ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Guests' Hotel Feature Preferences:

Analysis utilizing topic modelling and model based recursive partitioning

By

I.A. Grigoriev

MSc Thesis for Data Science & Marketing Analytics

Rotterdam, The Netherlands

5 / 11 / 2019

Student number: 435896

Supervisor: Prof. Dr. D. Fok

First Reader: Prof. Dr. V. Landsman-Schwartz

Abstract

In today's world of globally connected societies and ever-increasing ease of information retrieval, consumers have become accustomed to investigating peer reviews prior to making purchase decisions. Businesses in various sectors solidly value consumer ratings and maintaining a high rating from a business perspective. To maintain a high rating, knowledge about the determinants of review ratings can be very valuable to businesses, as products' and/or services' features can be adjusted accordingly. A growing collection of research concerning feature importance exists, but segmentation of reviewers is rarely incorporated. This is problematic, as it is highly unlikely that every reviewer rates a service or product the same way. In this research, I focus on hotel guests' reviews: I attempt to find out if different groups of guests have distinct feature preferences. I do this by deriving topics from *Booking.com* reviews using the *LDA* and *JST* models, use some of these topics as inputs in a regression model predicting the review rating, and segment reviewers into smaller groups/regression models with distinct hotel feature preferences using *Model Based Recursive Partitioning.* I find that the reviewers can be segmented into ten groups, all with some distinct feature preferences.

Contents

1 - Introduction

With the rise of web 2.0, and the general increasing ease of communication between various parties on a global scale, opinions matter more than ever. Often these opinions are expressed in the form of (web) reviews. Back as far as 2006, 30% of internet users rated products and services online (Pew Internet & American Life Project (2006), with consumers' contributions ever increasing. This shift to online reviews and associated eWOM (Electronic word-of-mouth) from traditional (face-to-face) WOM has significantly changed how consumers perceive and process opinions. Where traditional WOM is limited in its effect, which decreases over time and distance (Ellison & Fudenberg, 1995), eWOM generated through the continuous growth of the web content oversteps these limitations, and gives web users the ability to access information about products, services, and companies at any given moment (Duan, Gu, & Whinston, 2008).

With these developments, most websites offering products and/or services have developed a review system in some form, promoting users to share opinions and experiences in a transparent way. Maintaining good user review ratings is very important for businesses and their development: Consumer reviews are recognized as more credible than the information provided through marketing methods, and peer reviews are preferred over editorial reviews (Smith, Menon, & Sivakumar, 2005). Additionally, research by eMarketer (2007a) states that six out of ten consumers have a clear preference for websites with user reviews over websites without, and conversion rates on websites with peer written reviews are generally higher. Another study indicated that 80% of consumers consult reviews before making a purchase decision (Forrester Research, 2006), and 75% of US shoppers highly values reading reviews prior to a purchase decision (Bazaarvoice, 2007).

One of the industries affected by the increasing importance of online reviews is the hospitality industry. Research by eMarketer (2007b) states that 25% of infrequent travelers and 33% of frequent travelers that consult online hotel reviews have changed to a different hotel based on peer reviews. Park, Kim and Han (2007) found that consumers rely more on user ratings when the products and services being purchased are "high involvement" products, a label which applies to the hospitality industry. Research conducted by Compete Inc. (2006) found that 50% of travelers consulted reviews or read forums prior to booking their trip, and answers given during a survey conducted by Gretzel and Yoo (2008) indicated that 94.6% of participants found reviews to be a helpful tool to learn about new travel destinations, 91.9% used reviews to compare alternatives, and 91.8% used reviews to avoid a badly rated places. Clearly, findings indicate that hotel management (and consumers) can greatly benefit from the growing collection of reviews available.

The ever-increasing importance of user reviews has also caught the eye of computational linguists and machine learning specialists, some of whose research will be discussed in the literature section. Interestingly enough, existing research mostly

focusses on accurate prediction of review (star) ratings, and much less so on the identification of the underlying determinants of the rating. These determinants are important to businesses, as product/service features can be adjusted to achieve a higher rating. In academic literature that *does* deal with the importance of rating determinants in reviews, segmentation of reviewers is rarely included. This is problematic, as it is unlikely that all reviewers rate products/services the same way.

To try and fill this gap, I will focus my research on how opinions about hotel aspects (cleanliness/service/price/etc.) translate into ratings among different categories[1] of consumers, based on data obtained from *Booking.com*. The consumer categories are determined using a decision tree method, which can be used for segmentation based on significant differences in regression coefficients between two fitted General Linear Models (resulting from a split in the tree).

The relevance and usefulness of such research is twofold; on one hand, websites that have a review aggregation section like Booking.com, TripAdvisor and Trivago can incorporate this information into their recommendation systems. Given a user profile, his or her most likely determinants of a good hotel stay can be obtained, and subsequently hotels with high scores in those aspects can be recommended to the user. On the other hand, hotel management can use the information regarding consumers' hotel aspect preferences to tailor the consumers' stay to their needs, thereby increasing customer satisfaction/review ratings, and (hopefully) future visits and revenues.

The specific question I will try to answer in this research is:

"On what characteristics can hotel guests be segmented to reveal differences in their hotel-feature preferences, and what are these preferences?"

I try to answer this question in a few general steps. Topics are derived from the reviews. These topics do not necessarily translate into a single hotel feature: Some topics do, others translate into a combination of features, and others are vague and therefore hard to interpret as a feature. I do not predefine features in the same way Booking.com and similar websites do (in case of Booking.com these are: *Staff, Facilities, Cleanliness, Comfort, Value for money, Location* and *WiFi).* Instead, features and combinations of features follow naturally from the topic outcomes. This approach results in more specific features such as *Bed Size, Room view, Hotel design, etc.* From a managerial perspective this also seems more useful, because the predefined features can relate to anything. For example; good cleanliness can imply clean hallways, clean rooms, and clean dishes. Averaging these to derive general cleanliness seems counterproductive, and a waste of useful information.

[1] The Cambridge Dictionary defines a segment as *"one of the smaller groups or amounts that a larger group or amount can be divided into"* and a category as *"(in a system for dividing things according to appearance, quality, etc.) a type, or a group of things having some features that are the same"*. For the purpose of this research I will use the terms interchangeably, as both definitions apply.

The occurrences of these topics in reviews (i.e. what each review is about) are used as predictors in a number of regression models fitting them to the user review rating. The models should reflect meaningful differences in the rating generating process/feature preferences of users, so a segmentation method that splits reviewers based on these differences is applied.

A methodology that, to my knowledge, has not previously been adopted will be used: The topic occurrences in each review are created using topic models *LDA* and *JST*. The reasons why these topic models are appropriate for the used dataset are discussed in detail in section 4.2. The topic occurrences are used as input for a *Model Based Recursive Partitioning* model. Based on parameter instabilities in user/user stay characteristics, the model can segment reviewers into groups with similar aspect-to-rating generating processes (topic coefficients in regression).

The rest of this paper is structured as follows: Section 2 briefly discusses the idea of topic modelling, and evolution of topics models, new models, and research related to mine. In section 3 the dataset used for the research, and the preprocessing of the data is discussed. Section 4 provides extensive information about the methods used to answer the research question. Section 5 contains the obtained results from the application of the methods discussed in section 4, and in section 6 the research is concluded, by answering the research question, discussing limitations, and recommending ideas for future research.

2 – Literature Review

2.1 – Topic modelling

Topic models are methods within machine learning and natural language processing to discover hidden abstract topics in a selection of documents. The following example describes an intuitive hidden topic structure: A text containing the words *lion*, *giraffe*, *dog*, and *cat*, and consists of three topics; *pets*, *zoo* and *feline. Cat* and *Lion* are part of the *feline* topic, *cat* and *dog* are part of *pets*, and *lion* and *giraffe* are part of the topic *zoo.* From this example, an important characteristic of words in topic models follows: Words can belong to multiple topics. Topic models do not know the "correct" number of topics; the number is usually defined by the researcher. Topic models are based on the same assumption: each given topic consists of a collection of words, and each document consists of a mixture of topics.

2.2 – Evolution of topic models

Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) (Furnas, et al., 1988) is one of the principal techniques on which topic models are based. It introduced the idea of decomposing available textual information (the terms per document) into separate matrices: A document-topic matrix (to what latent topics is a document related?) and a topic-term matrix (to what topics are words related?). They are created using Singular Value Decomposition (SVD), a technique for the factorization of matrices in linear algebra; it is simply a low dimensional representation for the documents and words.

Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) built on the general idea of LSA, but replaced the SVD method with a *probabilistic* method. Instead of using decomposition, pLSA assumes that documents are generated by chance: For each given topic, each word has a certain probability to be drawn from it; in each document, each topic has a probability to be present in it. The probabilities are obtained through expectation maximization (EM), a method used to find the most likely parameters in certain models.

Currently, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is the most used model for topic modelling. It functions much like pLSA, but with the addition of Dirichlet priors that control the distributions of topics in documents and words in topics. These can be adjusted by hyperparameters to:

- Make documents focused on specific/many topics

- Make topics focused on specific/many words

LDA will be applied in this research. The reasons why the model is appropriate for the dataset are discussed in section 4.2, and the model itself is discussed in more detail in methodology section 4.3.1.

Most novel topic models in academic research are variations of the LDA model or build on the framework of LDA. One of these is the Joint Sentiment Topic model (JST) by Lin & He (2009). *Sentiment models* are models that capture positivity and negativity in text. JST functions like LDA, but with an additional "sentiment layer" that splits topics into positive, neutral, and negative subtopics. Regarding other sentiment topic modelling approaches, they state that:

- Sentiment models trained on one domain are not necessarily applicable to other domains.
- Most sentiment modelling methods are supervised, and text is pre-classified as Positive/Neutral/Negative. In practice, pre-classified texts are rare.

- Existing models first find features/topics, and only afterwards assign a sentiment label.
- Sentiment polarities are dependent on domains: the word *unpredictable* in relation to *driving* should be perceived as negative, while *unpredictable* in relation to the word *movie* is positive.

JST will also be applied in this research. The reasons why are discussed in section 4.2, and model specifics are discussed in section 4.4.1.

Other variations of LDA include:

- Multi-grain LDA (MG-LDA): assumes the existence of local and global topics in text (Titov & McDonald, 2008).
- Sentence LDA (sLDA): assumes that each sentence is limited to one topic (Balikas, Amini, & Clausel, 2016).

Qiang, Qian, Li, Yuan and Wu (2019) compare the performance of various newly proposed topic models aimed at dealing with short texts to LDA. The datasets used for comparison all have a relatively short average document length; table 1 shows the number of topics per document (*K*), the number of documents per dataset (*N*), the average and maximum length of the documents (*Len*), and the size of the vocabulary (*V*).

Dataset	*K*	*N*	*Len*	*V*
SearchSnippets	8	12,295	14.4/37	5,547
StackOverflow	20	16,407	5.03/17	2,638
Biomedicine	20	19,448	7.44/28	4498
Tweet	89	2,472	8.55/20	5,096
GoogleNews	152	11,109	6.23/14	8,110
PascalFlickr	20	4,834	5.37/19	3,431

Table 1. Performance comparison datasets (source; *Short Text Topic Modeling Techniques, Applications, and Performance: A Survey,* Qiang, Qian, Li, Yuan and Wu, 2019)

The proposed model types are;

1) *Dirichlet Multinomial Mixture based Methods. (DMM)*

These models are based on the assumption that each document is sampled by only one topic, which can often be a reasonable hypothesis for short texts. These models include The Gibbs Sampling DMM (GSDMM), Latent Feature DMM (LF-DMM), and Generalized Polya Urn Poisson DMM (GPU-PDMM)

2) *Global Word Co-occurrences based Methods*

These methods are based on the assumption that closer words are more relevant to each other than distant words in a text. Word co-occurrences are extracted using a sliding window. The models based on this are the Biterm Topic Model (BTM) and the Word Network Topic Model (WNTM).

3) *Self-aggregation-based Methods*

Self-aggregation-based Methods ease the problem of sparseness in short text, by merging by merging them into longer pseudo documents, prior to deriving the latent topics. Self-Aggregation based topic modeling (SATM) and Pseudo-document-based topic modeling fall into this category.

The average classification accuracy of the models can be seen in figure 1. Newer proposed models rarely (considerably) outperform the standard LDA approach, illustrated by the orange bar in the figures.

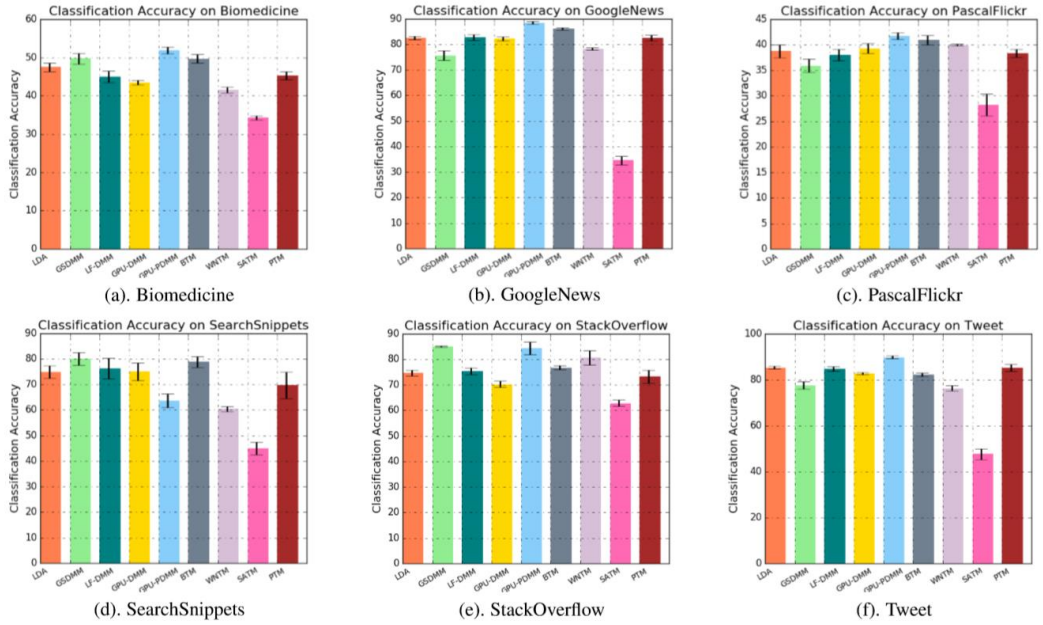


Figure 1. Average classification accuracy LDA vs. Short text topic models (source: *Short Text Topic Modeling Techniques, Applications, and Performance: A Survey*, Qiang, Qian, Li, Yuan and Wu, 2019)

2.3 – Similar & related research

Ganu, Elhadad and Marian (2009) state that user experience, when browsing reviews, would be greatly improved if the structure of the content and sentiment towards each aspect were taken into account in recommendation systems. Unfortunately, often only an aggregated rating and review is available, and users writing reviews regularly enter informal text containing poor spelling.

In Project URSA (User Review Structure Analysis) Ganu, Elhadad and Marian predict the rating of restaurant reviews by training a support vector machine to recognize aspects and polarity of sentences, based on annotated training set sentences.

Wang, Lu and Zhai (2010) introduce Latent Aspect Rating Analysis (LARA) problem. Given an overall rating of a product, the problem is defined as:

- Find ratable aspects of the product
- Assign a latent rating to each aspect
- Assign weights to each aspect (how strongly does the rating of that aspect influence the overall rating)

Wang, Lu and Zhai formally capture this problem in their Latent Rating Regression Model (LRR).

Jo and Oh (2011) state that for product and service reviews, a simple average rating does often not suffice. When buying a camera, one user might value the lens, and another the battery, thereby making the aspect-level sentiment of greater importance than the rating. Similarly to Lin and He, they observe that polarity is dependent on the domain; a *long* battery life is perceived as positive, but can be neutral or negative in other context. To tackle the problem of finding aspects, and associated sentiment words, Jo and Oh propose the *Aspect and Sentiment Unification Model* (ASUM).

ASUM is a modified version of Sentence Latent Dirichlet Allocation, which in turn is a special version of regular LDA. SLDA assumes that each sentence is constrained to a single topic. ASUM builds on this model by including sentence level sentiment.

Like other researchers in the field, Wang and Ester (2014) state that aggregated reviews do not provide detailed information; products and services can have the same rating, but can differ in their satisfactory and unsatisfactory aspects. Vice versa, two products or services can have different ratings, but be equal in quality in some aspects.

Wang and Ester present the Sentiment-Aligned Topic Model (SATM), which focuses on the alignment of sentiment labels (closely related to aspect ratings) with aspect phrases polarities, e.g. < *dirty, room*>.

Kamath, Ochi and Matsuo (2015) observe a similar problem in rating prediction as Wang and Ester. They add that different users may value different aspects of a product or service. In other words, for each person the contribution of each aspect to the total rating is distinct.

To predict the ratings of restaurant reviews, Kamath, Ochi and Matsuo use the latent topics obtained from Multi-grain Latent Dirichlet Allocation (which assigns a topic/aspect to each sentence) and the polarity score of each sentence. They calculate

the total polarity of each aspect in a review, by averaging the polarity scores of all sentences relating to that aspect.

To derive which people value which aspects, Kamath, Ochi and Matsuo use the idea of *representative users.* These are users who's polarity scores of an aspect correlate strongly with their given review ratings: Suppose user A wrote 5 reviews, and gave ratings [5, 5, 4, 5, 1], and the polarity scores of the topic *cleanliness* are [1, 1, 0.5, 1, -0.5] in his reviews. If the Pearson Correlation between the vectors is high (above some determined threshold), user A can be seen as a representative user for cleanliness.

Using this method, users can be segmented according to what aspects/topics have value in their rating determination. It can also be used to rate a certain unobserved aspect of a product or service, by only considering representative users' ratings.

Similarly, to Kamath, Ochi, and Matsuo, the goal of my research is to segment users based on their aspect preferences, but my segmentation approach, and the goal of the segmentation, is vastly different. As stated, the method applied by Kamath, Ochi and Matsuo is focused on the selection of groups of representative users for each aspect. However, there is no guarantee that the characteristics of the people in these groups will be interpretable, as representative users for a certain aspect could have nothing in common. Interpretability is important to my analysis, as it allows (future) guests to be placed easily in a segment, therefore easily revealing their preferences. Model Based Recursive Partitioning achieves this easy interpretability: By simply following the split rules down the tree, each segment's reviewer characteristics can be read. Additionally, Model Based Recursive Partitioning Focuses on *all* topics simultaneously in its segmentation. This implies that two user segments could value some topics' aspects equally, as long as at least some other valuations of aspects are (significantly) different between them.

2.4 – Review usefulness to management

Intuitively, receiving good reviews and maintaining a good average review rating seems desirable for managers. However, resources may be better spent on something else if positive reviews do not effect businesses' actual performance. Research findings regarding the connection between reviews and performance are varied.

Duan, Gu and Whinston (2008) analyzed the effect of user reviews on films' daily box office performance. They used a two-equation system; one equation predicts the daily revenue by cumulative user rating, last day's revenue, average rating, and a weekend dummy. The second predicts the number of posted reviews by daily revenue, last day's number of review posts, cumulative number of reviews, and a weekend dummy. After accounting for endogeneity, captured by the two-equation system, they

did not find a significant effect of user reviews on revenues. However, revenues were found to affect review post volume (awareness effect).

Ye, Law and Gu (2009) researched the effect of review ratings on room sales in three Chinese cities. They use the number of reviews as a proxy for room sales, and average rating, rating variance, stars, room rate, and city GDP rank (out of the three cities) as explanatory variables. They found a statistically significant positive effect of average rating, and a negative effect of rating variance on room sales.

Zhang, Ye, Law and Li (2010) investigated the effects of a collection of separate restaurant ratings (taste, environment and service) in China on restaurant popularity, using restaurant website visits as a proxy. Including a number of control variables (number of reviews, popularity index, average cost and an editors' comment dummy) Zhang et al. found varying effects of the ratings: All three affect popularity positively, but environment to a lesser degree and less significantly compared to taste and service.

Tuominen (2011) explored the relationship between hotel performance, reflected by revenue per available room, and the average review rating, number of reviews, recommendation percentage, and ranking on Trip Advisor. All but ranking are found to affect room revenue significantly.

Archak, Ghose and Ipeirotis (2011) researched the effect of Amazon camera and camcorder feature opinions, consisting of a feature paired with an evaluation (obtained from decomposing text into segments) on sales rank. Most feature opinions were found to have a significant effect on the sales rank of cameras and camcorders. Archak et al. argue that their developed method can be used by manufacturers to identify important determinants of sales, facilitating impactful change to the products. (Online) retailers can use the information to highlight product features in advertisements.

