Text Analysis of Hotel Reviews from Booking.com

Elisa Cangialosi

April 29, 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

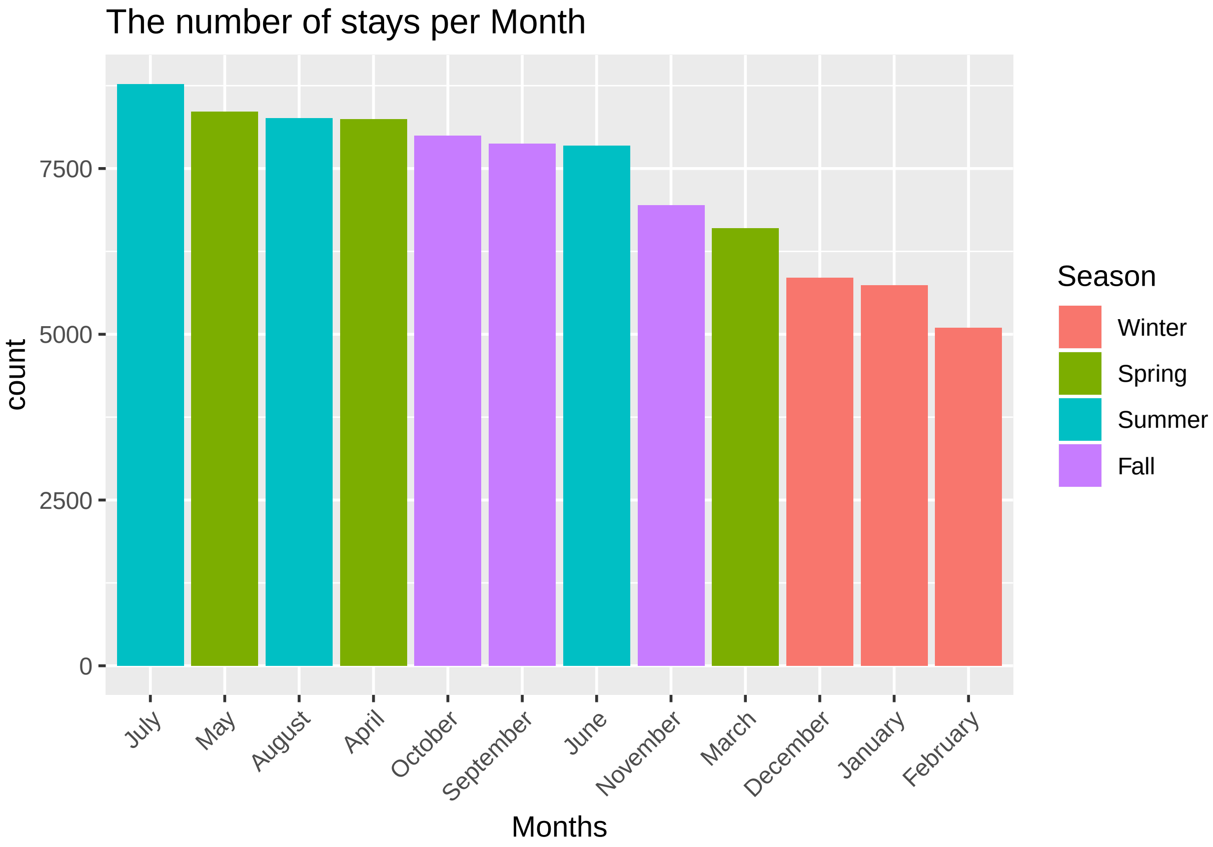
The dataset comprises 12 columns, which are the following:

```
colnames(en_reviewResponse)
```

