```



20 Countries with highest average score

```
avgscore_nation <- sqldf('SELECT reviewer_country, avg(review_score) as avg_score from en_reviewRespons

ggplot(avgscore_nation[1:20,],aes(x=reviewer_country, y=avg_score)) + geom_bar(stat = 'identity')+theme_
```

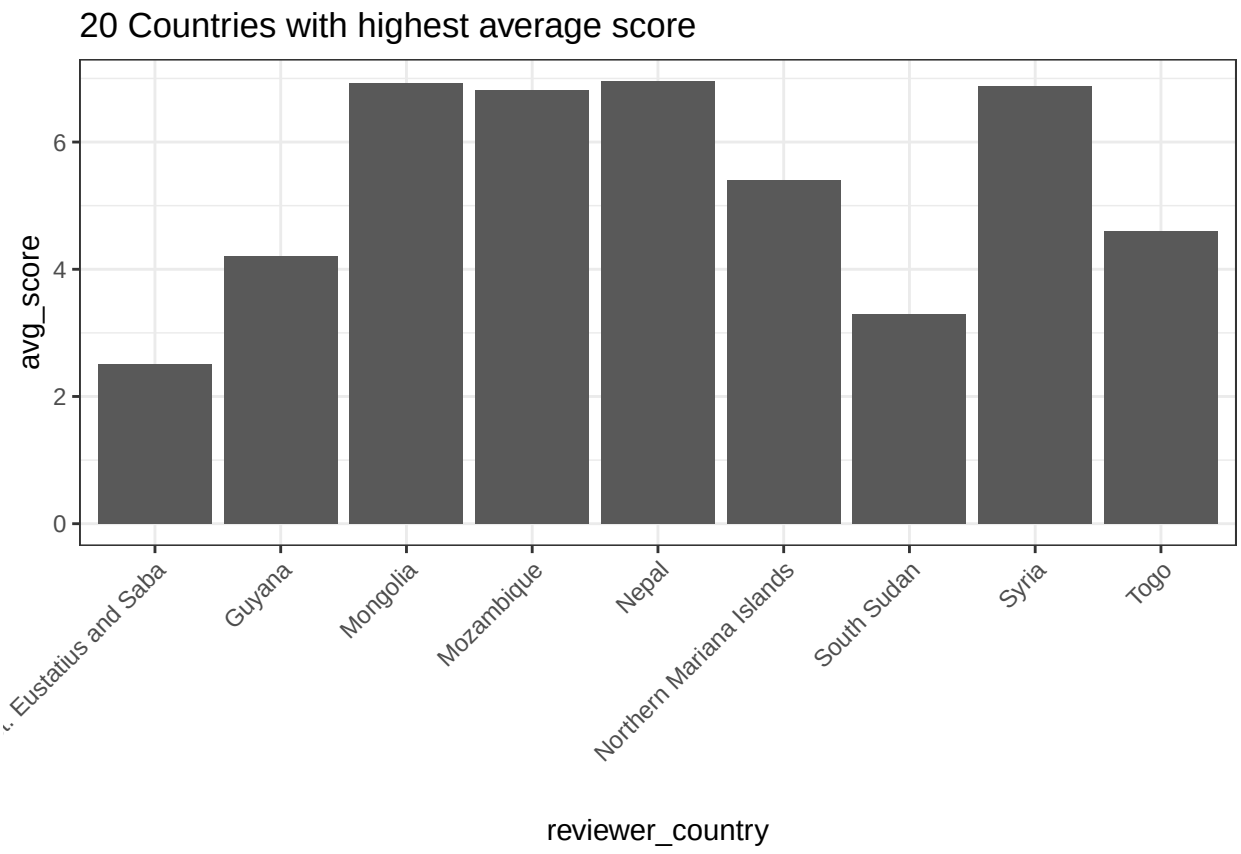


Italian average score nation

```
it_avgScore <- avgscore_nation %>% filter(reviewer_country == 'Italy')
```

Countries with lowest average score