With a few exceptions, most research seems to demonstrate a significant effect of reviews and review ratings on business performance indicators such as sales, revenue, and popularity.

3 – Data

3.1 – Booking.com

Booking.com started in 1996 as a small startup in Amsterdam with the mission: *"Empower people to experience the world."* (Booking.com, 2019) Nowadays, it counts 17,000 employees, spread over 198 offices in 70 countries around the globe.

It functions as a travel fare aggregator, and search engine for lodging reservations. A variety of accommodations are listed; apartments, luxury hotels, vacations homes and more. User reviews are readily available for most countries; an average Dutch hotel has more than 1,000 reviews.

To post a review, users need to have booked the accommodation through Booking.com. After their trip, users receive an e-mail from booking.com, encouraging them to write a review. This ensures the authenticity of reviews, and counteracts fake grading.

3.2 – Dataset

The used dataset is obtained from Kaggle.com, and was originally scraped from Booking.com's accommodations review section. It contains 515,738 reviews of hotels in the Netherlands, France, Spain, United Kingdom, Austria and Italy from August 2015 to August 2017. An example of a Booking.com review can be seen in figure 2.

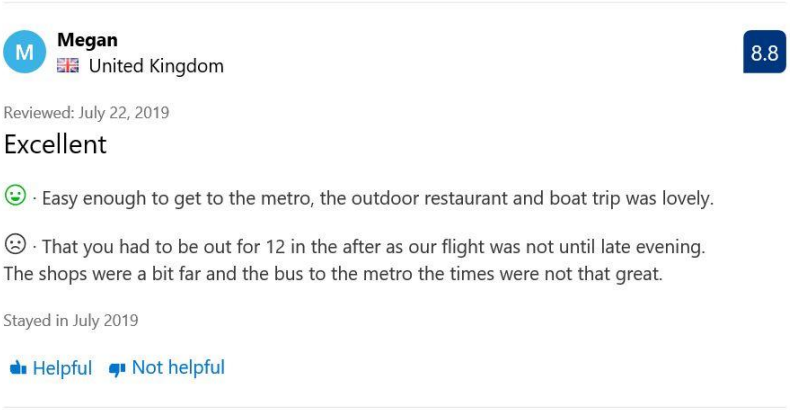


Figure 2. Review example (source: Booking.com)

The dataset contains two separate variables; *Positive_Review* and *Negative_Review*. Reviews on Booking.com have a predetermined positive and negative part. This alleviates problems with negations presented by Baccianella, Esuli and Sebastiani (2009). Negations become less problematic, because they can be deduced from the location of words (in *Positive_review* or in *Negative review).* For example; when *friendly* and *staff* occur in the *Negative_review*, it is clearly part of a negation, because *friendly staff* is a positive expression by itself. Both the positive and negative review sections tend to be relatively short; few contain more than 50 words, and almost none contain more than 100 words. The distribution of the positive review section word count can be observed in figure 3, and the negative word count in figure 4. Both distributions are skewed and gradually decreasing, but peak at different points; the most common

amount of words used for positive reviews is 6 or 7, while the negative review word count tends to be closer to 0. For both review sections, the average number of words used is +/- 18, indicated by the dotted red lines in the figures.

Missing entries in the *Positive_Review* and *Negative_Review variables* contain the text *"No Positive"* and *"No Negative"*, and do not count towards the total wordcount.

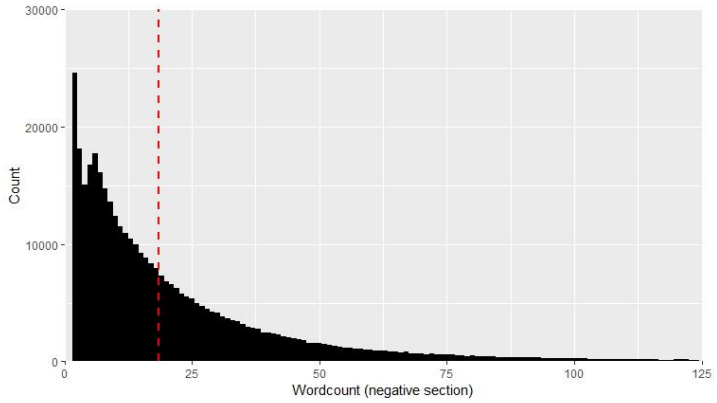


Figure 3. Negative reviews word count

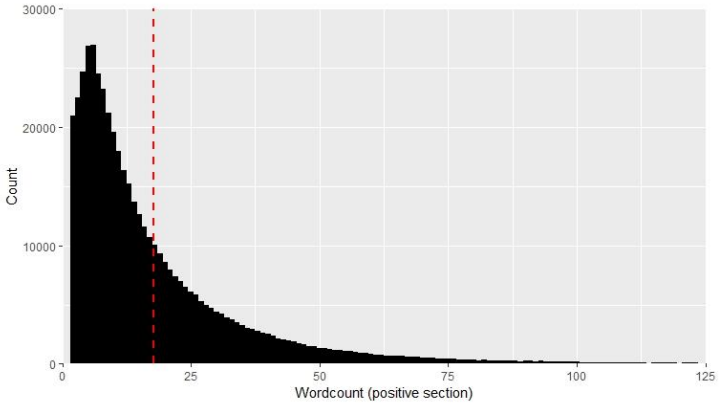


Figure 4. Positive reviews word count

Stopwords and interpuntion are removed from the reviews in *Positve_review,* and *Negative_review,* and words are stemmed using the *Porter Algorithm* (Porter, 1980)

The variable *Reviewer_Score* indicates the score given to the hotel by users. The score distribution can be observed in figure 5. It is strongly negatively skewed, with an average given rating of +/- 8.5 indicated by the dotted red line. Almost no ratings below score 2 are found in the dataset.

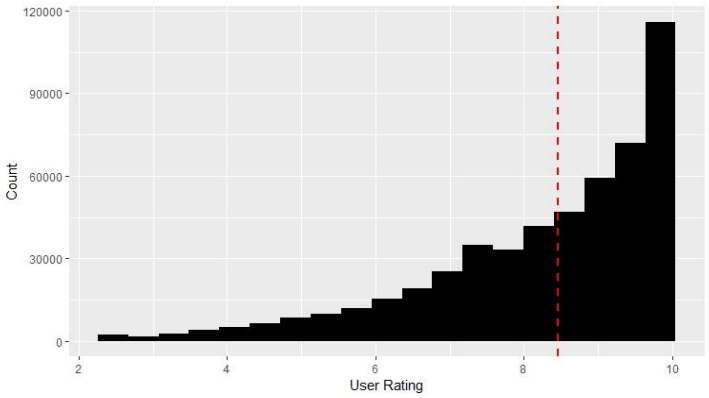


Figure 5. User Ratings

Other variables include; *Hotel name, address, State, City, Longitude, Latitude, Businesses in 100m/1km/5km, Room type, Bed type, Guest type, Trip type, Stay duration, Review date, day of the week/month/year, week of the month/year, Quarter of the year, Reviewer nationality, Total number of hotel reviews, Average hotel review score, Total number of reviewer reviews, Submitted from mobile phone (Yes/No)*. The variables *Is_Reviewer_Holiday*, and *Is_Hotel_Holiday* were added by the original poster of the dataset, indicating what days are holidays in the reviewer's home country, and the destination country.

The use of the Review date variable (and related variables such as day of the week, day of the month, holiday, etc.) can be problematic for data analysis, as it indicates the date when the review was posted, not when the trip took place. No data is available regarding the amount of time that passed between the trip and the posting of the review. *Quarter of the year* could possibly be used for research purposes, if the assumption is made that time between a trip and the posting of a review is limited.

The *City* variable (and therefore also *State*) is also of limited use, because some cities in the dataset are simply suburbs or regions within a larger city contained in the dataset (Amsterdam Zuid-oost for Amsterdam, Paris 13 for Paris, etc.). London is by far the most represented city in the dataset, with a total of 243,488 reviews (47%), not accounting for suburbs represented as full cities.

A subset of only Amsterdam hotels is used for the research; both methods used for topic construction experience problems when dealing with a dataset this big. The distributions of variables discussed and shown in figures 3, 4 and 5 are the same in this subset of the data. The Amsterdam subset consists of 57,107 reviews.

4 – Methodology

4.1 – Overview

As described in the Data section, (most) booking.com reviews are defined by the following characteristics, each of which have influenced some choice(s) made related to the methodology of this research:

- Short
- Not always structured in (correct) sentences
- Strongly negatively skewed user ratings (towards 10)
- Predefined negative and positive review sections.

In the following sections, I justify the choice of the LDA and JST models as opposed to other topic models. I discuss the theory behind the LDA and JST models, and the selection of the amount of topics and hyperparameters values. I explain *Tobit* regression, and the two regression methods it relates to, and how the output from the topic models is used in the Tobit model. Then I illustrate the basic ideas behind decision trees, and describe a special variation: Model Based Recursive Partitioning (tree). Lastly, I describe how MBRP can be used to answer my research question.

4.2 – LDA and JST vs. other topic modelling methods

Although newer methods like ASUM (Jo & Oh, 2011) LRR (Wang, Lu & Zhai, 2010) and steps in the URSA project (Ganu, Elhadad, & Marian, 2009) are specifically designed to work with ratable aspects, they are incompatible with the dataset used for this research: They assume that each sentence in a review relates to *one* aspect. The reviews written on Booking.com often have the form of a summary, where opinions about the aspects are separated by commas, spaces, or the word "and". Figure 6 is a good example of this sentence structure: The first sentence of the positive review (happy smile), and the negative review (sad smile) clearly relate to more than one aspect of the hotel stay.

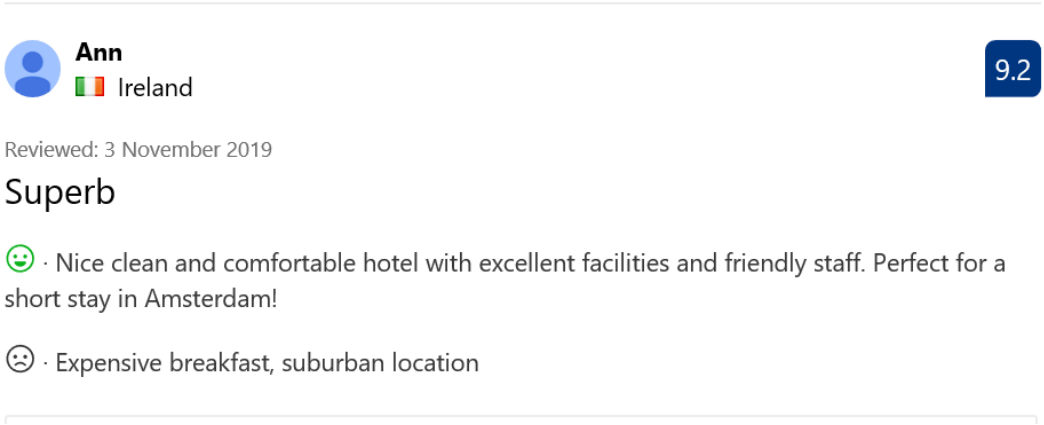


Figure 6. Example sentence structure Booking.com review

Qiang, Qian, Li, Yuan and Wu (2019) found only minor differences between the performance of LDA and short text topic models. Therefore, I assume LDA should suffice for effective topic modelling in my research.

JST can possibly add valuable information to the (LDA) results by introducing sentiment layers. Not only does JST give insight into what reviewers like or dislike, but *how strongly* did they like or dislike a certain hotel feature.

4.3 – LDA

4.3.1 – Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was first introduced by Blei, Ng and Jordan (2003). The algorithm treats documents as bags of words, meaning the order of the words is not taken into consideration. A fundamental assumption is at the center of LDA; documents are generated by picking a topic from a topic/document matrix, and from this topic a word from a word/topic matrix, and repeating until the document length is met. The word/topic matrix contains the probability of each word to be selected when sampling a topic (for each topic), and the topic/document matrix contains the probability of selecting each given topic when sampling a specific document (for each document).

The topic/document matrix is assumed to be constructed by the following generating process: Draw a sample from a Dirichlet distribution for each document, with a selected hyperparameter β as input, and fill in the obtained probabilities of sampling each topic.

The Dirichlet distribution can be regarded as a "distribution of distributions". Different distributions for various selected β's are illustrated in figure 7 with the three numbers above the planes indicating β's for each topic. In the first plane (and

generally for values where $\beta < 1$), distributions are concentrated near the corners, implying that distributions will strongly favor one topic ({1,0,0}, {0,1,0} and {0,0,1}). When $\beta=1$, any distribution can be sampled, as can be seen in the second plane. For increasing values of β after $\beta=1$, the probability of drawing a "balanced" distribution where in {a,b,c} a, b, and c are close to each other increases, eventually becoming {1/3,1/3,1/3}, meaning all topics have the same sampling probability.

The word/topic matrix is composed in a similar manner to the topic/document matrix, but a Dirichlet sample is drawn for each topic instead of each document. The hyperparameter for the Dirichlet distribution here is referred to as α. The distribution in this case refer to the probabilities of sampling words.

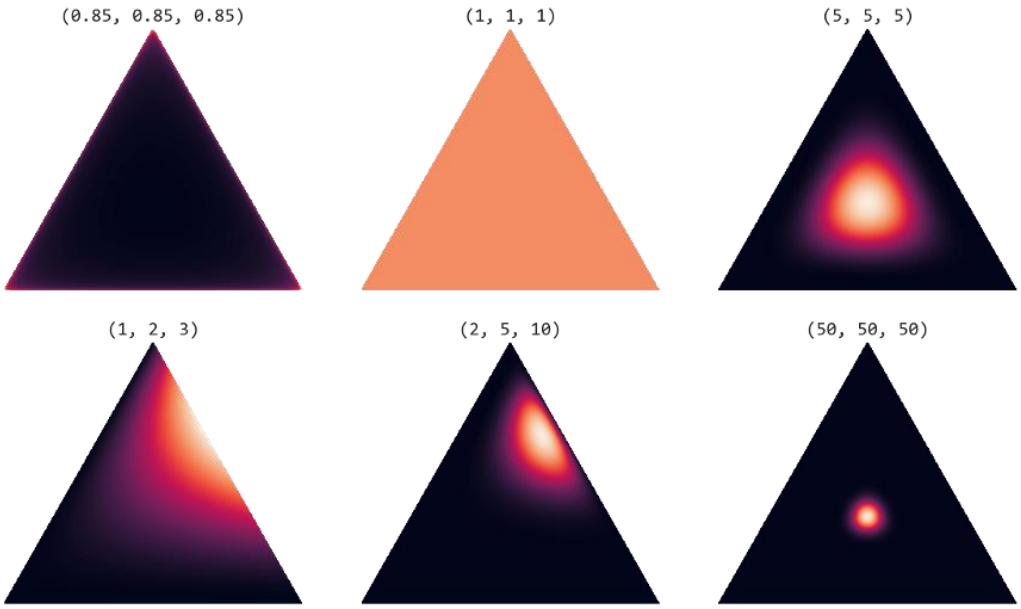


Figure 7. Three-topic Dirichlet distributions for different parameter values (source: towardsdatascience.com)

The real topic/document and word/topic distributions (drawn from the Dirichlet distribution) are unknown. We are interested in finding the posterior distribution of latent variables (topic assignments Z, topic/document distributions θ, and word/topic distributions φ) given the parameters α and β, and the observed data W:

$$P(Z,\theta,\varphi)|W,\alpha,\beta) = \frac{P(W,Z,\theta,\varphi|\alpha,\beta)}{P(W|\alpha,\beta)}$$

The form of the denominator makes this distribution untraceable, because of the coupling between θ and φ in the summation over latent topic assignments (shown after marginalizing over the hidden variables (Blei, Ng, & Jordan, 2003)):

$$P(W|\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

However, the distributions can be estimated using various methods. One of these is Gibbs sampling, which is usually used for the construction of LDA models.

Gibbs Sampling is useful when direct sampling is not possible/proves difficult: it requires sampling from conditional distributions, which we will call $p(x|y)$ and $p(y|x)$ in this two-variable example. $(x_0|y_0)$ is set to some starting value. $x_1 \sim (x|y_0)$ is sampled from the first conditional distribution, followed by $y_1 \sim (y|x_1)$, $x_2 \sim (x|y_1)$, $y_2 \sim (y|x_2)$, etc. This process is repeated N times. At some point, the sample averages converge. Samples prior to the last sample are discarded, which is referred to as the *burn-in.*

4.3.2 – Application of LDA

Using the *topicmodels* package from CRAN, a separate LDA model is constructed for the negative and positive sections of reviews, as combining these into one LDA model would lead to information loss regarding polarity: In a review without predetermined positive and negative parts, a reviewer could have something good to say about one thing (e.g. *clean floor)* and something bad about another (e.g. *dirty window)*. Because LDA treats reviews as *bags of words,* it would be unclear which one is clean; the floor or the window (similarly for *dirty).* Moreover, sentiment words may not even be needed for the interpretation; the presence of *floor* in either the positive or negative part of the review reveals the associated sentiment.

Both resulting LDA models contain the topic/document matrix (θ from 4.3.1); it details the topic distribution of each document/review in the corpus.

The number of topics (K) for the positive and negative reviews are selected using the *ldatuning* package, which combines four academic methods that qualify how well the model fits the data. Additionally, the package allows for multi-processor computation, greatly increasing speed. Two of the four methods used are discussed in section 4.3.3. The number of topics does not necessarily need to be the same for the positive and negative reviews. For instance, *construction* could be a topic in negative parts of reviews, but would never appear in positive parts of reviews.

From manual observation and the average review length (18 words), reviews appear to usually contain one or two aspects (that would desirably translate into topics). For this reason, only low values of the β hyperparameter are considered, concentrating the topic distribution of reviews towards a few specific topics.

The hyperparameter β is selected automatically in accordance with the selected hyperparameter α in *topicmodels*, so an α that results in a low β (i.e. a very specific topic distribution per review) is selected, which happens at small values of α.

4.3.3 – Finding optimal topic number using *ldatuning*

Ldatuning contains multiple methods that indicate an appropriate K for LDA models. Two of these are discussed and applied.

The method proposed by Cao, Xia, Li, Zhang & Tang (2008) aims to minimize intra-cluster differences and maximize inter-cluster differences, based on topic correlation. When words belong to many topics, the score of the models' performance score should be penalized. This counteracts having too few topics in your model. Figure 8 shows such a case (K = 2). W's indicate words, Z's topics and the lines and numbers to which topics words belong (proportions). Topics Z'_1 and Z'_2 overlap in three out of five words, and the dependence degrees of words W_1 and W_2 on the topics are almost the same, resulting in little discrimination between the topics.

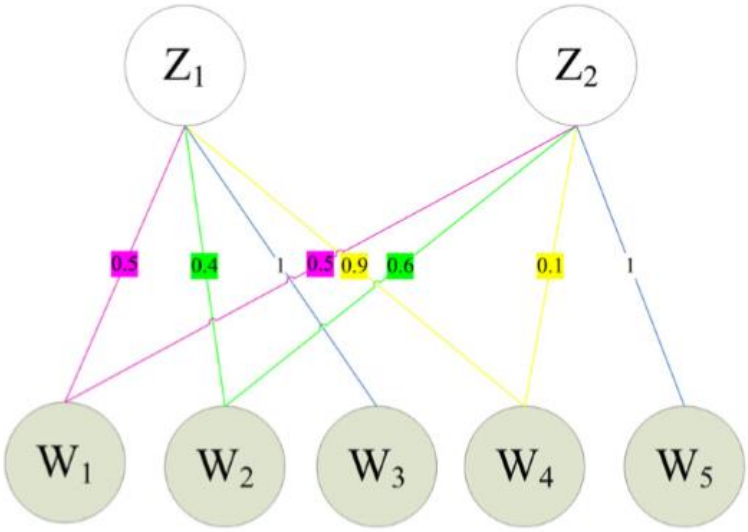


Figure 8. LDA model with too few topics (source: *A density-based method for adaptive LDA model selection*, Cao, Xia, Li, Zhang & Tang, 2008)

On the other hand, models' scores should get penalized when the correlation between topics is large. An example is displayed in figure 9 (K = 4); if Z'_3 is removed from the model, word 4, which is the only word in Z'_3, moves to Z'_2, indicating strong correlation between Z'_2 and Z'_3.