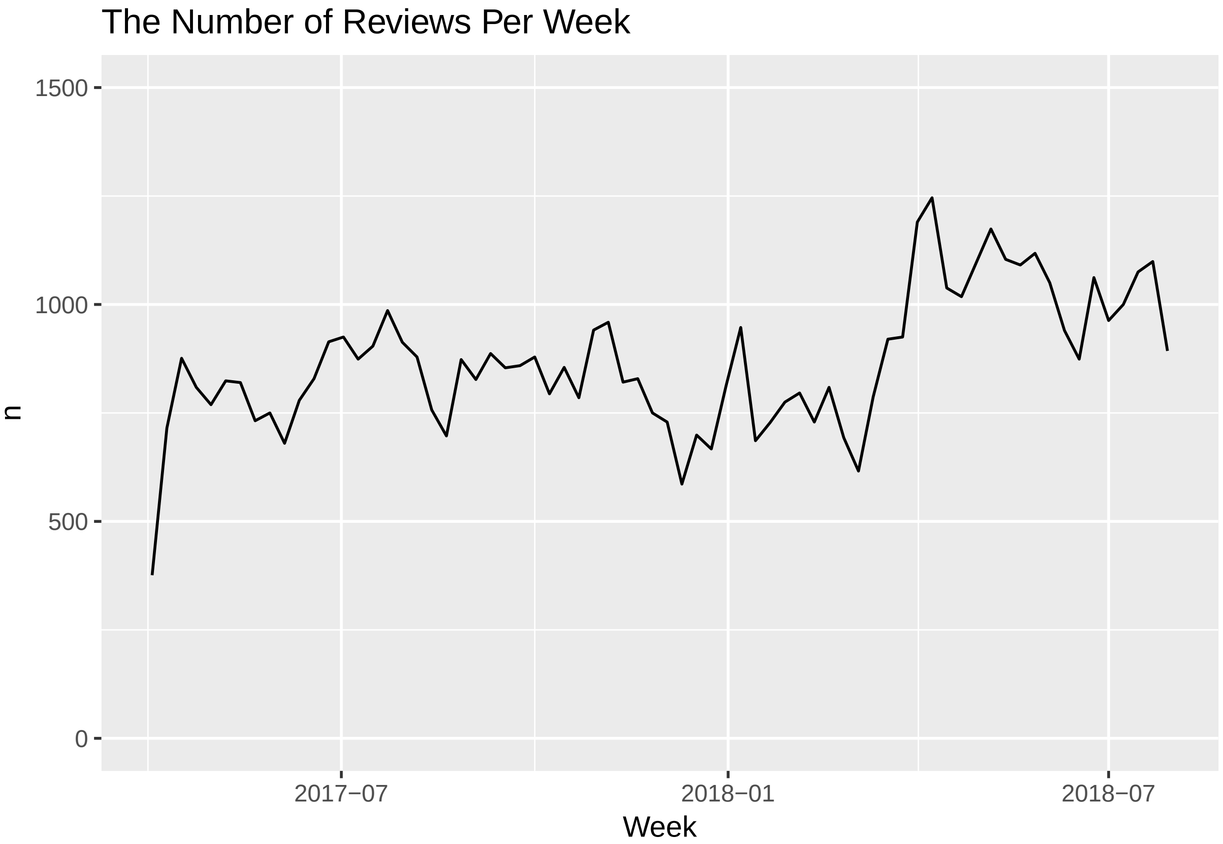
```
##  [1] "hotelid"          "reviewer_name"    "reviewer_country"
##  [4] "review_title"     "review_date"      "stay_date"
##  [7] "review_score"     "fullText"         "response"
## [10] "helpfulvote"      "city"             "reviewID"
```

There are 87633 english reviews in the dataset. The reviews range from 2016-12-09 to 2019-01-09.

Which are the busiest months of the year for hotels?