```
ggplot(avgscore_nation[183:191,],aes(x=reviewer_country, y=avg_score)) + geom_bar(stat = 'identity')+th
```



We often want to understand the relationship between words in a review. What sequences of words are common across review text? Given a sequence of words, what word is most likely to follow? What words have the strongest relationship with each other? Therefore, many interesting text analysis are based on the relationships. When we exam pairs of two consecutive words, it is called "bigrams".

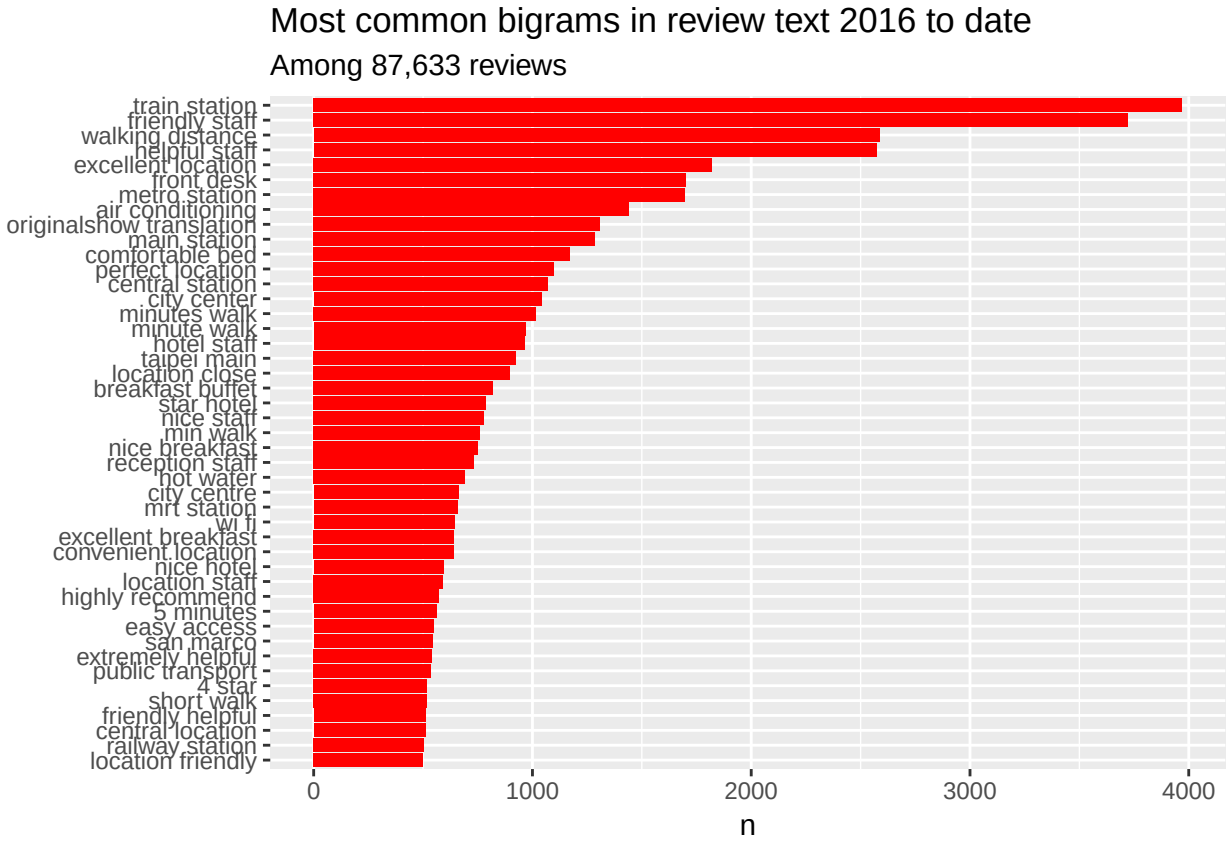
What are the most common bigrams in our reviews?

```
review_bigrams <- en_reviewResponse %>%
  unnest_tokens(bigram, fullText, token = "ngrams", n = 2)

bigrams_separated <- review_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")
bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)
bigrams_united <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")
bigrams_united %>%
  count(bigram, sort = TRUE)
```