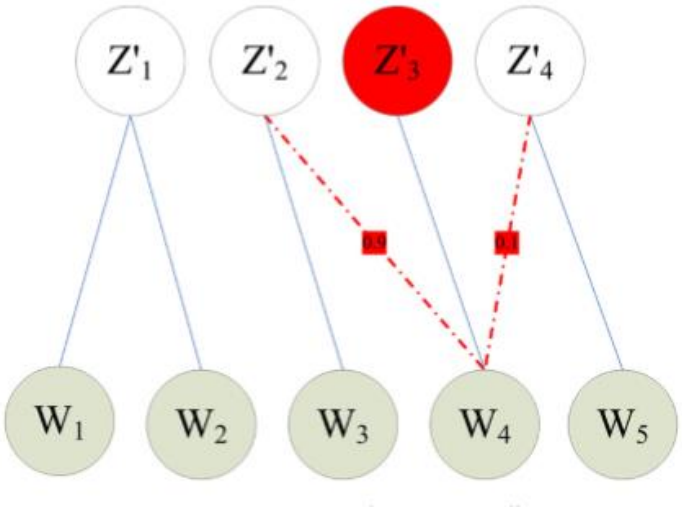


Figure 9. LDA model with too many topics (source: *A density-based method for adaptive LDA model selection*, Cao, Xia, Li, Zhang & Tang, 2009)

The *cosine similarity* between the documents is used to describe the correlation between the topics:

$$Corr(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}}$$

A simple example can make this clear: Suppose there are two topics (Item 1 & Item 2), and two words (A & B). Axis X_1 in figure 10 represents the probability of drawing word A from a topic's word distribution, and $X_2$the probability for word B. The topics are projected into this two-dimensional space. The cosine of the angle between the two topics (θ) is interpreted as the correlation between the topics. Most texts contain more than two words, so the number of axes will be equal to the total number of words (n), and the topics are projected in n-dimensional space.

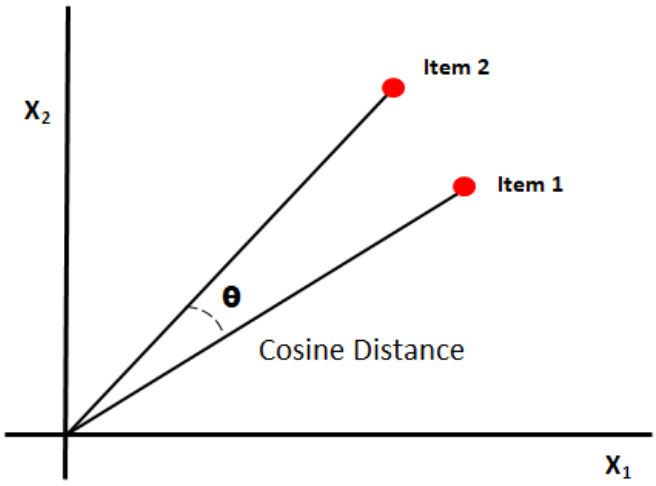


Figure 10. Cosine Distance (source: Oreilly.com)

The final metric used in the determination of the optimal number of topics (K) is the *average cosine distance* between all pairs of topics:

$$Average\ distance = \frac{\sum_{i=0}^{K}\sum_{j=i+1}^{K} Corr(T_i, T_j)}{K * (K-1)/2}$$

A low *Average distance* implies a good number of topics.

Deveaud, Sanjuan & Bellot (2014) follow a similar logic; the optimization score of a model is calculated by the formula below:

$$\frac{1}{K(K-1)} \sum_{(k,k')\epsilon T_k} D(k||k') = Average\ divergence$$

The average divergence is calculated by adding the divergences for each topic pair, and dividing by the total number of pairs. $D(k||k')$ is the Jensen-Shannon divergence (Lin J. , 1991), which indicates differences between topics:

$$D(k||k') = \frac{1}{2} \sum_{w\epsilon W_k \cap W_{k'}} P_{TM}(w|k)\ log \frac{P_{TM}(w|k)}{P_{TM}(w|k')} + \frac{1}{2} \sum_{w\epsilon W_k \cap W_{k'}} P_{TM}(w|k')\ log \frac{P_{TM}(w|k')}{P_{TM}(w|k)}$$

$P_{TM}(w|k)$ and $P_{TM}(w|k')$ are the probabilities of drawing word *w* from topic *k* and topic *k'*.

4.4 – JST

4.4.1 – Joint Sentiment/Topic Model (JST)

The Joint Sentiment/Topic Model is an extended version of LDA developed by Lin & He (2009). As discussed, standard LDA has hierarchical layers; topics are related to documents, and words are related to topics. The Joint Sentiment/Topic Model proposes adding an additional sentiment layer. In JST, topics are associated with sentiment labels, and words are both associated with topics and sentiment labels. In other words, JST assumes the following generative process for words in documents:

1) A sentiment label is selected from a document specific sentiment label distribution.
2) A topic is selected from the topic distribution. The topic distribution is conditional on the selected sentiment label in step 1; where LDA only has one topic distribution per document, JST has a number of topic distributions per document equal to the number of sentiment labels defined by the user. The number of sentiment layers is three by default; *positive/neutral/negative*. However, this number can be adjusted by the user.
3) A word is selected from the final sentiment/topic distribution.

Like with LDA, hyperparameters are set prior to running the model: α, β and γ. A visual comparison of the generative processes of LDA (a) and JST (b) are displayed in figure 11. In the JST model, γ defines the sentiment distribution (l), α defines the topic distribution (θ, conditional to l), and β defines the distribution of words over topics (φ), again conditional on sentiment.

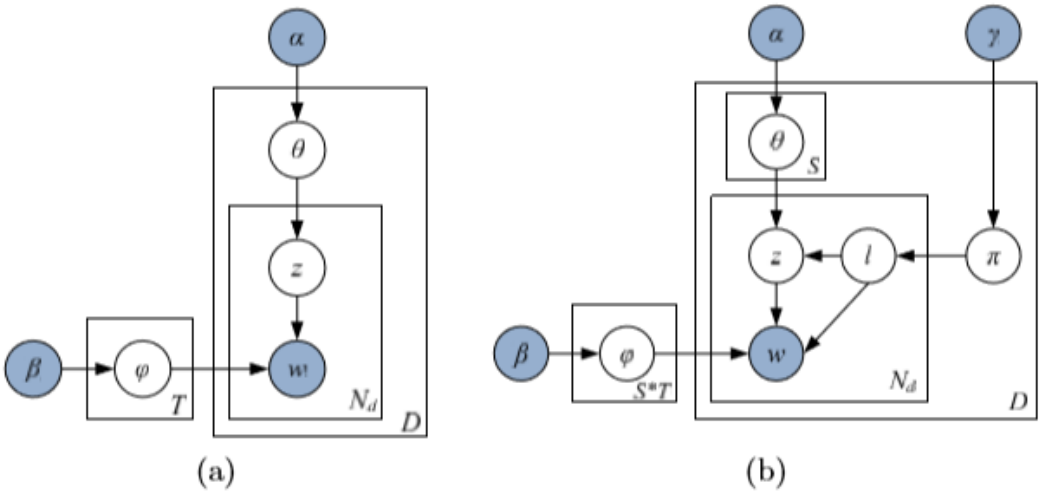


Figure 11. LDA (a) and JST (b) generative processes (source: *Joint Sentiment/Topic Model for Sentiment Analysis.* Lin & He, 2009)

The sentiment labels of JST are defined by distributions of sentiment words; these are provided to the model by a sentiment dictionary/lexicon. As discussed in section 2., words can have positive or negative sentiment depending on the context, which severely impacts the performance of sentiment analysis/JST. For example, *unpredictable* is positive in the context of movies, but *unpredictable* driving can be seen as negative. Field specific and custom dictionaries consider these differences in sentiment depending on the context of the text.

4.4.2 – Application of JST

The JST input variables are created similarly to the LDA variables described in section 4.2.2, with a few key differences. γ and β are adjusted manually to obtain:

1) Interpretable topics
2) Few topics per document

In the used R package (*rJST)* α is updated automatically every N iterations in the Gibbs-sampling process (the frequency and total iterations in the Gibbs sampling process are determined by the user (Lin, He, Everson, & Ruger, 2012). As (to my

knowledge) no standard method(s) exist to determine the optimal number of topics for JST, I will run the model using 10, 15, and 20 topics.

For the purpose of this research, I use a custom dictionary for the different sentiment levels in the Joint Sentiment Topic model, as sentiment tends to be domain-specific. To construct the sentiment levels, I gave a list of 1000 most frequently occurring words in both the positive and negative sides of reviews (2000 words in total) to two persons and myself. They were asked to categorize words into *"positive"* and *"very positive"* (for the 1000 most occurring words in positive reviews), and *"negative"* and *"very negative"* (1000 most occurring words in negative reviews). If two out of three persons put the word in a specific category, it was labeled as such. This dictionary is displayed in figure 7 in Appendix B.

4.5 – Predicting the rating

A number of regression methods exist for the prediction of continuous and binary response variables. In this research I use the Tobit model. Linear Regression and the Probit Model are shortly explained, as the Probit model relates to the *censored* parts of the Tobit model (unobserved values beyond a maximum or minimum set value, that take the maximum or minimum value). Linear Regression relates to the *uncensored* part (regular observed values).

4.5.1 – Linear Regression

Linear Regression describes a relationship between independent variable(s) X_i and dependent variable Y. The equation $Y = \alpha + \beta_i X_i + \varepsilon$ represents the best fit line, where α is the intercept, β_i the coefficient of variable X_i, and ε the error term. One of the most used methods to obtain the best fitting line is the *Ordinary Least Squares* method, which minimizes squared differences between actual and predicted $Y's$.

Linear Regression is strict model: It is very susceptible to outliers, which can significantly affect the predicted Y and predictor coefficients. Multicollinearity, an event in which on predictor can be linearly predicted by others, increases the variance of coefficient estimates.

4.5.2 – Probit Regression

The Probit model is a useful tool when the outcome variable is binary, e.g. takes values of 0 and 1. The Probit function has the following form:

$$\Pr(Y = 1|X) = \Phi(X^T\beta)$$

X is the vector of regressors in the model, and Φ stands for the *cumulative distribution function* of the standard normal distribution, which gives the probabilities for randomly selected X's to be below each value of X. The normal CDF is shown in figure 12.

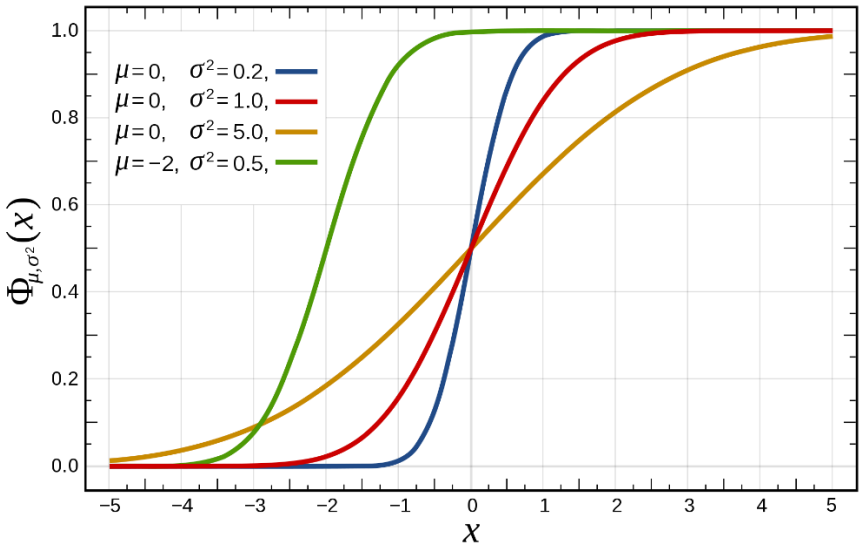


Figure 12. Cumulative Distribution Function (source: Wikipedia)

This function limits the probability outcomes of the model between 0 and 1, which is the desired effect for a binary dependent variable. The β's in the model are estimated using maximum likelihood estimation, which is the maximization of:

$$LogL(\beta; Y, X) = \sum_{i=1}^{n} (y_i log\Phi(x'_i\beta) + (1 - y_i)\log(1 - \Phi(x'_i\beta))$$

4.5.3 – Tobit Regression

Tobit Regression was first introduced by James Tobin (1958). The model assumes that the dependent variable has an upper or lower limit (or both). In his paper, Tobin states that Multiple Linear Regression could be used to model the relationship between the regressors and the (censored) dependent variable, but that this would not be a good representation of the actual relationship. On the other hand, the Probit model could be used to calculate the probabilities of the dependent variable to be under the lower limit (or above the upper limit), but in such a model all information available about the uncensored observations is lost. The Probit model proposed by Tobin is, in essence, a linear regression model with elements of the Probit model.

The user-rating variable in this research is a good example of a censored variable, with 10 being the upper limit. Not all 10's are 'equal'; some reviewers give their hotel stay

the rating 10, but still argue what was bad/could have been better, and some reviewers state that *"everything was perfect"*, and provide no downsides in their reviews. This implies that some users would like to rate the hotel stay above 10, but cannot because of the upper limit of the rating scale. Similarly, some reviewers providing the rating 1 still provide some positive feedback, while others write something along the lines of *"nothing was good/everything was terrible/etc."*, implying that they would like to rate the stay lower than 1. The Tobit model only has added value when dense concentrations of observations at the limiting value(s) are present.

An excellent visual example of the goal of the Tobit model is provided in a lecture from the Economics Department of the University of Copenhagen, shown in figure 13 (Munk-Nielsen, 2016). In this example, a lower limit of $y = 0$ exists in the data. The aim of the model is to construct a regression line that accounts for theoretical (or latent) values of y below the lower limit (the orange observations).

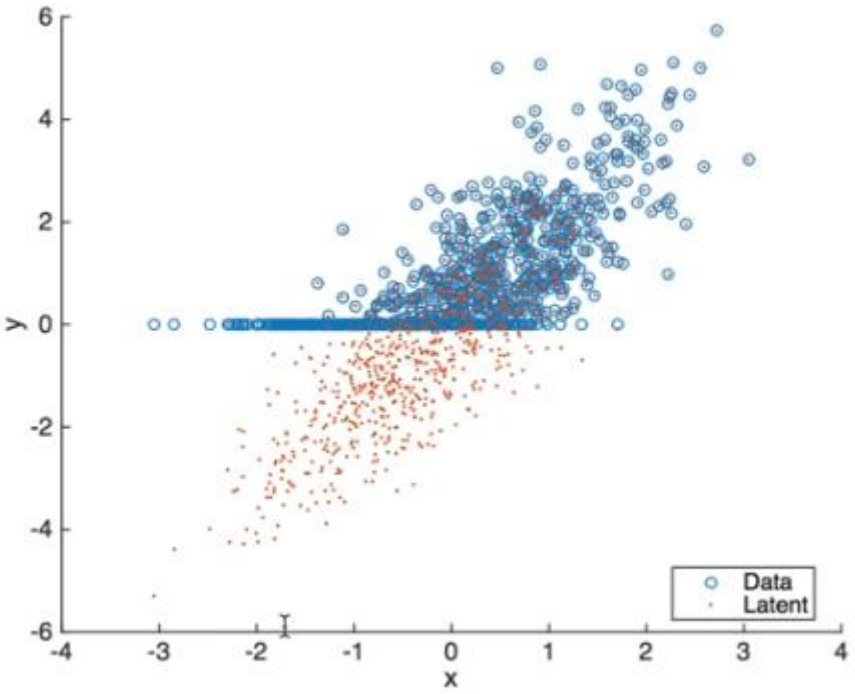


Figure 13. Tobit model example (source: *Maximum Likelihood: The Tobit Model*, Munk-Nielsen, A., 2016)

With an upper limit and lower limit, the uncensored $y's$ are given by:

$$y_i = x'_i\beta + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

But when y_i is greater than the upper limit *b*, *b* is the observed value:

$$w_i = \min\{y_i, b\}$$

Similarly, when y_i is smaller than the lower limit *a*, limit *a* is observed:"

$$w_i = \max\{y_i, a\}$$

The obtain values for β and σ, the *loglikelihood* function is maximized. It can be decomposed into three parts: the likelihood for observations *i* above the upper limit:

$$P(y_i > a) = 1 - \Phi\left(\frac{a - x_i^T\beta}{\sigma}\right)$$

The likelihood for observations *i* below the lower limit:

$$P(y_i \leq b) = \Phi\left(\frac{b - x_i^T\beta}{\sigma}\right)$$

And the likelihood for observations *i* of the uncensored part of the data:

$$f(y_i; x_i^T\beta, \sigma) = \frac{1}{\sigma}\varphi\left(\frac{y_i - x_i^T\beta}{\sigma}\right)$$

Φ is the cumulative distribution function of the normal distribution, and φ the probability density function of the normal distribution. The log-likelihood function adds up the logarithms of all observations across the three categories:

$$logL(\beta, \sigma) = \sum_{i\,\epsilon\,nc} log\left(\frac{1}{\sigma}\Phi\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\right) + \sum_{i\,\epsilon\,tc} log\left(1 - \Phi\left(\frac{a - x_i^T\beta}{\sigma}\right)\right) + \sum_{i\,\epsilon\,bc} log\left(\Phi\left(\frac{b - x_i^T\beta}{\sigma}\right)\right)$$

With *nc = not censored, tc = top censored,* and *bc = bottom censored.*

4.6 – Regression with LDA & JST input variables

Using the obtained LDA *or* JST (senti-)topics as predictor variables, and the review rating as the dependent variable, a Tobit model is fitted on the whole dataset. Censoring is applied at reviewer rating 10, one of the limiting values, and the only value where observations are densely concentrated. High topic coefficients indicate a strong effect on the reviewer rating, whereas low coefficients suggest a weak/no effect.

4.7 – MBRP

A general model predicting review ratings using ratings provided by all reviewers may not be a good representation of the relationship; certain groups of people probably value different aspects/topics differently than others. To find groups with distinct preferences/rating functions, *Model Based Recursive Partitioning* (Zeileis, Hothorn, & Hornik, Model-based Recursive Partitioning, 2008) is used. This method is based on decision trees, so a short introduction is provided in the following section.

4.7.1 – Decision Trees

Decision trees are supervised learning methods used for regression and classification. Using available data features to create simple decision rules, the models predict a target variable's value. The model starts off including the whole provided dataset; this first *node* is referred to as the root node. Based on some condition, the data is subsequently split into two partitions, or leaves, which should yield a better prediction for the target variable. The split is a simple decision rule. For example: in a model predicting income this could be $AGE > 40$ → leaf 1, $AGE < 40$ → leaf 2, or in case of a categorical variable *HIGHLY EDUCATED* → leaf 1, NOT EDUCATED & SOMEWHAT EDUCATED → leaf 2. In the leaves the same process is repeated, until some stop condition is met. These conditions can be;

- The number of observations in the terminal nodes need to be more than a set number.
- The *depth* (length of the path from root- to terminal node) reaches a maximum.
- The *purity* (similarity between observations' target variable) of a node is more than the limit for splits.

Alternatively, a full-grown tree can be *pruned:* terminal nodes are removed based on certain conditions.

When using the model to predict the target variable value of a new observation, the observation starts off in the root node. Following the path down the tree, the observation is assigned to leaves based on the existing decision rules. In the terminal node, the target variable value of the observation can be predicted. For regression trees, this is simply the average of the target variable values of observations (used to train the model) in that node. In classification trees, it is the majority category represented in the node (by observations used to train the model).

Decision trees are *greedy;* A good split at a node is the local optimum, but may not be the global optimum.

4.7.2 – Model Based Recursive Partitioning

Model Based Recursive Partitioning is a variation of the decision tree; each leaf in the tree yields a full Regression model. This (Generalized Linear) Regression model is specified by the user prior to the training of the Recursive Partitioning Model. Two sets of variables need to be provided; the *predictor set* and the *partitioning set.* Variables in the predictor set are used to create the Generalized Linear Model in each leaf. The variables in the *partitioning set* are tested for significant parameter instability. When instabilities in the parameters are observed, the data is split. An *instability* implies that the model is too simple, and the relationship between the predictors and dependent variable would be better described by (two) separate models. Measuring instability is explained in more detail by Andrews, Zeileis, Hothorn & Hornik (2007, 2008).

An example of a two-leaf model (with one of the coefficients shown) is displayed in figure 14.