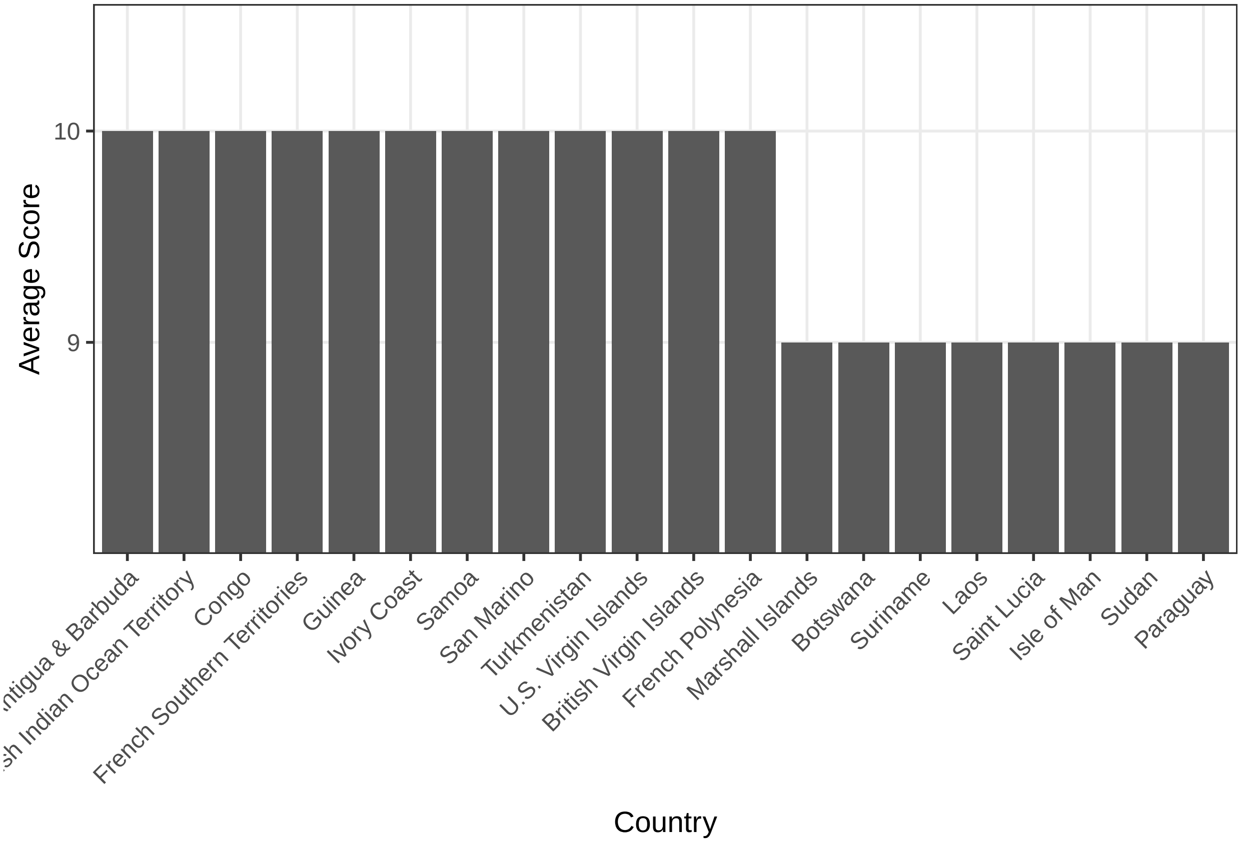


As the graph shows, hotels during nsummer and spring months received a higher number of tourists compared to winter months.