```
bigrams_united %>%
  count(bigram, sort = TRUE) %>%
  filter(n > 500) %>%
  mutate(bigram = reorder(bigram, n)) %>%
  ggplot(aes(bigram, n)) +
  geom_col(fill = "red") +
  xlab(NULL) +
  coord_flip() +
  labs(title = "Most common bigrams in review text 2016 to date",
       subtitle = "Among 87,633 reviews")
```



The above visualizes the common bigrams in Booking reviews, showing those that occurred at least 3000 times and where neither word was a stop-word.

The network graph shows strong connections between the top several words ("friendly", "staff", "excellent" and "location", "train" and "station"). However, we do not see clear clustering structure in the network.

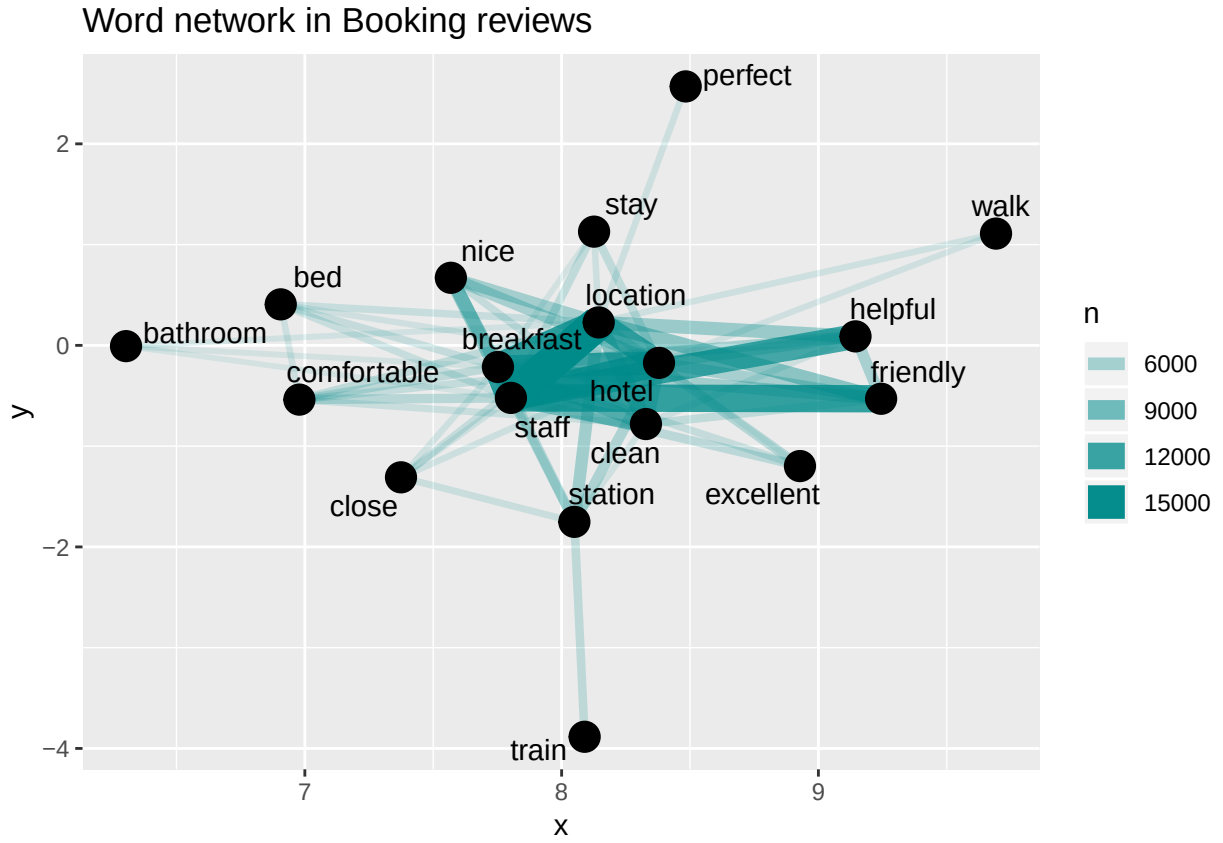
Visualize bigrams in word networks:

```
review_subject <- en_reviewResponse %>%
  unnest_tokens(word, fullText, token = "ngrams", n = 1) %>%
```

```
  anti_join(stop_words)

my_stopwords <- data_frame(word = c(as.character(1:10)))
review_subject <- review_subject %>%
  anti_join(my_stopwords)

title_word_pairs <- review_subject %>%
  pairwise_count(word, reviewID, sort = TRUE, upper = FALSE)
set.seed(1234)
title_word_pairs %>%
  filter(n >= 3000) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "cyan4") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  ggtitle('Word network in Booking reviews')
```



The importance of words can be illustrated in a wordcloud.

The word cloud clearly shows that “room”, “staff”, “hotel”, “breakfast” and “location” are the five most important words in Booking reviews in italian tourist cities.

```
freqWords <- review_subject %>% count(word, sort = TRUE)
set.seed(1234)
wordcloud(words = freqWords$word, freq = freqWords$n, min.freq = 3000,
          max.words=400, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```