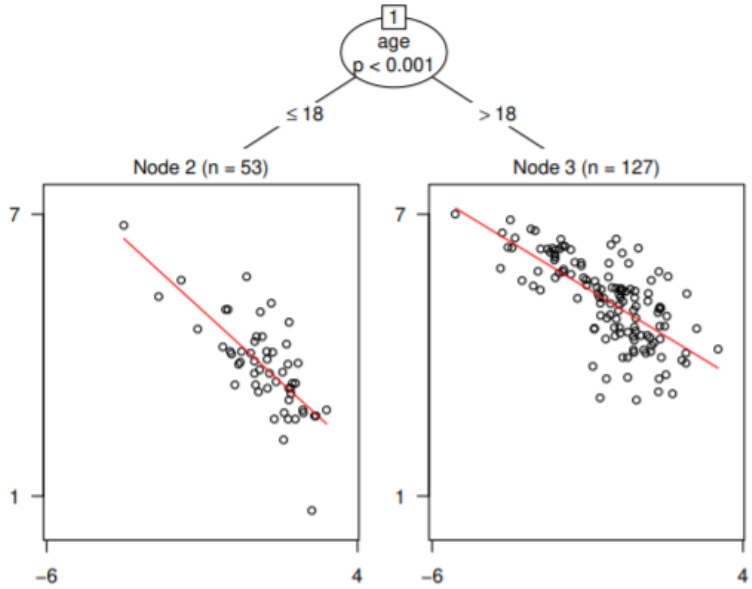


Figure 14. Linear Regression based tree from journals data *(source: Parties, Models and Mobsters: New implementation of Model-Based Recursive Partitioning in R, Zeileis & Hothorn, 2015)*

4.7.3 – Model Based Recursive Partitioning applied

The same Tobit regression used on the whole dataset in section 4.6 is used for the purpose of Model Based Recursive Partitioning.

As described in section 4.7.2, two sets of variables need to be defined; the *partitioning set* and the *predictor set.* The predictor variables of interest are the Negative and Positive LDA and JST topics, analogous to section 4.3.5. For the purpose of this research I am interested in segmenting users and their respective rating-generating-

processes based on characteristics observable *prior* to their stay; hotel management should be able to observe a booking order, and tailor the stay to the user's preference.

The variables matching these criteria are; *Room type, Bed type, Guest type, Trip type, Nationality* and *Stay duration, Average hotel rating,* and *Businesses in 100m/1km/5km.* The addition of *Quarter of the Year* is debatable, as this variable indicates when the review was written, not when the trip took place (which becomes even more problematic for shorter timeframes like *Month of the year* or *Week of the year).* I will make the simplifying assumption that the reviews were written in the same quarter of the year as the trip, and include it in the regression. *Nationality* and *Room type* are omitted, because the model cannot processes categorical variables with that many levels. *Average hotel rating, Businesses in 100m,* and *Businesses in 1km* are omitted, because they crowd out the other, more interesting partitioning variables (The total number of terminal leaves increases more than tenfold, and other partitioning variables move to the bottom of the tree). A solution to this would be converting them categorical variables with two or three levels, thereby limiting splits. However, I have decided not to do this in favor of the partitioning variables directly related to the trip (as opposed to *hotel variables* like *Average hotel rating, Businesses in 100m,* and *Businesses in 1km,* which are fixed per hotel*).*

Reviews where *Trip type* equals anything other than *Business trip* and *Leisure trip* are omitted from MBRP. These are duplicates of the categories found in the *Guest type* variable: *Couple, Family with older children, Family with young children, Solo traveler, with a pet (*and *NULL).* Leaving them in the dataset results in strange, uninterpretable splits in the tree. Figure 1 in Appendix D shows a partitioning tree where these variables are included. Groups [1,2,3] include Trip types *Couple, Family with young children* and *Solo traveler* based on the first split. However, a split is made further in the tree (Groups [4,5,6] and [7,8]) based on the variable *Guest type,* which also includes categories *Couple, Family with young children* and *Solo traveler.* Intuitively this makes no sense, and it is not clear from the data what the differences between these categories in *Trip type* and *Guest type* are. The omitted observations form only 3% of the data, so relatively little information is lost.

The results obtained from MBRP tie the research together, and enable me to answer the research question; the resulting regression models in each terminal leaf can be analyzed to obtain important feature rating determinants for each segment.

5 – Results

5.1 – LDA topic results

The optimal number of topics, derived using the methods proposed by Cao, Xia, Li, Zhang & Tang (2009) and Deveaud (2014) are displayed in figure 15 (metrics for positive LDA topics parts) and figure 16 (metrics for negative LDA topics). Based on the Cao metrics (the Deveaud metric seems to be constantly decreasing), 20 positive and 20 negative topics appear to be good values for K. According to the metrics, more topics could be added, but this would produce a model with double the amount of predictors (35 topics is the second optimal amount of positive topics according to the CaoJuan metric, and 45 negative topics) and make interpretation significantly harder.

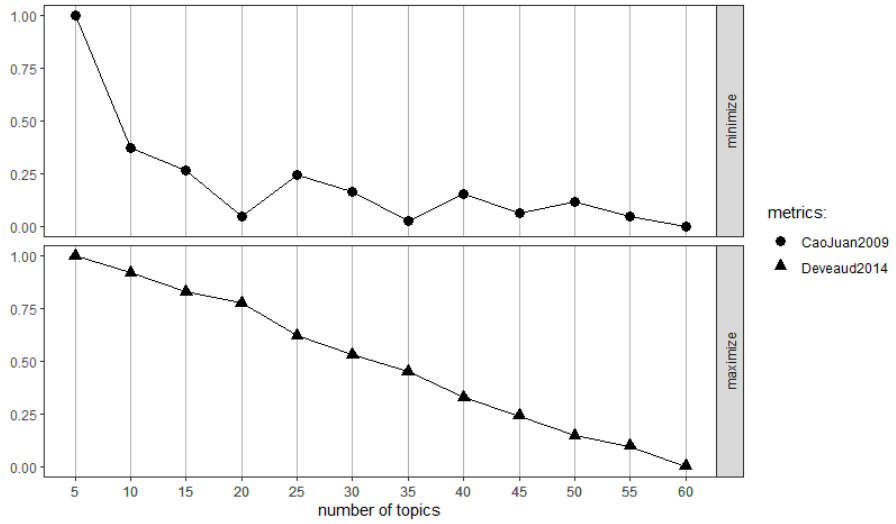


Figure 15. Optimization positive number of topics (LDA)

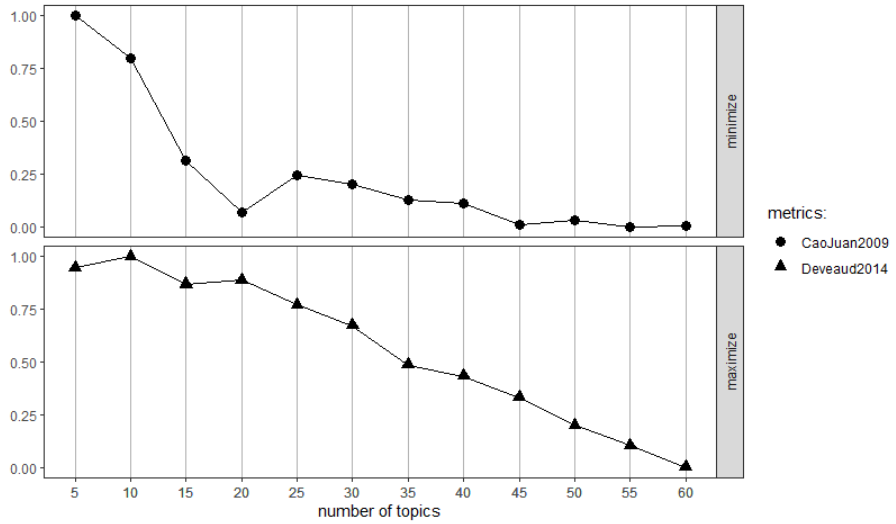


Figure 16. Optimization negative number of topics (LDA)

The resulting topics can be found in Appendix A. The numbers at the bottom of each figure indicate the probability of drawing a certain word from that topic. Some topics are easily interpretable, while others can be more ambiguous. β is automatically adjusted to fit the data by *lda* function from the topicmodels package, according to the α selected; a large value for hyperparameter α results in a low β and vice versa. When the selected α is too large documents consist of too many topics, and when α is too small the topics become too ambiguous for interpretation. After manual tuning, a good trade-off point between these two seems to be at $\alpha = 0.05$: The topics are (mostly) interpretable, and reviews consist of relatively few topics, which is expected after manually observing the structure of some reviews, and the average length of the reviews. Even though some topics seem to overlap (or simply seem ambiguous at first), looking into their respective wordclouds often reveals more details that are valuable for their interpretation.

The most important obtained topics are labeled below in list 1 (Positive Topics) and list 2 (Negative Topics). The "most important" topics are defined by the outcome of Model Based Recursive Partitioning, discussed in section 5.1.3.

Topic 5 appears to be about pleasant additions to the stay on special occasions, indicated by *"free"*, *"upgrade"* and *"birthday"*. Some minor words in the wordcloud in figure 17 also point at this fact: *"anniversary"*,*"surprise"*,*"celebrate"*,*"cake"* and *"balloon"*.

Negative topics 8, 9, and 12 all appear to relate to money in some way. The comparison cloud for the three topics is shown in figure 18; it presents the words with the highest frequency *relative* to frequencies in other topics. Topic 9 is focused primarily on payment problems/unexpected payments/payment policy, indicated by words such as *extra, deposit, (not) clear, advance, refund, inform, problem, mistake,* etc. *Pricey* or *expensive* are not present for topic 9 in either the top 15 words or the comparison cloud, so it does not appear to be about the price/value in itself. Topic 12 appears to relate to the value (mainly size) of the room, indicated by the strong presence of *room,* and minor words such as *price, worth, disappointed, small, smaller, tiny, little* and *photo* ("*not as big as expected from the photo*"). Topic 8's words in the comparison cloud still give little insight into the topics (specific) theme. Its wordcloud is shown in figure 19 (the information it contains is different from the comparison cloud, because the wordcloud shows the absolute frequency of words, as opposed to relative frequency to another topic). It also appears to relate to value (*price, worth, value, money).* I will treat this topic as the hotel "general worth", because *room (*or any other concrete aspect) does not hold a high position in the topic.

Positive Topics

2. Bar/restaurant, emphasis on staff (*staff, bar, restaurant, excellent, food, breakfast)*

3. Access (*tram, stop, easy, minute, walk, airport & station)*

5. Anniversary/Birthday/other occasion additions (figure 17)

7. Everything good *(best, everything, amazing, perfect)*

8. Reception staff (indicated by high position of *staff,* and *front, desk, reception, welcoming)*

14. Hotel design (*hotel, beautiful, decoration, modern, design, style*)

16. Everything good, emphasis on staff (*love, recommend, great, staff, friendly, nice)*

17. Bed *(bed, comfort, comfi)*

19. Room design (similar to positive topic 14, but instead of *hotel*, *room* appears the most in this topic)

20. Room facilities (*room, free, coffee, machine, tea, clean, water)*

List 1. Main positive topics

Negative Topics

1. Check-in problems (appears somewhat ambiguous, but indicated by *arrive, check, time, wait, night, ask, reception* → *"I asked if the receptionist could wait, so I could check in at night when I arrive"*

4. Reception staff (*front, desk, reception & rude)*

8. General value (figure 18)

9. Payment problems/unexpected payments (figure 18)

11. Room state *(room, old, dirty & smell)*

12. Room value (figure 18)

17. Housekeeping *(housekeeping, left, didn't, clean, service* & *staff)*

List 2. Main negative topics

Figure 17. Positive topic 5 wordcloud

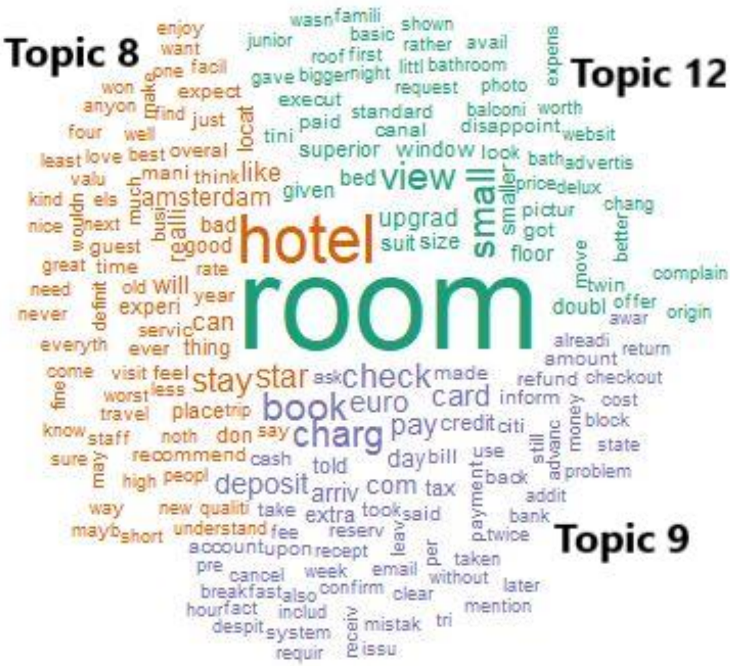


Figure 18. Comparison cloud topics 8, 9 & 10

Figure 19. Negative topic 8 wordcloud

5.2 – Tobit regression results general dataset (LDA topics)

The results for the whole sample are displayed in figure 1 in Appendix C. The intercept in the regression models in itself is meaningless; a review always contains either a positive part, a negative part, or both. The intercept implies the score of a theoretical 'empty' review. The coefficients in the regressions can therefore not be compared to a baseline, because it does not exist. They can however be used to:

1) Predict the rating (*Intercept* + *Coefficient* * *Topic*). This has limited value, as other machine learning methods like neural networks and random forest are much better equipped for the purpose of accurate prediction.
2) The coefficients can be used to compare the effects of the topics on the rating, which is the goal of this research. For example, the coefficients of topic 1 and topic 2 in the model are 0.519 and 1.175 respectfully. This implies that for a review that is fully about topic 2, the reviewer rating would on average be 0.656 higher compared to a review fully about topic 1. The coefficients can be seen as artificial dummies; a review that is *fully* about a topic contains a 1 for that (positive or negative) topic.

Observing the regression output for the whole dataset (table 1 in Appendix C), a number of things can be noted.

Overall, the signs of the coefficients make sense; most positive review topics have a positive effect on the rating, and most negative review topics have a negative effect on

the rating. There are two exceptions; positive review topic 11, and negative review topic 3. Positive topic 11 seems ambiguous from just looking at the top-15 most occurring words. The wordcloud including additional words for positive topic 11 is shown in figure 20. Even the top 100 occurring words are not really revealing anything about positive topic 11's subject. However, this ambiguity is in line with the results; the topic's coefficient is the smallest in the data (-0.146), implying it has almost no effect on the reviewer rating (relative to the theoretical "empty review")

Figure 20. Positive Topic 11 wordcloud

However, negative topic 3 has a strong coefficient of 1.856, and clearly influences the predicted rating. Looking at the top-15 words, the topic is mostly comprised of the word *"nothing"*, followed by "everything" and extremely positive polarity laden words like *"perfect"*, *"great"* and *"love"*. This implies sentences like: "*I disliked nothing; everything was perfect"*, i.e. the reviewer listed the lack of downsides of the hotel stay in the negative review section. This is logically in line with the found positive coefficient.

Another notable observation regarding the general dataset is the overall difference between positive and negative review topics' coefficients; negative topics tend to have a much stronger effect on the review ratings compared to positive review topics. Only one positive topic has a coefficient larger than 2 (positive topic 16), while 7 negative topics have a coefficient smaller than -2 (topics 1, 4, 8, 9, 11, 12 & 17).

The five strongest review rating determinants for the whole dataset are, in descending order: "Everything good, emphasis on staff" (2.56), "Anniversary/Birthday/other occasion additions" (1.89), "Everything good" (1.88), "Hotel design" (1.74) and "Room facilities" (1.36),

The strongest negative review rating determinants are: "Room state" (-2.972), "Housekeeping problems" (-2.917), "Reception staff" (-2.769), "Check-in problems" (-2.560), and "Room value" (-2.386)

5.3 – JST topic results

Unfortunately, the results obtained from the JST model do not serve the intended "purpose" of the model. The goal is to find three stages of polarity within topics; *neutral* (sent1), *positive* or *negative* (sent2), and *very positive* or *very negative* (sent3). However, the model fails to capture clear, interpretable topics. This is illustrated in figures 1-6 in Appendix B . Figures 1-3 show some positive sentiment topics when the total number of topics produced is 10/15/20 ($\beta = 0.01$ & $\gamma = 0.3$, the standard setting of *jst*), and figures 4-6 show some negative sentiment topics when the total number of topics is 10/15/20 ($\beta = 0.01$ & $\gamma = 0.3$). Most *sentiment topics* are interpretable on their own, but the three sentiment topics per topic have very little in common, and make the labeling/interpretation of topics hard. Adjusting β and γ does not alleviate this problem. Changing α to a small value does, but simultaneously makes topics indistinguishable, and therefore uninterpretable.

Because the added value of the JST model is lost on this dataset, and because LDA is a more academically proven model, only the LDA output will be used in the following sections.

5.4 – Segmentation using MBRP (LDA topics)

The tree obtained by Model Based Recursive Partitioning is shown in figure 2 in Appendix D. As described in the methodology section, the variables used for partitioning are, *Bed type, Guest type, Trip type, Stay duration,* and *Quarter of the year.* It is *fully-grown.* No instabilities in the parameters exist in the terminal nodes of the tree; the relationship between the topics and the reviewer rating cannot be explained any better by dividing the reviewers further based on partitioning variables/adding more regression models. On first glance the tree is dominated by *Bed type, Guest type, Trip type* and *Stay duration. Quarter of the year* is found only in one split at the bottom near two terminal leaves. This (possibly) reflects one of the limitations of decision trees; continuous variables and categorical variables with many levels have a higher chance of being selected for splitting.

The guest categories obtained from following every path down the tree in Appendix D, figure 2 are shown in list 3 below:

Guest categories

1. **Business trip; Guest type**: Family with older children, Couple, Solo traveler
2. **Business trip; Guest type:** Family with young children, Group, Travelers with friends, With a pet
3. **Leisure trip; Bed:** King-sized bed
4. **Leisure trip; Guest type:** Couple, Family with older children, Group; **Bed:** Single, Double, Triple, Queen, Unknown;
   Stay duration <= 2
5. **Leisure trip; Guest type:** Couple, Family with older children, Group; **Bed:** Single, Double, Triple, Queen, Unknown;
   2 < **Stay duration** <= 20
6. **Leisure trip; Guest type:** Couple, Family with older children, Group; **Bed:** Single, Double, Triple, Queen, Unknown;
   20 < **Stay duration** <= 29
7. **Leisure trip; Guest type:** Couple, Family with older children, Group; **Bed:** Single, Double, Triple, Queen, Unknown;
   29 < **Stay duration**
8. **Leisure trip; Guest type:** Family with young children, Travelers with friends, Solo traveler, With a pet ; Bed: Single, Double, Triple, Queen, Unknown;
   Quarter of the year: 1
9. **Leisure trip; Guest type:** Family with young children, Travelers with friends, Solo traveler, With a pet ; **Bed:** Single, Double, Triple, Queen, Unknown;
   Quarter of the year: 2, 3, 4
10. **Leisure trip; Bed:** Twin

List 3. Model Based Recursive Partitioning categories

The results for each segment can be found in the tables in Appendix E. Table 2 shows a more summarized version of this information:

	Cat 1	***Cat 2***	***Cat 3***	***Cat 4***	***Cat 5***	***Cat 6***	***Cat 7***	***Cat 8***	***Cat 9***	***Cat 10***
Pos 1	16 (2.10)	7 (3.21)	16 (2.95)	16 (2.71)	16 (2.61)	16 (2.37)	5 (2.62)	16 (2.67)	16 (2.34)	16 (2.38)
Pos 2	14 (1.76)	16 (2.32)	7 (2.94)	7 (2.31)	7 (2.14)	5 (1.82)	16 (2.60)	7 (2.48)	14 (1.80)	5 (2.01)
Pos 3	5 (1.64)	14 (2.28)	5 (2.20)	5 (1.69)	5 (1.84)	14 (1.55)	7 (2.10)	14 (1.72)	5 (1.79)	14 (1.72)
Pos 4	7 (1.39)	2 (1.74)	14 (2.20)	14 (1.61)	14 (1.71)	7 (1.54)	14 (1.92)**	2 (1.41)	7 (1.78)	8 (1.64)
Pos 5	19 (1.38)	5 (1.57)*	3 (1.66)	20 (1.53)	20 (1.34)	20 (1.39)	8 (1.79)	17 (1.41)	19 (1.43)	20 (1.36)
Neg 1	1 (-3.11)	11 (-3.95)	9 (-2.77)	11 (-3.29)	11 (-3.08)	17 (-2.85)	11 (-3.95)	4 (-2.83)	17 (-2.99)	17 (-3.40)
Neg 2	11 (-3.06)	4 (-3.43)	1 (-2.48)	17 (-2.69)	17 (-3.07)	11 (-2.70)	4 (-3.87)	11 (-2.77)	4 (-2.79)	11 (-2.88)
Neg 3	17 (-2.96)	17 (-3.38)	4 (-2.46)	12 (-2.58)	4 (-2.91)	4 (-2.65)	1 (-3.76)	8 (-2.24)	11 (-2.79)	1 (-2.71)
Neg 4	12 (-2.94)	8 (-3.22)	17 (-2.36)	4 (-2.57)	1 (-2.79)	1 (-2.40)	17 (-3.43)	12 (-2.19)	8 (-2.70)	8 (-2.69)
Neg 5	4 (-2.60)	9 (-2.98)	11 (-2.19)	1 (-2.21)	12 (-2.22)	12 (-2.35)	12 (-3.39)	9 (-2.05)	1 (-2.39)	4 (-2.57)

* P < 0.05 ** P < 0.1 (all other coefficients at least P < 0.01)

Table 2. Top regression coefficients

The categories (columns) in the table refer to the different segmentation categories from the Model Based Recursive Partitioning tree in Appendix E. The first five rows indicate the top five largest positive topic coefficients for each category in descending order. The bottom five rows indicate the top largest negative topic coefficients. The first number in the inner cells of the table is the topic number, and the number below it refers to the size of the topic's coefficient.