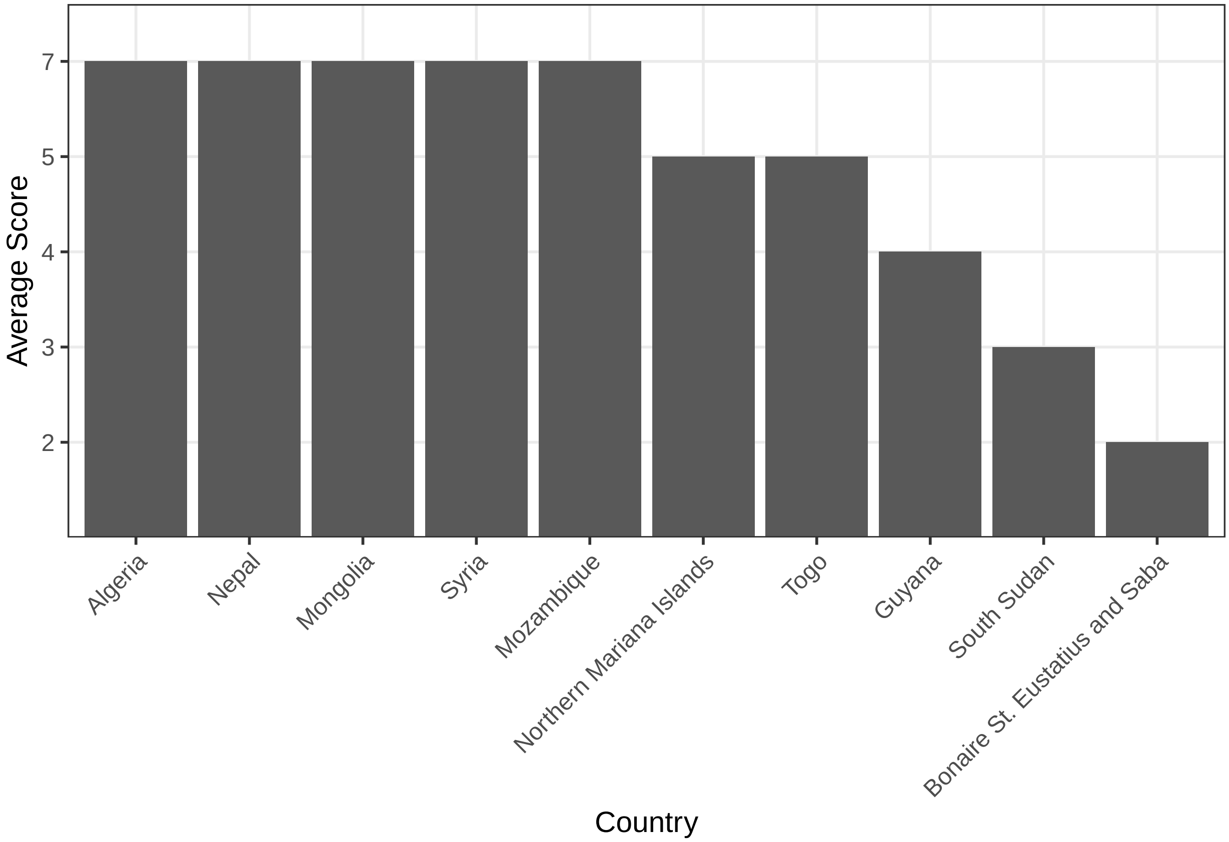


The highest number of weekly reviews was received within the half of 2018. The hotels received almost 1250 reviews in that week.

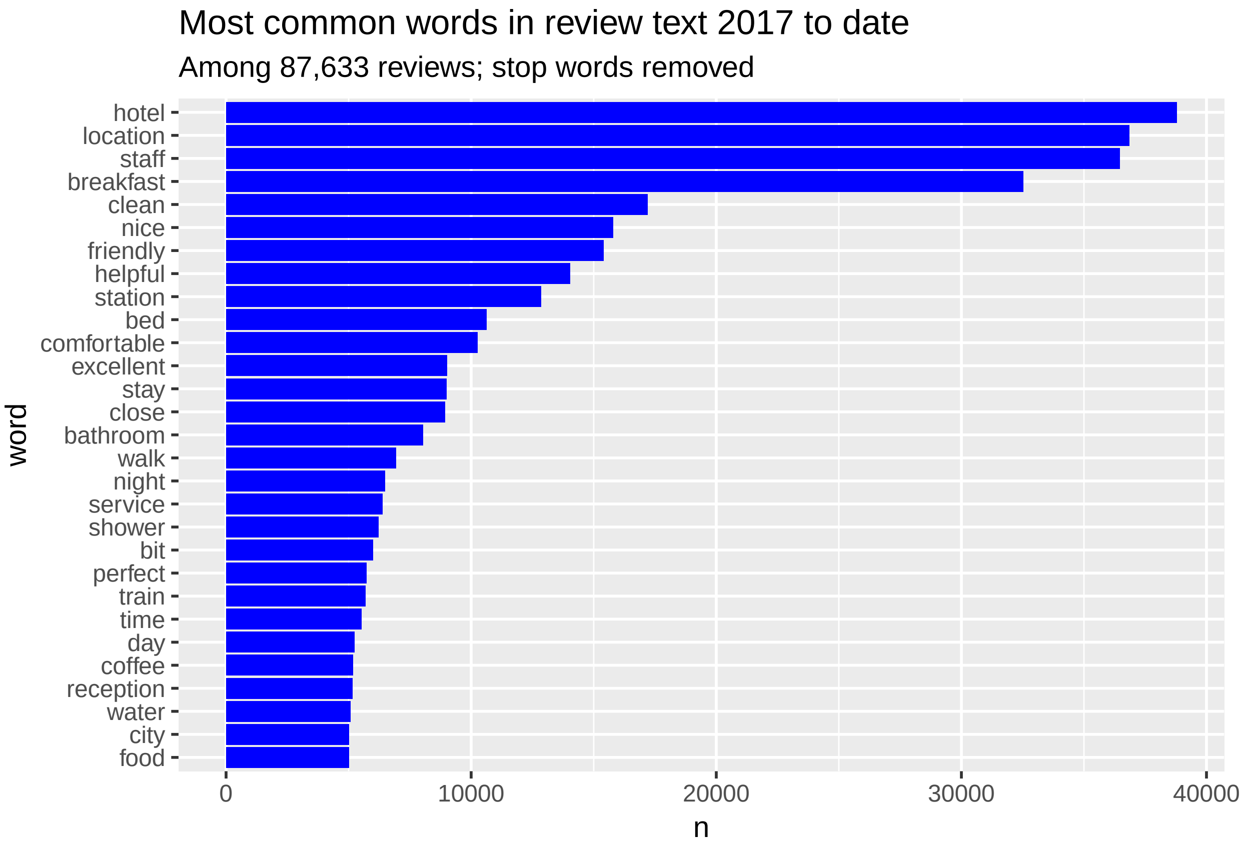
Countries with highest average score



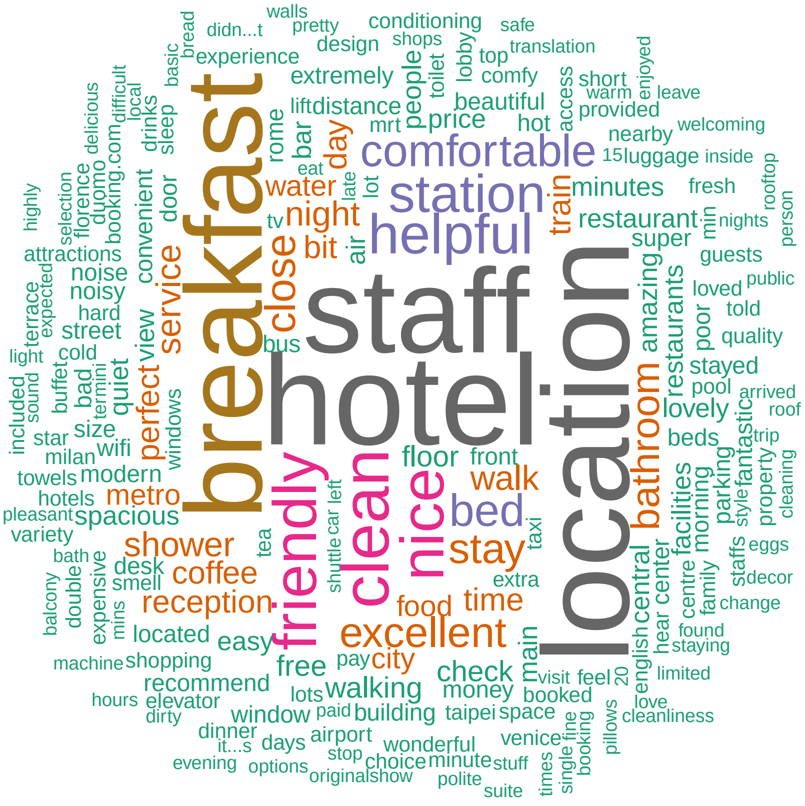
Countries with lowest average score



What are the most commonly occuring words in English reviews?