Certain facts can be readily observed from the data in table 2: Similarly to the regression for the whole dataset, negative coefficients are mostly larger than positive coefficients. The only exception to this are the first few topic coefficients of category 3; the leisure trip with a king sized bed. There is a little more variety in the positive top topics present in the top 5 compared to the negative topics: Nine out of twenty positive topics appear in the top five of at least one of the categories (topics 2, 3, 5, 7, 8, 14, 16, 19 & 20), whereas only seven distinct negative topics appear in the top 5 of at least one of the categories (topics 1, 4, 8, 9, 11, 12 & 17). However, the top negative topic coefficients are more varied: Eight out of ten *Pos 1* positions (Highest positive coefficients in categories) are occupied by topic 16, whereas the most occurring topic in *Neg 1* only occupies 4 spots (negative topic 11).

Figures 21 and 22 display the coefficient strengths of topics that appear at least once in summary table 2 in a visually easier interpretable way. Each corner of the decagon represents a guest category, and each point shows the respective coefficient strength of a topic, indicated by the distance between the point and the center of the circle. Statistically insignificant coefficients ($p>0.05$) are omitted. The spread of the points per category clearly illustrates the hierarchy of hotel feature preferences.

Figure 21. Positive topics (occur in top 5)

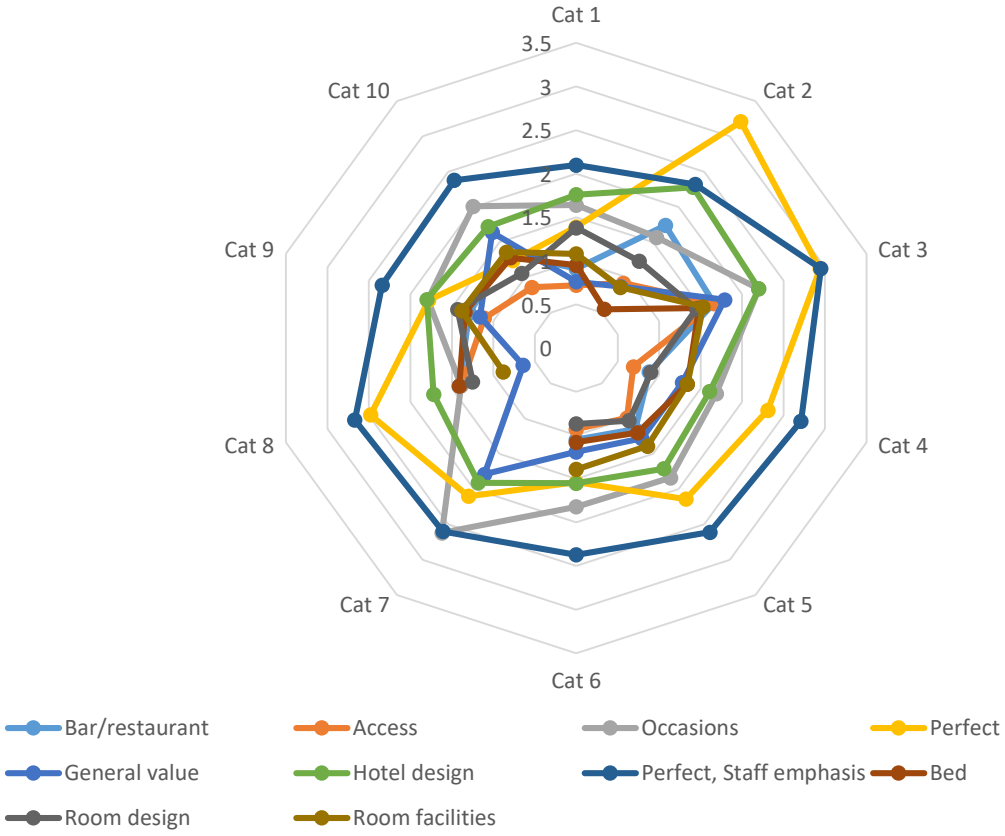
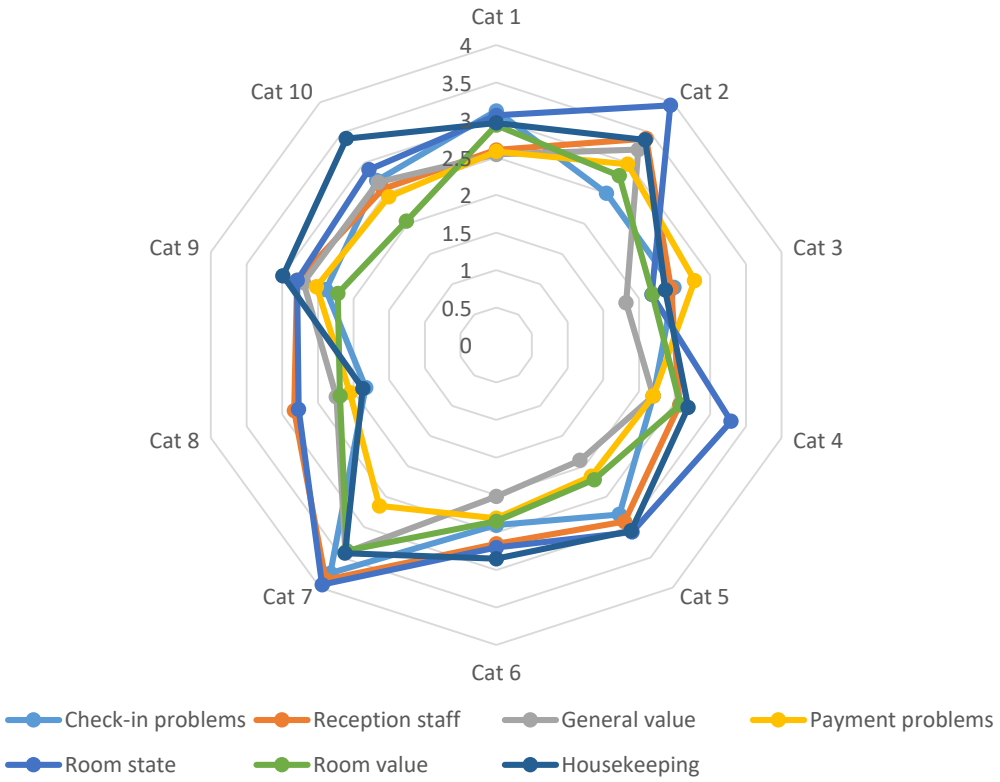


Figure 22. Negative topics (occur in top 5)



While some of the topics obtained from LDA are ambiguous and do not correspond to specific features, all of the topics in figures 22 and 23, defined as "important topics", could be labeled to at least some degree.

6 – Conclusion, Discussion, & Limitations

With the available results, I can now answer the question central to this research:

" On what characteristics can hotel guests be segmented to reveal differences in their hotel-feature preferences, and what are these preferences?"

All of the partitioning variables provided to the model are used; *Trip type, Guest type, Bed type, Stay duration,* and *Quarter of the year*. This answers the first part of the research question; segmentation can be done on any of the variables/characteristics, as they all appear at some point in the Recursive Partitioning algorithm. However, some hierarchy in the partitioning variables can be observed; stay duration and quarter of the year only appear at the bottom of the tree. For quarter of the year this is somewhat expected, as decision trees prefer continuous variables and variables with a high number of levels, and quarter of the year only has four levels. Duration is a continuous variable, so this limitation should not apply to it; we can state that duration is most likely less useful for the segmentation of hotel customers based on their preferences. The scope of this research was limited to segmentation variables based on easily

observable consumer information and trip time details (*Quarter of the year* & *Stay duration*), but future research does not necessarily need to be. The list of partitioning variables could easily be extended to include other details like the average hotel rating (do certain categories of guests have different expectations related to hotel aspects when the rating is high/low?), number of reviews given by the reviewer (does a more seasoned traveler rate hotels differently?), etc. depending on the goal the research.

Figures 21 and 22 in the results section (5.4) give some clear insights into the rating generating processes/hotel-feature preferences of reviewers. The information in the figures can be summarized in words, although somewhat limited, as follows:

- **Category 1** guests value hotel design and special occasion additions, followed by room design to a lesser degree. Check-in problems, bad room value, a bad room state and housekeeping problems are disliked equally. Rude reception staff, payment problems and general value, all equally disliked, follow these features.
- **Category 2** guests value hotel design, followed by a nice bar/restaurant (with emphasis on staff) and occasion additions (equally). They dislike rooms in bad states the most, followed by rude reception staff and housekeeping problems.
- **Category 3** guests appreciate hotel design and occasion additions. They dislike payment problems.
- **Category 4** guests value hotel design and occasion additions the most, followed by general value, a comfortable bed, and room facilities. They dislike a bad room state by far the most.
- **Category 5** guests value occasion additions, followed closely by hotel design. They dislike a bad room state, housekeeping problems, rude reception, and check-in problems (in descending order, but very closely).
- **Category 6** guests value occasion additions, room design, and room facilities in descending order. They dislike housekeeping problems the most, followed closely by rude reception staff and room state (equally).
- **Category 7** guests appreciate occasion additions the most. They dislike a bad room state, rude reception staff, and check-in problems in descending order (but very closely).
- **Category 8** guests value hotel design, followed by a comfortable bed, good access, and room design. The dislike housekeeping problems and rude reception (much more than other aspects).
- **Category 9** guests enjoy hotel design and occasion additions, and dislike housekeeping problems, followed closely by rude receptions staff and room state (equally).

- **Category 10** guests value occasion additions, followed by hotel design and general value (equally). They dislike housekeeping issues by far the most.

The coefficients of positive topic 7 ("Everything good") can appear confusing; on a number of occasions it is lower than other coefficients, which is intuitively not possible (everything good > something good). This could possibly be caused by the distinct lack of the word "staff", implicitly implying that service is bad/not as good as other topics' reviewers' experiences. A concrete explanation is beyond the scope of this research.

On a more general note, the interpretation of topics is subjective. I label the topics based on a number of important keywords occurring in the topics, but all topics consist of a mixture of *all words* in the corpus. In order to gain a perfect understanding of the topic coefficients' meaning, in-depth analyses of *every* topic's word consistency (and the differences in word consistencies between all moderately related topics) are needed.

The use of a Tobit regression is debatable; although it is a clear improvement over standard Ordinary Least Squares, it assumes a Gaussian distribution of the data. The occurrences of the ratings in the data are strictly increasing, and it is doubtful that when including hypothetical values past rating 10 the distribution would look like the normal bell-curve.

For possible future research, I recommend further experimentation with combining topic models and regressions. In this research I have selected the number of topics through the use of topic number optimization methods, and hyperparameters α and β manually for easy interpretation. Regression followed as a "second step" after topic modelling. A comprehensive overview of how the different hyperparameters directly influence regression performance and results could be interesting.

Appendix A: LDA topics

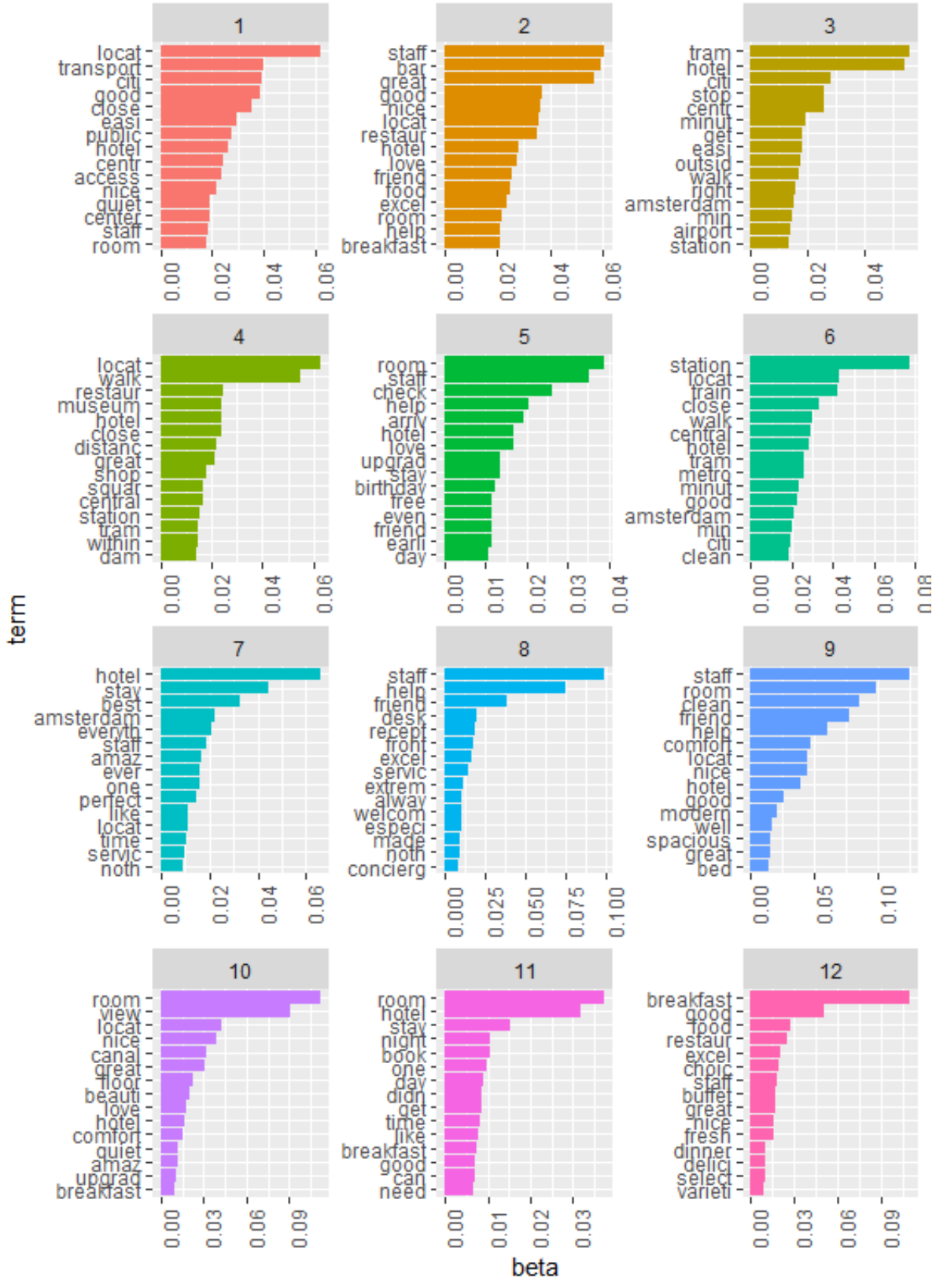


Figure 1. Positive LDA topics (topics 1 – 12)

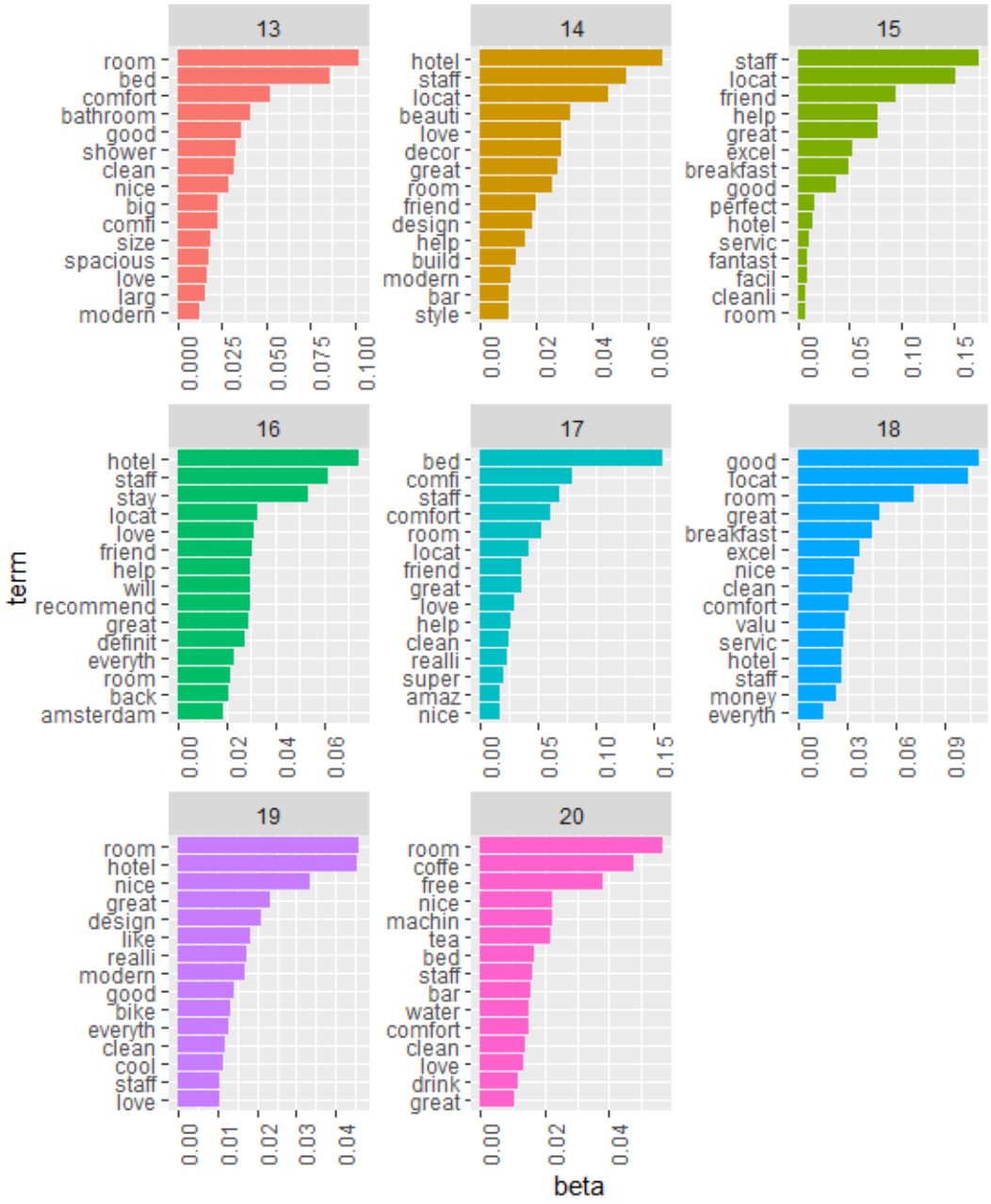


Figure 2. Positive LDA topics (topics 13 – 20)