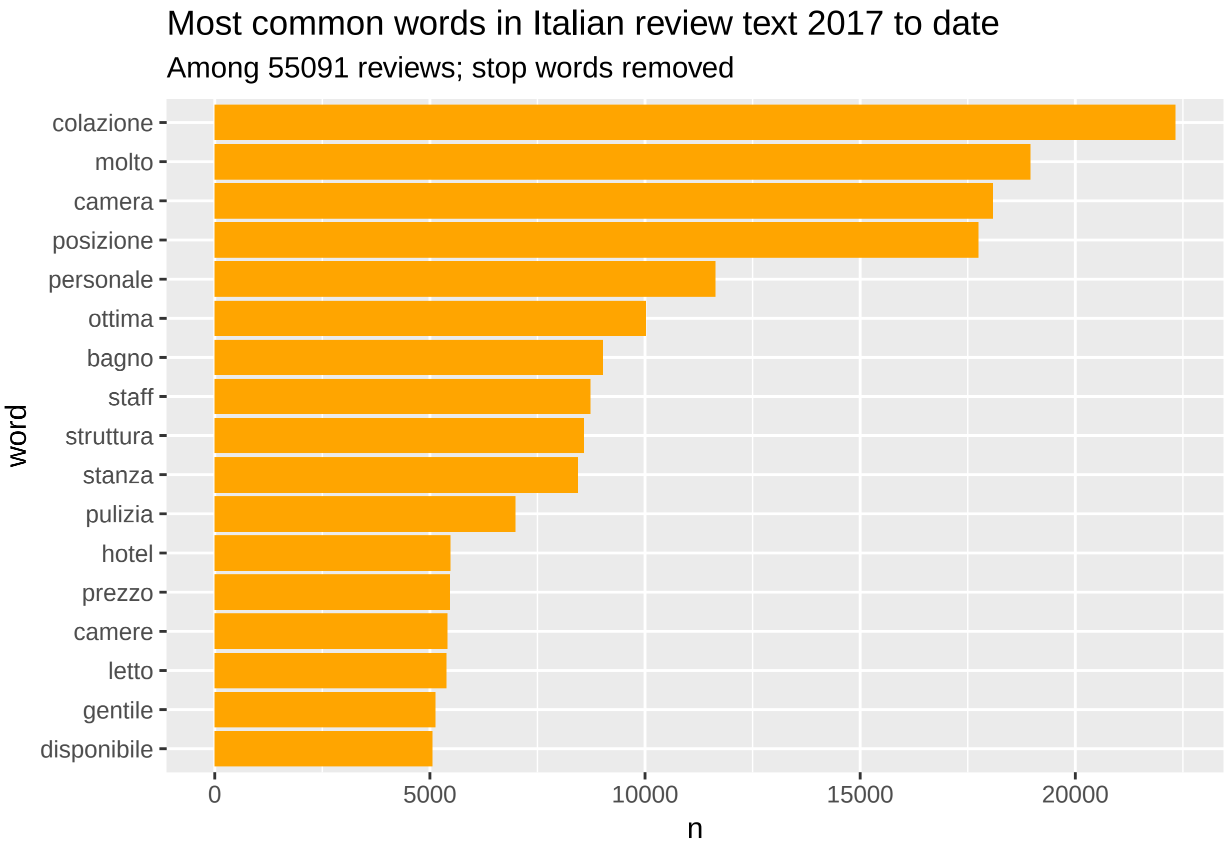


The importance of words can be illustrated in a wordcloud. The wordcloud clearly shows that “hotel”, “location”, “breakfast” and “staff” are the four most important words in Booking reviews in italian tourist cities.



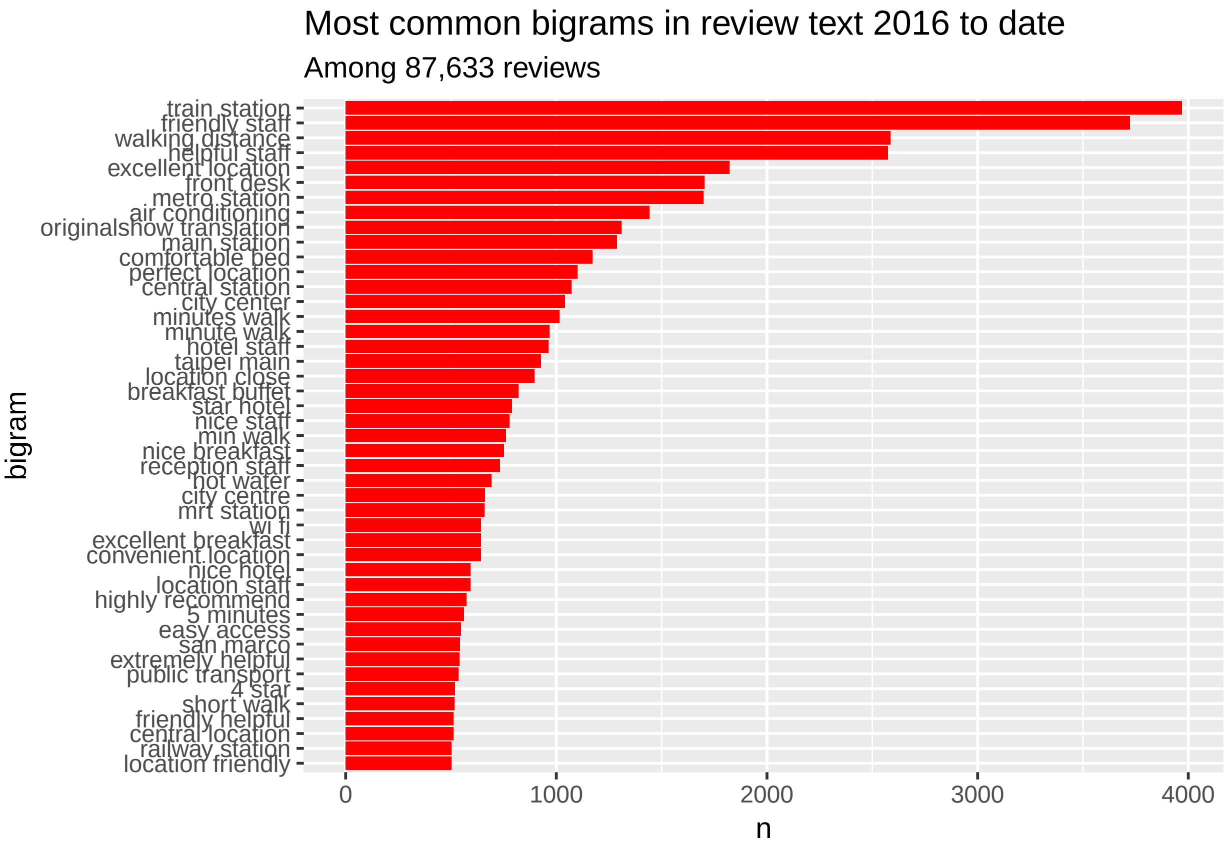
What are the most commonly occuring words in Italian reviews?