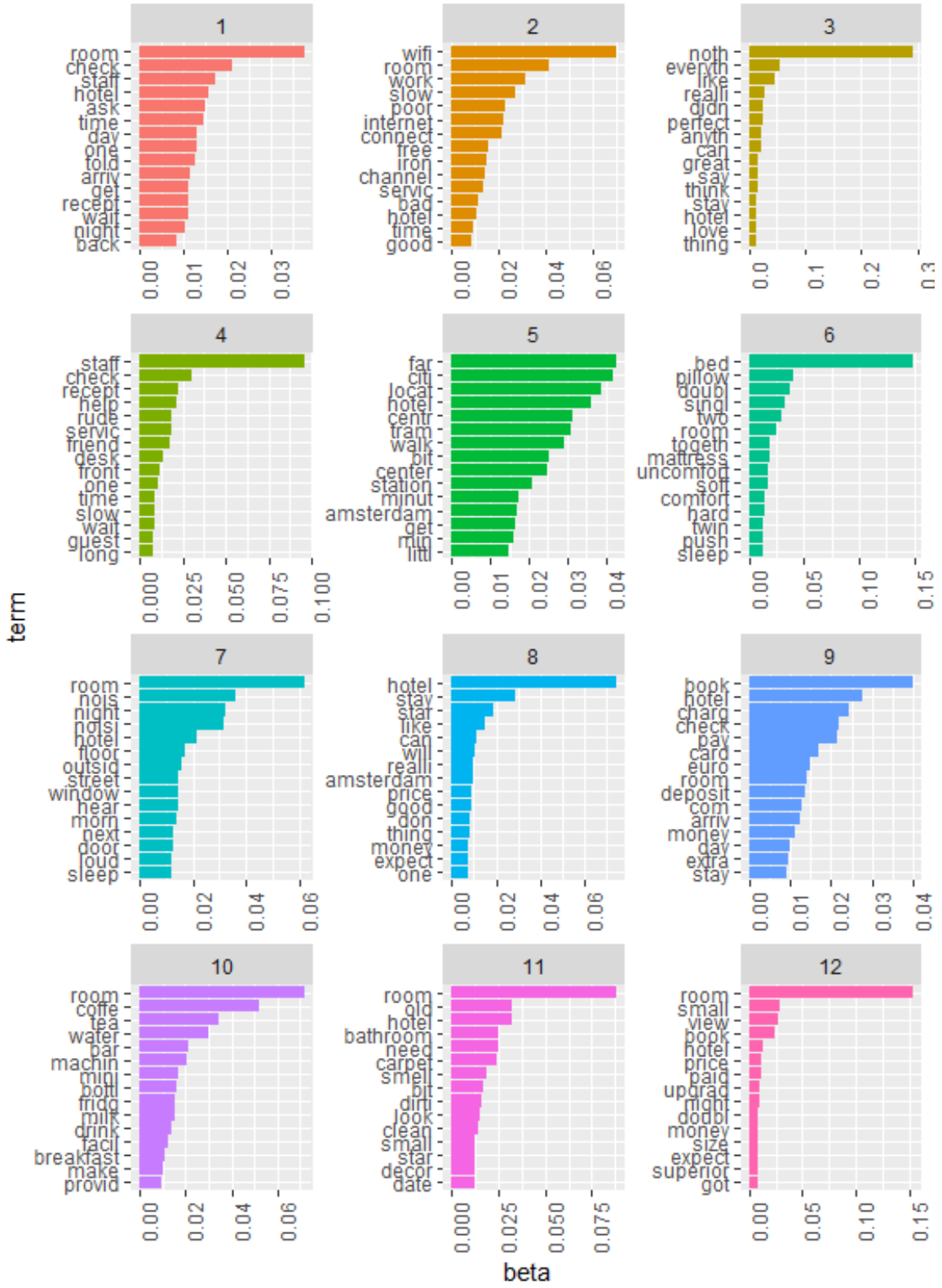


Figure 3. Negative LDA topics (topics 1 – 12)

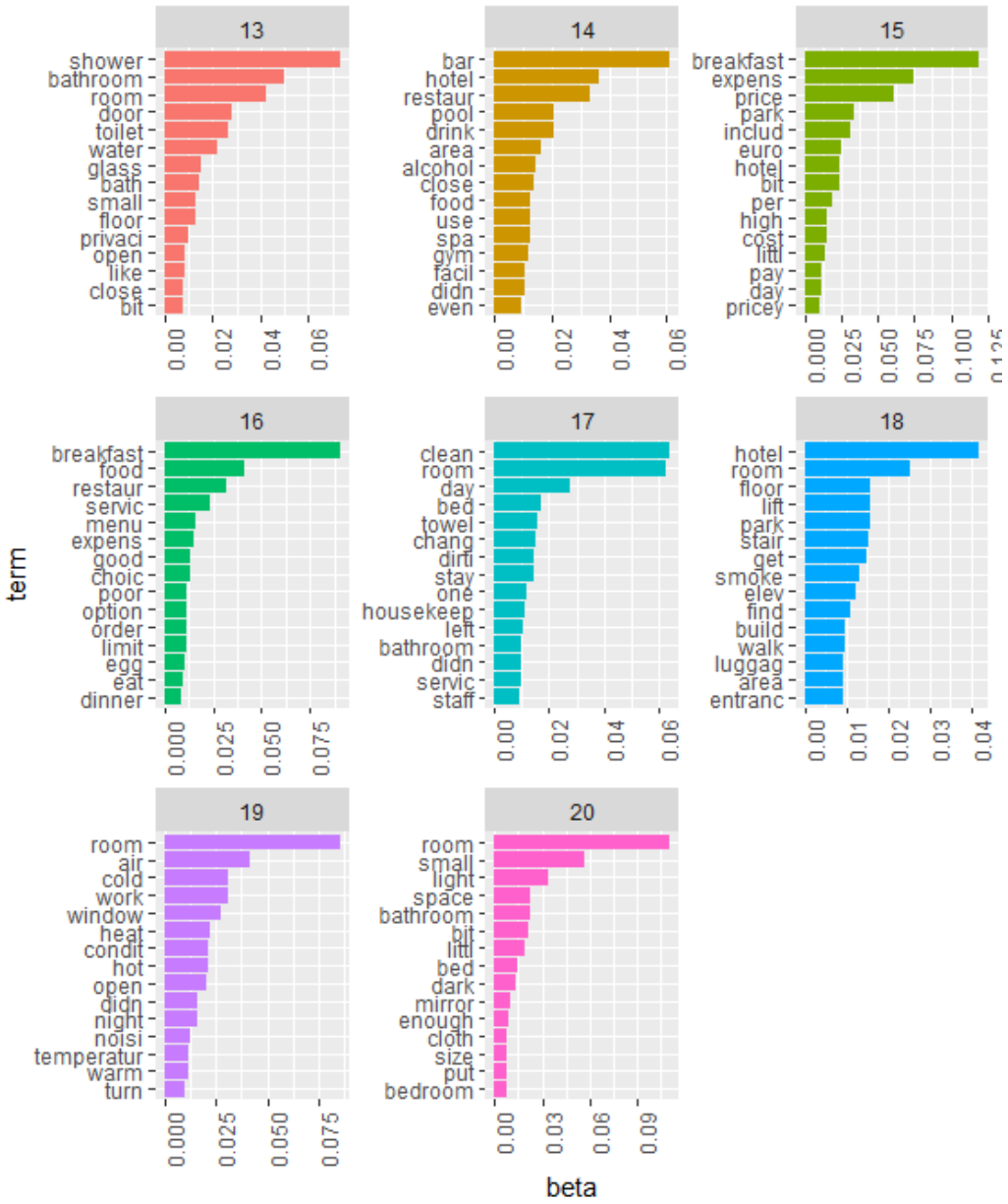


Figure 4. Negative LDA topics (topics 13 – 20)

Appendix B: JST sentiment topics & Dictionary

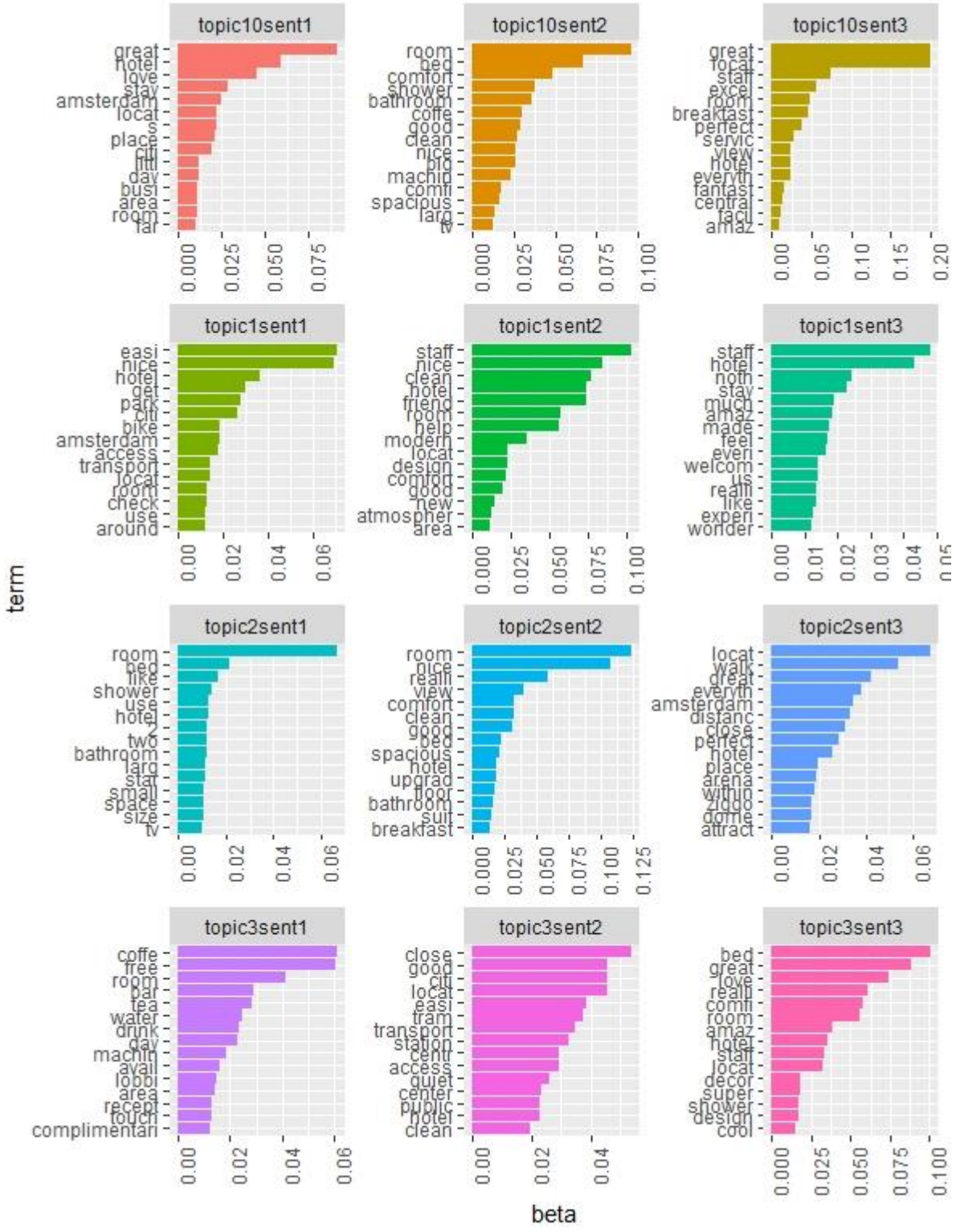


Figure 1. Some JST topics (10 topics, positive review section)

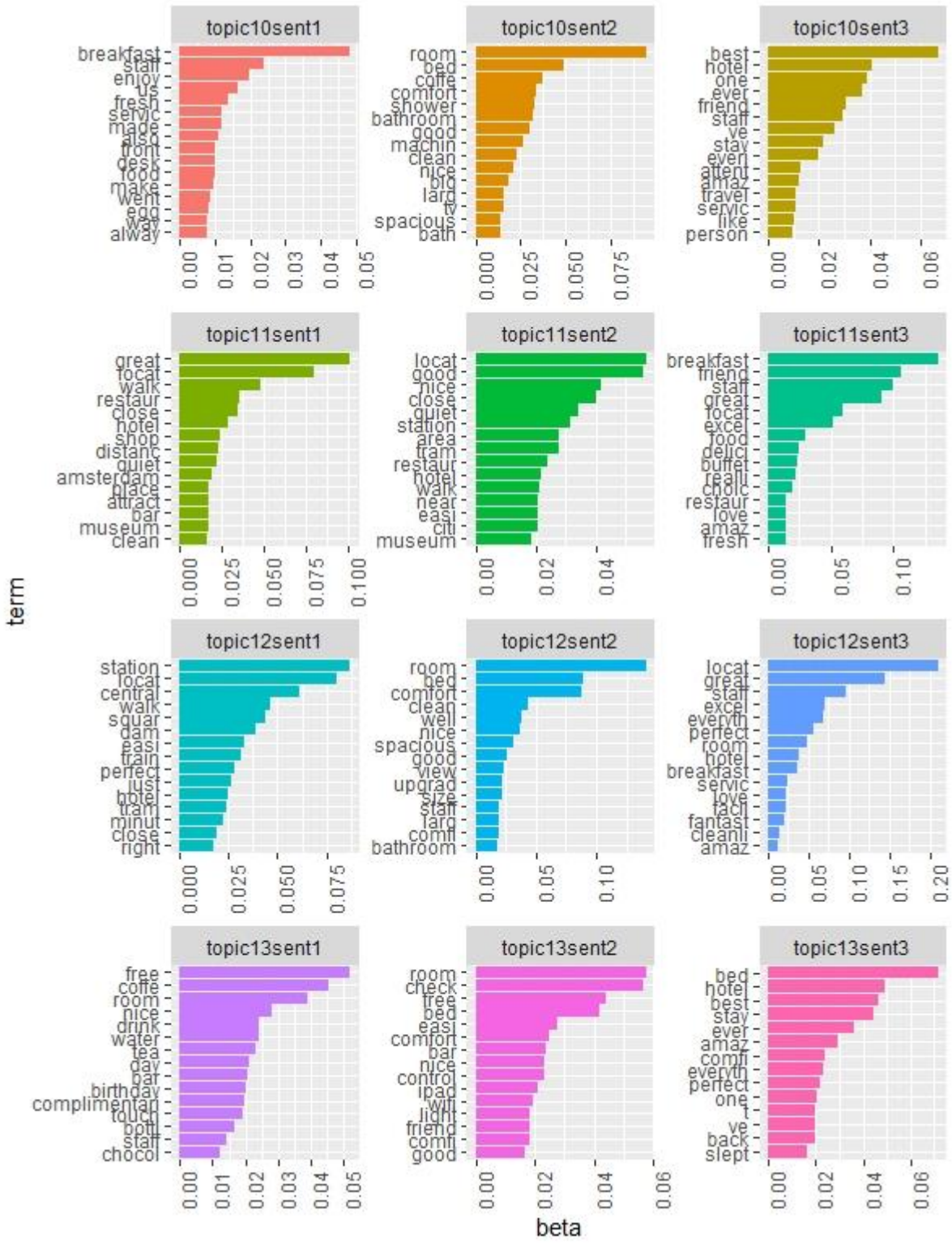


Figure 2. Some JST topics (15 topics, positive review section)

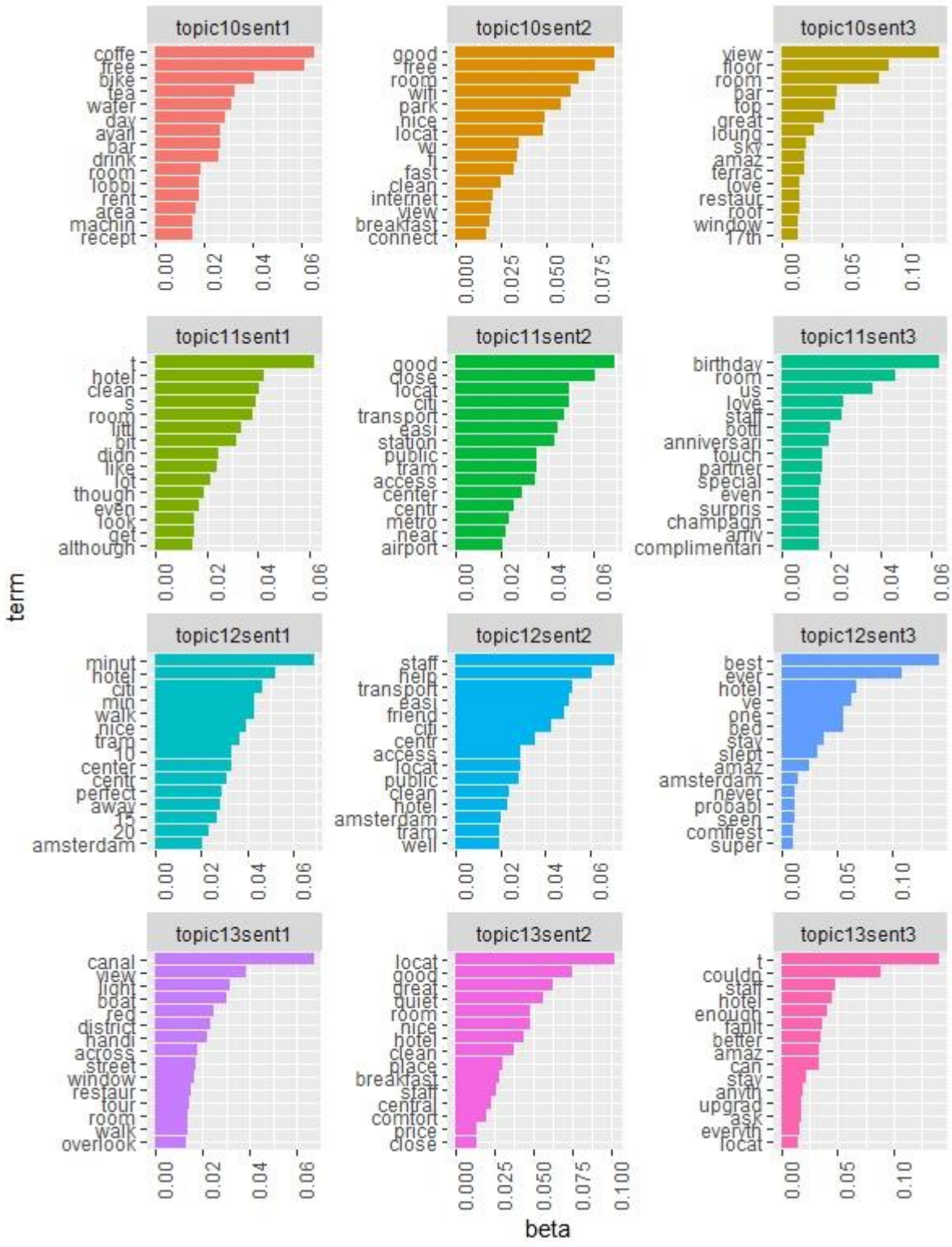


Figure 3. Some JST topics (20 topics, positive review section)