Conversly to english reviews, the most important word in italian reviews is "breakfast". For Italian speaking people, food topic appears to be particularly valuable in the context of hospitality.



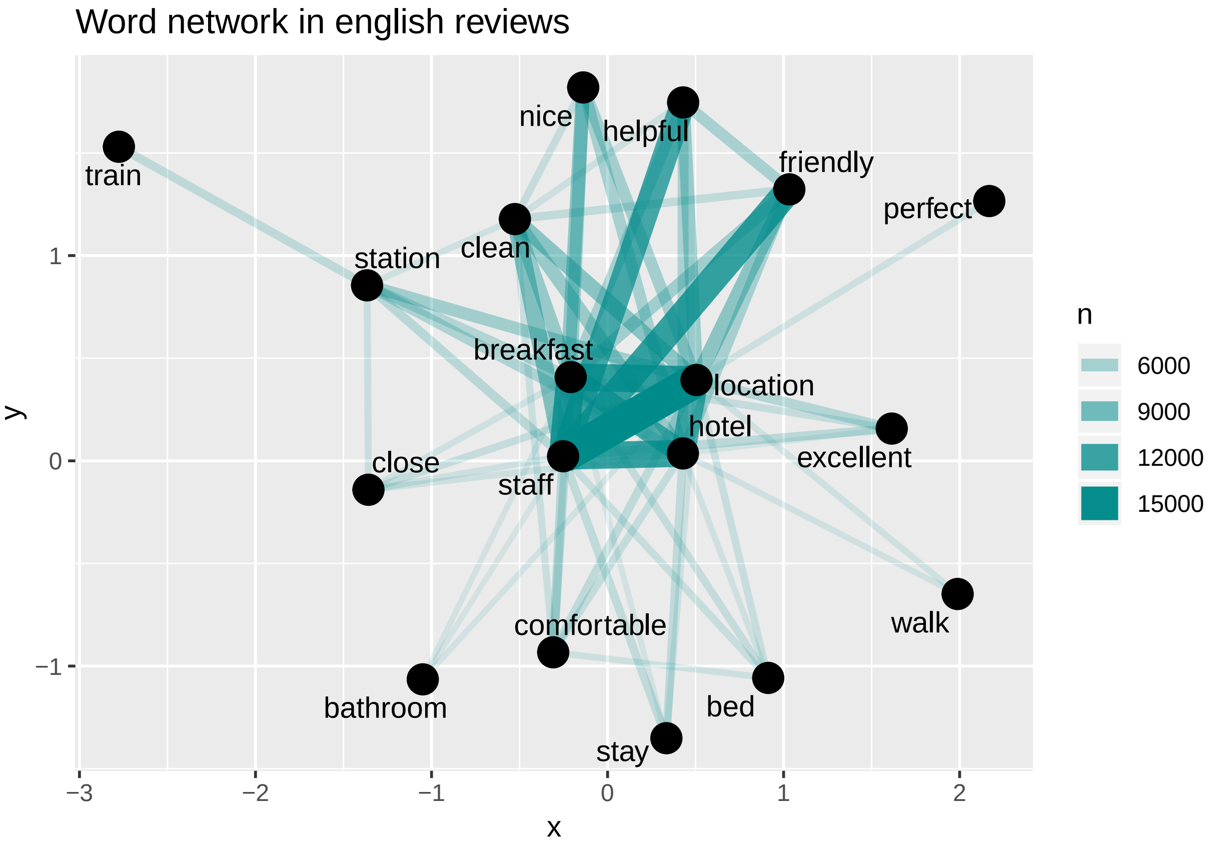
What are the most common bigrams in our reviews?

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?



The above graph visualizes the common bigrams in English reviews, showing those that occurred at least 500 times and where neither word was a stop-word.

Visualize bigrams in word networks:



The network graph shows strong connections between the top several words (“friendly”, “staff”, “excellent” and “location”, “train” and “station”).

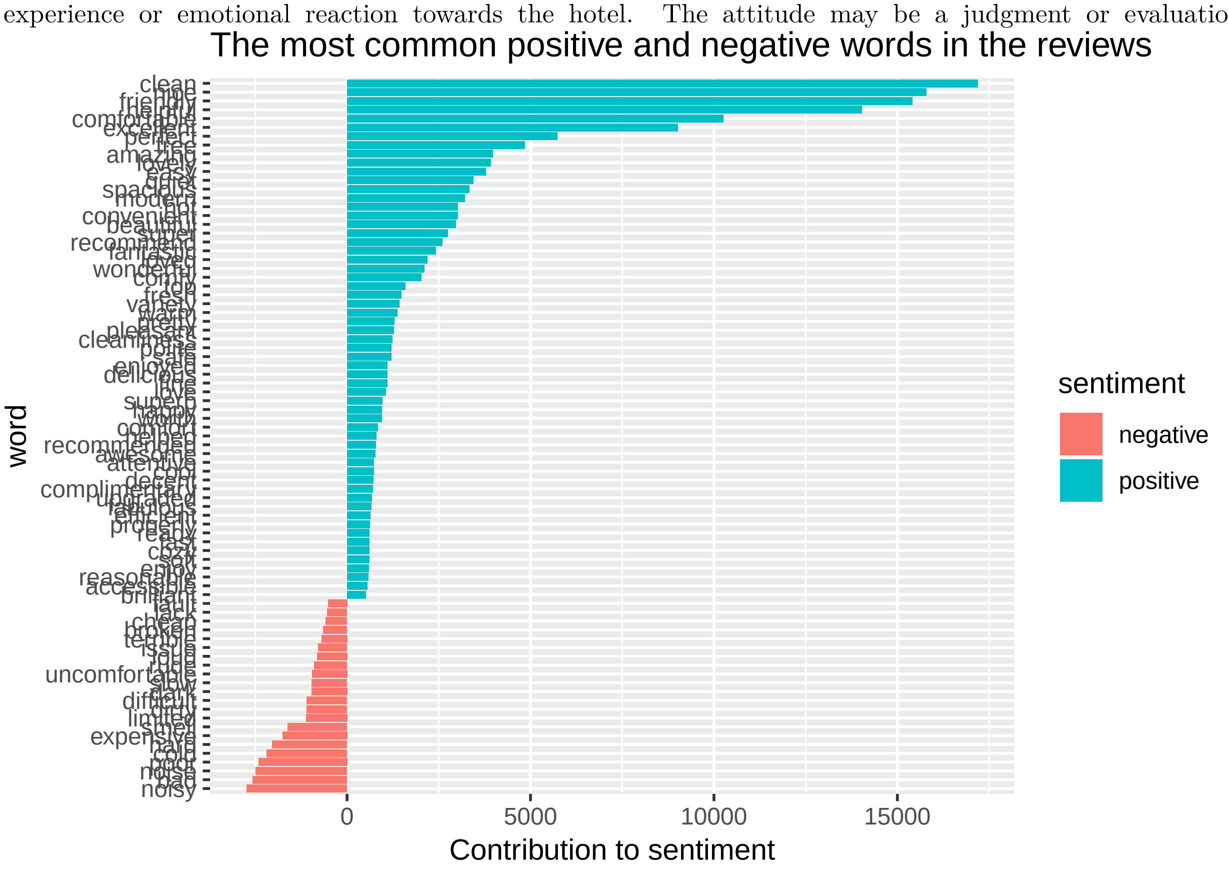
Sentiment Analysis

One way to analyze the sentiment of a text is to consider the text as a combination of its individual words and the sentiment content of the whole text as the sum of the sentiment content of the individual words. Sentiment analysis can be done as an inner join. Three sentiment lexicons are available via the get_sentiments() function. Let’s look at the words with a joy score from the NRC lexicon.

What are the most common joy words?

```
## # A tibble: 436 x 2
##    word          n
##    <chr>     <int>
##  1 clean     17190
##  2 friendly  15401
##  3 helpful   14035
##  4 excellent  9010
##  5 perfect    5730
##  6 food       5004
##  7 lovely     3925
##  8 money      2968
##  9 beautiful  2960
## 10 wonderful  2107
## # ... with 426 more rows
```

The aim is to determine the attitude of a reviewer (i.e. hotel guest) with respect to his (or her) past experience or emotional reaction towards the hotel. The attitude may be a judgment or evaluation.


The most common positive and negative words in the reviews
word
clean
nice
friendly
helpful
comfortable
excellent
perfect
free
amazing
lovely
easy
quiet
spacious
modern
hot
convenient
beautiful
super
recommend
fantastic
loved
wonderful
comfy
top
fresh
variety
warm
pretty
pleasant
cleanliness
polite
safe
enjoyed
delicious
fine
love
superb
happy
worth
comfort
helped
recommended
awesome
attentive
cool
decent
complimentary
upgraded
fabulous
efficient
properly
ready
fast
cozy
soft
enjoy
reasonable
accessible
brilliant
fault
lack
cheap
broken
terrible
issue
loud
rude
uncomfortable
slow
dark
difficult
dirty
limited
smell
expensive
hard
cold
poor
noise
bad
noisy
Contribution to sentiment
sentiment
negative
positive