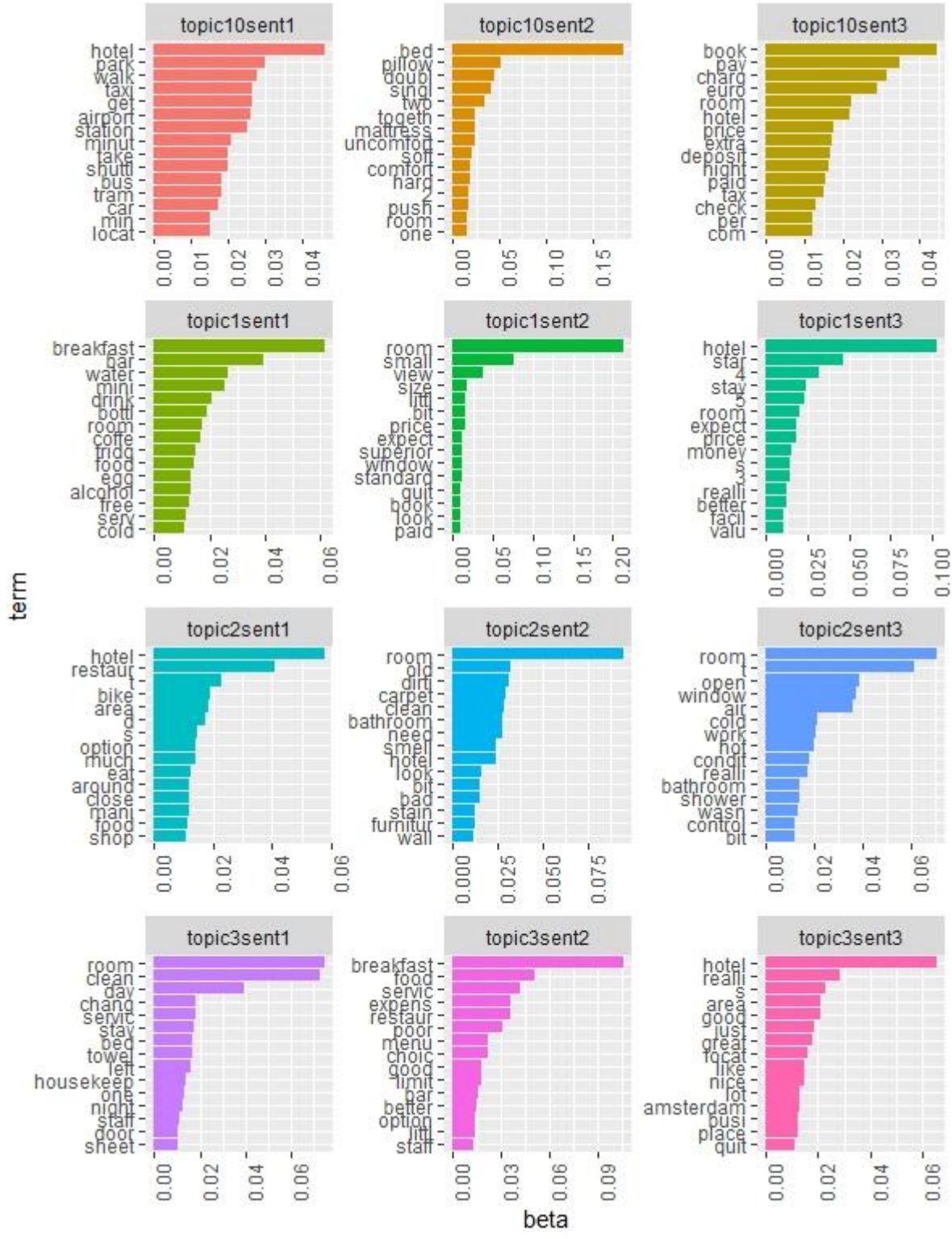


Figure 4. Some JST topics (10 topics, negative review section)

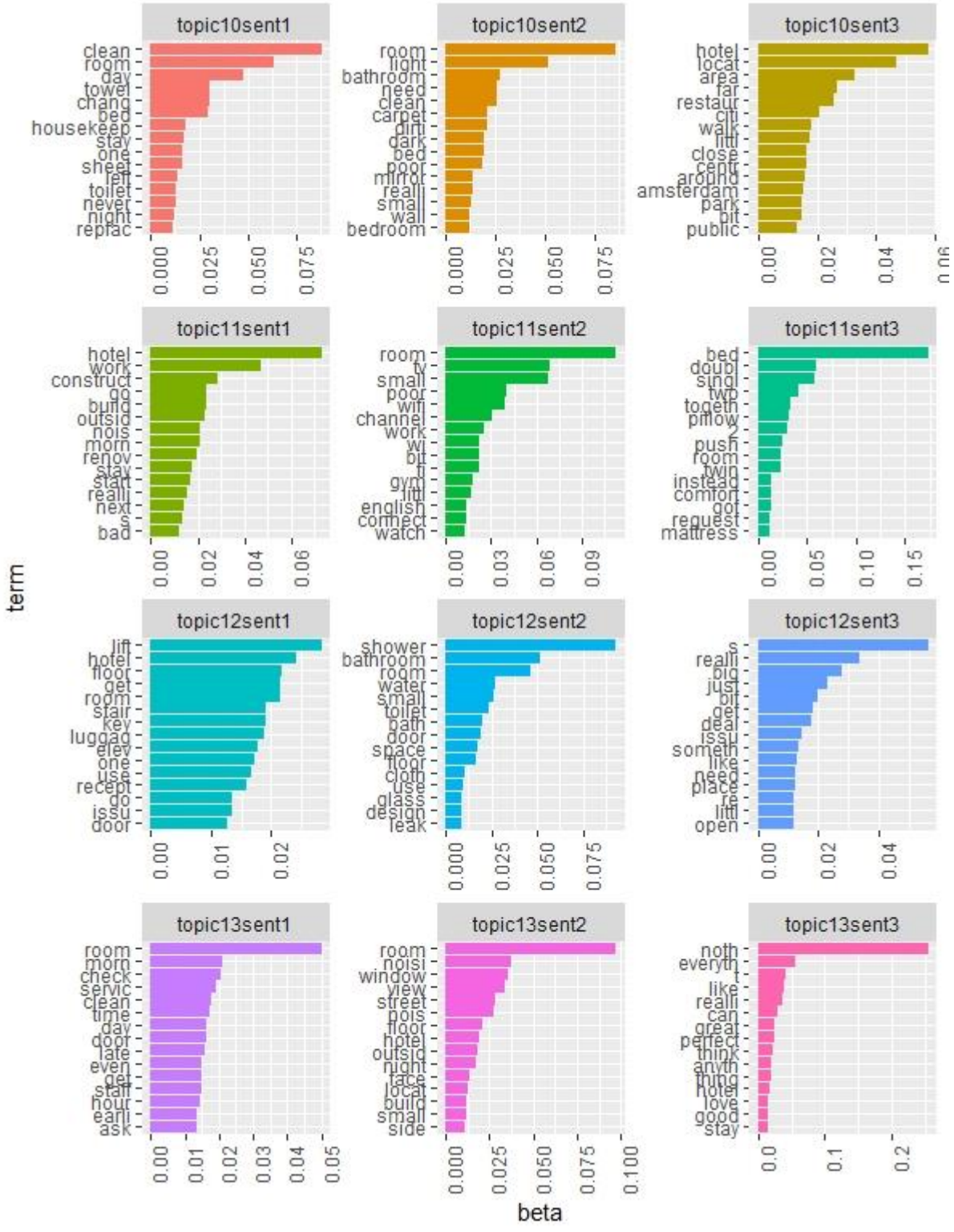


Figure 5. Some JST topics (15 topics, negative review section)

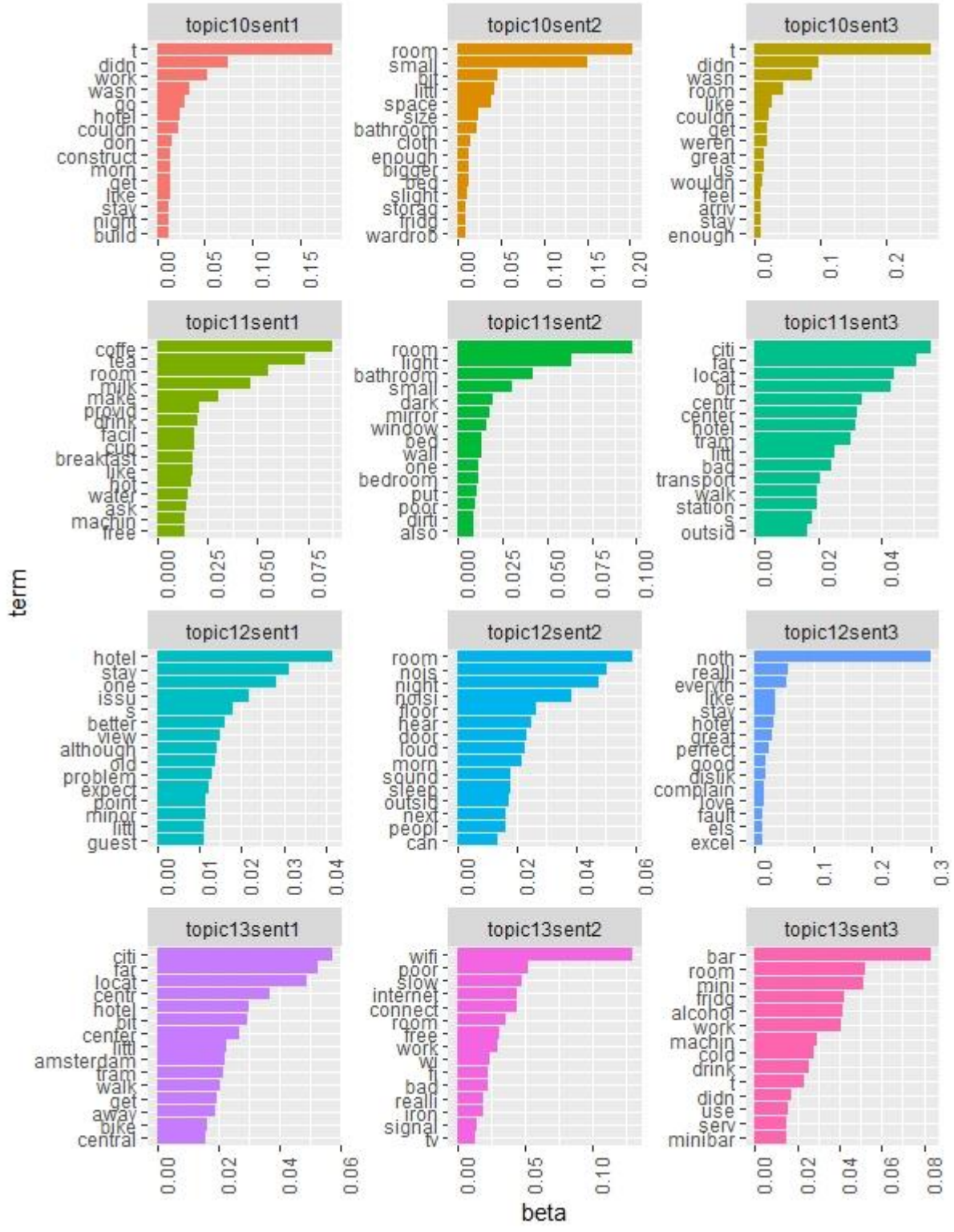


Figure 6. Some JST topics (20 topics, negative review section)

	Positive sentiment words (stems)	Negative sentiment words (stems)
+/- (sent 2)	friend, good, help, nice, clean, comfort, comfi, wel, quiet, easi, spacious, big, large, enjoy, lot, upgrad, valu, polit, cool, effic, pretty, posit, safe, bigger, fancy, satisfy, cheap, confort, friendli, handi, charm, helpful, easili, interest, cozi, cosi	small, expens, poor, noisi, bad, rude, dirti, uncomfort, difficult, overpr, loud, inconveni, unfriend, unhelp, outdat, awkward, unpleas
++/-- (sent 3)	great, excel, love, realli, perfect, amaz, beauti, fantast, best, awesom, ever, absolut, fabul, incred, luxuri, ideal, spotless, stun, spectacular, unbeliev, generous	extrem, serious

Figure 7. JST custom sentiment dictionary

Appendix C: Tobit regression total data

Results: Tobit All Data

	Dependent variable:
	Reviewer_Score
LDA_POS_TOPIC_1	0.519*** (0.055)
LDA_POS_TOPIC_2	1.175*** (0.057)
LDA_POS_TOPIC_3	1.032*** (0.057)
LDA_POS_TOPIC_4	0.991*** (0.057)
LDA_POS_TOPIC_5	1.890*** (0.067)
LDA_POS_TOPIC_6	0.836*** (0.055)
LDA_POS_TOPIC_7	1.880*** (0.064)
LDA_POS_TOPIC_8	1.236*** (0.058)
LDA_POS_TOPIC_9	1.148*** (0.050)
LDA_POS_TOPIC_10	0.968*** (0.059)
LDA_POS_TOPIC_11	-0.146** (0.065)
LDA_POS_TOPIC_12	0.883*** (0.067)
LDA_POS_TOPIC_13	0.836*** (0.054)
LDA_POS_TOPIC_14	1.748*** (0.057)
LDA_POS_TOPIC_15	1.146*** (0.049)
LDA_POS_TOPIC_16	2.560*** (0.059)
LDA_POS_TOPIC_17	1.262*** (0.052)
LDA_POS_TOPIC_18	0.827*** (0.051)
LDA_POS_TOPIC_19	1.091*** (0.063)
LDA_POS_TOPIC_20	1.356*** (0.063)
LDA_NEG_TOPIC_1	-2.560*** (0.054)
LDA_NEG_TOPIC_2	-0.958*** (0.071)
LDA_NEG_TOPIC_3	1.856*** (0.050)
LDA_NEG_TOPIC_4	-2.769*** (0.058)
LDA_NEG_TOPIC_5	-1.266*** (0.042)
LDA_NEG_TOPIC_6	-1.482*** (0.059)
LDA_NEG_TOPIC_7	-1.489*** (0.053)
LDA_NEG_TOPIC_8	-2.320*** (0.065)
LDA_NEG_TOPIC_9	-2.382*** (0.055)
LDA_NEG_TOPIC_10	-0.933*** (0.062)
LDA_NEG_TOPIC_11	-2.972*** (0.056)
LDA_NEG_TOPIC_12	-2.386*** (0.059)
LDA_NEG_TOPIC_13	-1.410*** (0.057)

LDA_NEG_TOPIC_14	-0.902*** (0.060)
LDA_NEG_TOPIC_15	-0.999*** (0.049)
LDA_NEG_TOPIC_16	-0.699*** (0.055)
LDA_NEG_TOPIC_17	-2.917*** (0.061)
LDA_NEG_TOPIC_18	-1.125*** (0.067)
LDA_NEG_TOPIC_19	-1.527*** (0.060)
LDA_NEG_TOPIC_20	-1.309*** (0.058)
Constant	8.699*** (0.031)
Observations	56,738
Log Likelihood	-93,777.080
Wald Test	23,162.220*** (df = 40)
Note:	*p**p***p<0.01

Appendix D: LDA partitioning trees

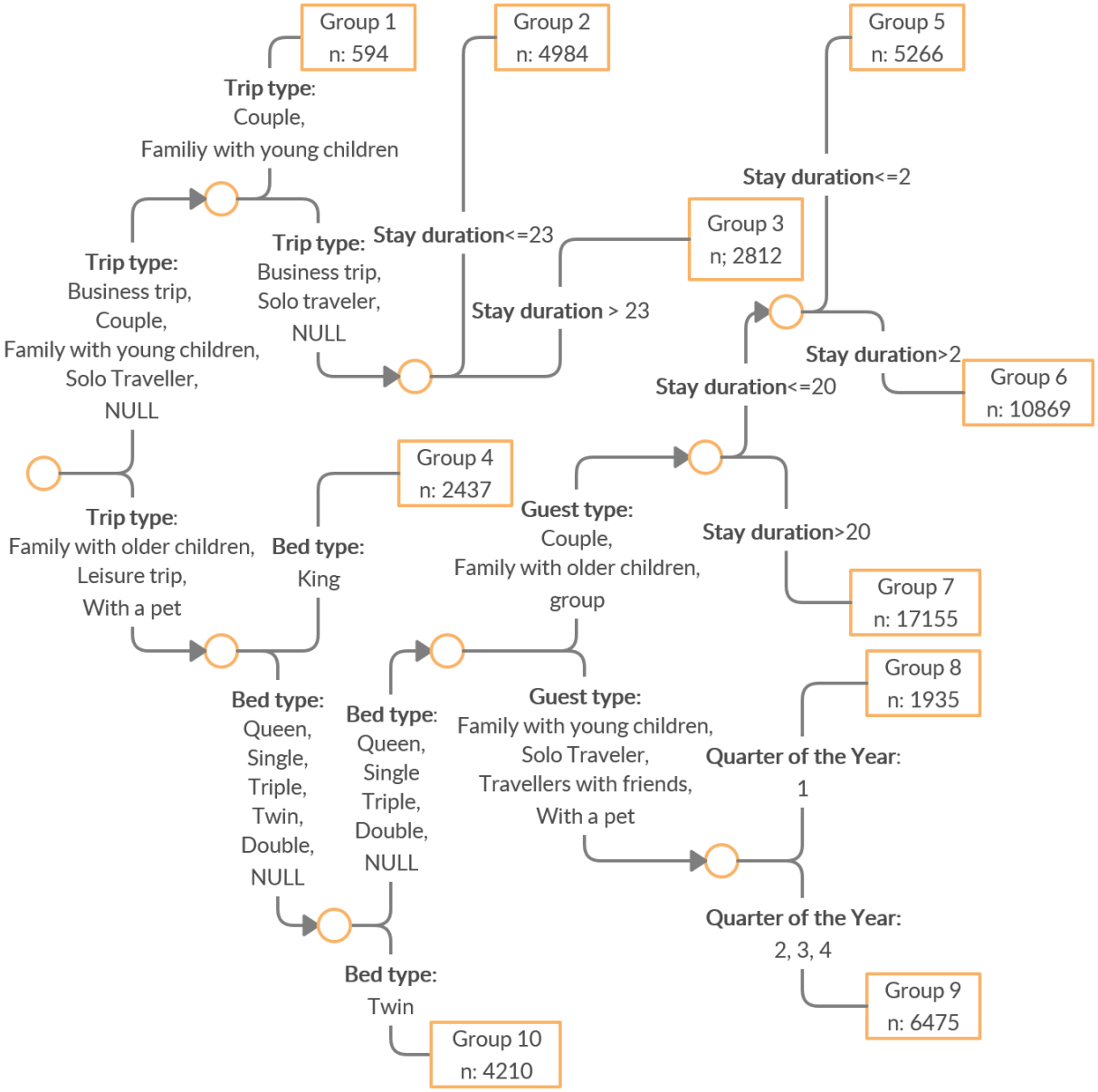


Figure 1. Uncorrected example LDA Partitioning Tree (*Trip type: All*)

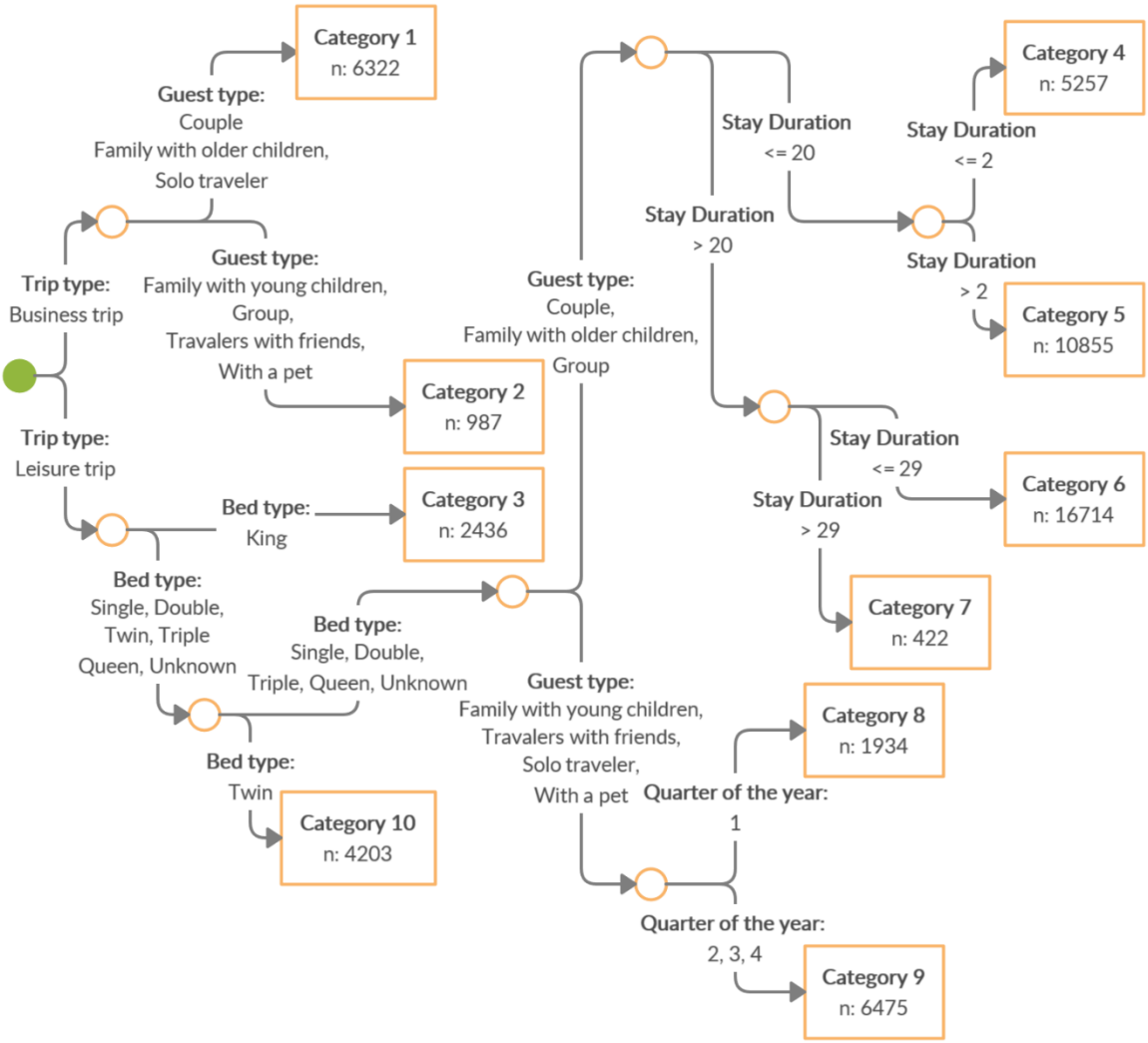


Figure 2. Corrected LDA Partitioning Tree (*Trip type: Business trip* or *Leisure Trip*)

Appendix E: Partitioning tree regressions

1. Business trip; Guest type: Family with older children, Couple, Solo traveler

```
Observations:
         Total  Uncensored Right-censored
          6322        5388            934

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
x(Intercept)       8.560592   0.084382 101.450  < 2e-16 ***
xLDA_POS_TOPIC_1   0.431846   0.148286   2.912  0.00359 **
xLDA_POS_TOPIC_2   0.916973   0.157765   5.812 6.16e-09 ***
xLDA_POS_TOPIC_3   0.727173   0.228111   3.188  0.00143 **
xLDA_POS_TOPIC_4   0.685341   0.189586   3.615  0.00030 ***
xLDA_POS_TOPIC_5   1.644921   0.251620   6.537 6.26e-11 ***
xLDA_POS_TOPIC_6   0.916822   0.161003   5.694 1.24e-08 ***
xLDA_POS_TOPIC_7   1.392324   0.180373   7.719 1.17e-14 ***
xLDA_POS_TOPIC_8   0.763812   0.181601   4.206 2.60e-05 ***
xLDA_POS_TOPIC_9   1.285722   0.143358   8.969  < 2e-16 ***
xLDA_POS_TOPIC_10  1.220003   0.171670   7.107 1.19e-12 ***
xLDA_POS_TOPIC_11 -0.174679   0.175282  -0.997  0.31898
xLDA_POS_TOPIC_12  1.017853   0.178923   5.689 1.28e-08 ***
xLDA_POS_TOPIC_13  1.021889   0.163981   6.232 4.61e-10 ***
xLDA_POS_TOPIC_14  1.764230   0.161273  10.939  < 2e-16 ***
xLDA_POS_TOPIC_15  1.078701   0.136921   7.878 3.32e-15 ***
xLDA_POS_TOPIC_16  2.099000   0.215774   9.728  < 2e-16 ***
xLDA_POS_TOPIC_17  0.956731   0.176514   5.420 5.95e-08 ***
xLDA_POS_TOPIC_18  0.774559   0.128288   6.038 1.56e-09 ***
xLDA_POS_TOPIC_19  1.377093   0.153772   8.955  < 2e-16 ***
xLDA_POS_TOPIC_20  1.077904   0.197294   5.463 4.67e-08 ***
xLDA_NEG_TOPIC_1  -3.118746   0.168025 -18.561  < 2e-16 ***
xLDA_NEG_TOPIC_2  -1.104549   0.166580  -6.631 3.34e-11 ***
xLDA_NEG_TOPIC_3   1.725100   0.170016  10.147  < 2e-16 ***
xLDA_NEG_TOPIC_4  -2.601090   0.154268 -16.861  < 2e-16 ***
xLDA_NEG_TOPIC_5  -1.212041   0.149681  -8.098 5.61e-16 ***
xLDA_NEG_TOPIC_6  -1.559042   0.195582  -7.971 1.57e-15 ***
xLDA_NEG_TOPIC_7  -1.861634   0.150603 -12.361  < 2e-16 ***
xLDA_NEG_TOPIC_8  -2.543444   0.177034 -14.367  < 2e-16 ***
xLDA_NEG_TOPIC_9  -2.586022   0.178108 -14.519  < 2e-16 ***
xLDA_NEG_TOPIC_10 -1.008664   0.204907  -4.923 8.54e-07 ***
xLDA_NEG_TOPIC_11 -3.062824   0.146258 -20.941  < 2e-16 ***
xLDA_NEG_TOPIC_12 -2.937865   0.189758 -15.482  < 2e-16 ***
xLDA_NEG_TOPIC_13 -1.441155   0.169825  -8.486  < 2e-16 ***
xLDA_NEG_TOPIC_14 -0.967592   0.193404  -5.003 5.65e-07 ***
xLDA_NEG_TOPIC_15 -1.060622   0.147786  -7.177 7.14e-13 ***
xLDA_NEG_TOPIC_16 -0.795343   0.155817  -5.104 3.32e-07 ***
xLDA_NEG_TOPIC_17 -2.966420   0.179031 -16.569  < 2e-16 ***
xLDA_NEG_TOPIC_18 -1.181192   0.196174  -6.021 1.73e-09 ***
xLDA_NEG_TOPIC_19 -1.583757   0.154663 -10.240  < 2e-16 ***
xLDA_NEG_TOPIC_20 -1.448760   0.167693  -8.639  < 2e-16 ***
Log(scale)         0.483779   0.009874  48.998  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.622

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -1.107e+04 on 42 Df
Wald-statistic: 1.538e+05 on 40 Df, p-value: < 2.22e-16
```

2. Business trip; Guest type: Family with young children, Group, Travelers with friends, With a pet

```
Observations:
         Total  Uncensored Right-censored
           987         822            165

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
x(Intercept)       8.58632    0.22375  38.374  < 2e-16 ***
xLDA_POS_TOPIC_1   0.36475    0.41296   0.883 0.377091
xLDA_POS_TOPIC_2   1.74079    0.51645   3.371 0.000750 ***
xLDA_POS_TOPIC_3   0.92389    0.64517   1.432 0.152142
xLDA_POS_TOPIC_4   1.12710    0.49020   2.299 0.021488 *
xLDA_POS_TOPIC_5   1.56687    0.74998   2.089 0.036689 *
xLDA_POS_TOPIC_6   0.85399    0.40304   2.119 0.034102 *
xLDA_POS_TOPIC_7   3.20993    0.50135   6.403 1.53e-10 ***
xLDA_POS_TOPIC_8   0.87228    0.43420   2.009 0.044545 *
xLDA_POS_TOPIC_9   1.07520    0.41802   2.572 0.010108 *
xLDA_POS_TOPIC_10  0.46019    0.49088   0.937 0.348520
xLDA_POS_TOPIC_11 -1.00269    0.49740  -2.016 0.043814 *
xLDA_POS_TOPIC_12 -0.39790    0.43748  -0.910 0.363077
xLDA_POS_TOPIC_13  1.14163    0.43358   2.633 0.008463 **
xLDA_POS_TOPIC_14  2.28280    0.46904   4.867 1.13e-06 ***
xLDA_POS_TOPIC_15  1.11899    0.38597   2.899 0.003741 **
xLDA_POS_TOPIC_16  2.31831    0.50486   4.592 4.39e-06 ***
xLDA_POS_TOPIC_17  0.55319    0.48438   1.142 0.253428
xLDA_POS_TOPIC_18  1.27360    0.33023   3.857 0.000115 ***
xLDA_POS_TOPIC_19  1.23394    0.46883   2.632 0.008489 **
xLDA_POS_TOPIC_20  0.86134    0.51945   1.658 0.097282 .
xLDA_NEG_TOPIC_1  -2.50226    0.44169  -5.665 1.47e-08 ***
xLDA_NEG_TOPIC_2  -1.03738    0.50812  -2.042 0.041191 *
xLDA_NEG_TOPIC_3   1.69416    0.47251   3.585 0.000336 ***
xLDA_NEG_TOPIC_4  -3.43133    0.39741  -8.634  < 2e-16 ***
xLDA_NEG_TOPIC_5  -0.70834    0.40456  -1.751 0.079968 .
xLDA_NEG_TOPIC_6  -2.59148    0.58268  -4.448 8.69e-06 ***
xLDA_NEG_TOPIC_7  -1.07363    0.46872  -2.291 0.021990 *
xLDA_NEG_TOPIC_8  -3.21968    0.51583  -6.242 4.33e-10 ***
xLDA_NEG_TOPIC_9  -2.97789    0.39419  -7.555 4.20e-14 ***
xLDA_NEG_TOPIC_10 -2.08432    0.57483  -3.626 0.000288 ***
xLDA_NEG_TOPIC_11 -3.95463    0.42996  -9.198  < 2e-16 ***
xLDA_NEG_TOPIC_12 -2.79539    0.49627  -5.633 1.77e-08 ***
xLDA_NEG_TOPIC_13 -1.32092    0.46502  -2.841 0.004503 **
xLDA_NEG_TOPIC_14 -1.01140    0.55178  -1.833 0.066808 .
xLDA_NEG_TOPIC_15 -0.85792    0.35123  -2.443 0.014581 *
xLDA_NEG_TOPIC_16 -0.53875    0.35937  -1.499 0.133836
xLDA_NEG_TOPIC_17 -3.37590    0.47429  -7.118 1.10e-12 ***
xLDA_NEG_TOPIC_18 -0.71055    0.51463  -1.381 0.167375
xLDA_NEG_TOPIC_19 -1.32110    0.45072  -2.931 0.003378 **
xLDA_NEG_TOPIC_20 -1.54193    0.48037  -3.210 0.001328 **
Log(scale)         0.53915    0.02534  21.279  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.715

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -1751 on 42 Df
Wald-statistic: 2.128e+04 on 40 Df, p-value: < 2.22e-16
```

3. Leisure trip; Bed: King-sized bed

```
Observations:
         Total  Uncensored Right-censored
          2436        1658            778

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
x(Intercept)       8.74582    0.16436  53.211  < 2e-16 ***
xLDA_POS_TOPIC_1   0.16718    0.32761   0.510 0.609842
xLDA_POS_TOPIC_2   1.58494    0.23851   6.645 3.03e-11 ***
xLDA_POS_TOPIC_3   1.66404    0.37526   4.434 9.23e-06 ***
xLDA_POS_TOPIC_4   1.02400    0.29035   3.527 0.000421 ***
xLDA_POS_TOPIC_5   2.19871    0.32104   6.849 7.46e-12 ***
xLDA_POS_TOPIC_6   0.31812    0.30691   1.037 0.299958
xLDA_POS_TOPIC_7   2.93696    0.30987   9.478  < 2e-16 ***
xLDA_POS_TOPIC_8   1.79695    0.29659   6.059 1.37e-09 ***
xLDA_POS_TOPIC_9   1.65083    0.26184   6.305 2.89e-10 ***
xLDA_POS_TOPIC_10  1.01305    0.29133   3.477 0.000506 ***
xLDA_POS_TOPIC_11 -0.65917    0.32967  -1.999 0.045559 *
xLDA_POS_TOPIC_12  1.54875    0.35992   4.303 1.68e-05 ***
xLDA_POS_TOPIC_13  0.91785    0.29687   3.092 0.001990 **
xLDA_POS_TOPIC_14  2.19972    0.24474   8.988  < 2e-16 ***
xLDA_POS_TOPIC_15  1.33375    0.23115   5.770 7.93e-09 ***
xLDA_POS_TOPIC_16  2.95078    0.27630  10.680  < 2e-16 ***
xLDA_POS_TOPIC_17  1.48141    0.25383   5.836 5.34e-09 ***
xLDA_POS_TOPIC_18  1.36124    0.27504   4.949 7.45e-07 ***
xLDA_POS_TOPIC_19  1.45428    0.32435   4.484 7.33e-06 ***
xLDA_POS_TOPIC_20  1.00781    0.33044   3.050 0.002289 **
xLDA_NEG_TOPIC_1  -2.48093    0.27204  -9.120  < 2e-16 ***
xLDA_NEG_TOPIC_2  -0.58763    0.42021  -1.398 0.161986
xLDA_NEG_TOPIC_3   1.89498    0.22913   8.270  < 2e-16 ***
xLDA_NEG_TOPIC_4  -2.45951    0.26335  -9.339  < 2e-16 ***
xLDA_NEG_TOPIC_5  -1.62801    0.31450  -5.176 2.26e-07 ***
xLDA_NEG_TOPIC_6  -1.61008    0.32695  -4.924 8.46e-07 ***
xLDA_NEG_TOPIC_7  -1.37145    0.25160  -5.451 5.01e-08 ***
xLDA_NEG_TOPIC_8  -1.81745    0.32198  -5.645 1.66e-08 ***
xLDA_NEG_TOPIC_9  -2.77892    0.30848  -9.008  < 2e-16 ***
xLDA_NEG_TOPIC_10 -1.19080    0.30927  -3.850 0.000118 ***
xLDA_NEG_TOPIC_11 -2.18770    0.32086  -6.818 9.22e-12 ***
xLDA_NEG_TOPIC_12 -2.18530    0.27469  -7.956 1.78e-15 ***
xLDA_NEG_TOPIC_13 -1.37671    0.27674  -4.975 6.54e-07 ***
xLDA_NEG_TOPIC_14 -0.91323    0.27602  -3.309 0.000938 ***
xLDA_NEG_TOPIC_15 -1.25683    0.26102  -4.815 1.47e-06 ***
xLDA_NEG_TOPIC_16 -0.69442    0.23318  -2.978 0.002901 **
xLDA_NEG_TOPIC_17 -2.37571    0.28478  -8.342  < 2e-16 ***
xLDA_NEG_TOPIC_18 -1.31637    0.34598  -3.805 0.000142 ***
xLDA_NEG_TOPIC_19 -1.14048    0.33654  -3.389 0.000702 ***
xLDA_NEG_TOPIC_20 -1.13704    0.28912  -3.933 8.40e-05 ***
Log(scale)         0.46275    0.01813  25.520  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.588

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -3647 on 42 Df
Wald-statistic: 6.66e+04 on 40 Df, p-value: < 2.22e-16
```

4. Leisure trip; Guest type: Couple, Family with older children, Group ; Bed: Single, Double, Triple, Queen, Unknown; Stay duration <= 2

Observations:

Total	Uncensored	Right-censored
5257	3996	1261

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.6859	0.1038	83.641	< 2e-16	***
xLDA_POS_TOPIC_1	0.4856	0.1900	2.555	0.010612	*
xLDA_POS_TOPIC_2	0.8846	0.1975	4.480	7.46e-06	***
xLDA_POS_TOPIC_3	0.6926	0.2001	3.461	0.000538	***
xLDA_POS_TOPIC_4	0.8051	0.1923	4.186	2.83e-05	***
xLDA_POS_TOPIC_5	1.6900	0.1908	8.855	< 2e-16	***
xLDA_POS_TOPIC_6	1.0178	0.1778	5.724	1.04e-08	***
xLDA_POS_TOPIC_7	2.3074	0.2067	11.164	< 2e-16	***
xLDA_POS_TOPIC_8	1.2792	0.2031	6.298	3.02e-10	***
xLDA_POS_TOPIC_9	1.2972	0.1602	8.096	5.69e-16	***
xLDA_POS_TOPIC_10	0.9867	0.1703	5.794	6.89e-09	***
xLDA_POS_TOPIC_11	-0.1957	0.2095	-0.934	0.350232	
xLDA_POS_TOPIC_12	0.7825	0.2383	3.284	0.001025	**
xLDA_POS_TOPIC_13	0.8562	0.1719	4.981	6.31e-07	***
xLDA_POS_TOPIC_14	1.6080	0.1834	8.767	< 2e-16	***
xLDA_POS_TOPIC_15	1.1256	0.1638	6.871	6.38e-12	***
xLDA_POS_TOPIC_16	2.7064	0.2014	13.439	< 2e-16	***
xLDA_POS_TOPIC_17	1.3368	0.1694	7.891	2.99e-15	***
xLDA_POS_TOPIC_18	0.8133	0.1646	4.942	7.72e-07	***
xLDA_POS_TOPIC_19	0.8976	0.1905	4.713	2.44e-06	***
xLDA_POS_TOPIC_20	1.5313	0.2139	7.160	8.06e-13	***
xLDA_NEG_TOPIC_1	-2.2098	0.1692	-13.064	< 2e-16	***
xLDA_NEG_TOPIC_2	-0.7296	0.2425	-3.009	0.002625	**
xLDA_NEG_TOPIC_3	2.0055	0.1572	12.754	< 2e-16	***
xLDA_NEG_TOPIC_4	-2.5666	0.1807	-14.206	< 2e-16	***
xLDA_NEG_TOPIC_5	-1.0048	0.1336	-7.521	5.43e-14	***
xLDA_NEG_TOPIC_6	-1.0589	0.2153	-4.918	8.76e-07	***
xLDA_NEG_TOPIC_7	-1.5944	0.1798	-8.866	< 2e-16	***
xLDA_NEG_TOPIC_8	-2.2060	0.2057	-10.724	< 2e-16	***
xLDA_NEG_TOPIC_9	-2.2023	0.1754	-12.559	< 2e-16	***
xLDA_NEG_TOPIC_10	-0.9564	0.2121	-4.510	6.49e-06	***
xLDA_NEG_TOPIC_11	-3.2870	0.1779	-18.477	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.5838	0.1901	-13.591	< 2e-16	***
xLDA_NEG_TOPIC_13	-1.5495	0.1884	-8.226	< 2e-16	***
xLDA_NEG_TOPIC_14	-0.8505	0.2215	-3.839	0.000123	***
xLDA_NEG_TOPIC_15	-1.0745	0.1457	-7.376	1.63e-13	***
xLDA_NEG_TOPIC_16	-0.5806	0.2158	-2.691	0.007133	**
xLDA_NEG_TOPIC_17	-2.6878	0.2498	-10.761	< 2e-16	***
xLDA_NEG_TOPIC_18	-1.1743	0.1953	-6.012	1.83e-09	***
xLDA_NEG_TOPIC_19	-1.5271	0.2030	-7.522	5.38e-14	***
xLDA_NEG_TOPIC_20	-1.2407	0.2050	-6.052	1.43e-09	***
Log(scale)	0.4520	0.0116	38.976	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.572

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -8444 on 42 Df
Wald-statistic: 1.467e+05 on 40 Df, p-value: < 2.22e-16

5. Leisure trip; Guest type: Couple, Family with older children, Group ; Bed: Single, Double, Triple, Queen, Unknown; 2 < Stay duration <= 20

Observations:

Total	Uncensored	Right-censored
10855	8195	2660

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.779244	0.071809	122.258	< 2e-16	***
xLDA_POS_TOPIC_1	0.522553	0.133100	3.926	8.64e-05	***
xLDA_POS_TOPIC_2	1.153757	0.125054	9.226	< 2e-16	***
xLDA_POS_TOPIC_3	0.992976	0.120011	8.274	< 2e-16	***
xLDA_POS_TOPIC_4	0.886676	0.129875	6.827	8.66e-12	***
xLDA_POS_TOPIC_5	1.836369	0.134343	13.669	< 2e-16	***
xLDA_POS_TOPIC_6	0.657307	0.125764	5.227	1.73e-07	***
xLDA_POS_TOPIC_7	2.147493	0.145210	14.789	< 2e-16	***
xLDA_POS_TOPIC_8	1.283561	0.132242	9.706	< 2e-16	***
xLDA_POS_TOPIC_9	1.086804	0.108955	9.975	< 2e-16	***
xLDA_POS_TOPIC_10	0.936689	0.134390	6.970	3.17e-12	***
xLDA_POS_TOPIC_11	0.038896	0.149429	0.260	0.794633	
xLDA_POS_TOPIC_12	0.610938	0.158575	3.853	0.000117	***
xLDA_POS_TOPIC_13	0.817720	0.121185	6.748	1.50e-11	***
xLDA_POS_TOPIC_14	1.708634	0.126978	13.456	< 2e-16	***
xLDA_POS_TOPIC_15	1.147443	0.111199	10.319	< 2e-16	***
xLDA_POS_TOPIC_16	2.605904	0.125857	20.705	< 2e-16	***
xLDA_POS_TOPIC_17	1.205284	0.107324	11.230	< 2e-16	***
xLDA_POS_TOPIC_18	0.800993	0.118693	6.748	1.49e-11	***
xLDA_POS_TOPIC_19	1.028698	0.145975	7.047	1.83e-12	***
xLDA_POS_TOPIC_20	1.344155	0.142300	9.446	< 2e-16	***
xLDA_NEG_TOPIC_1	-2.790606	0.118271	-23.595	< 2e-16	***
xLDA_NEG_TOPIC_2	-0.884295	0.173977	-5.083	3.72e-07	***
xLDA_NEG_TOPIC_3	1.877235	0.106426	17.639	< 2e-16	***
xLDA_NEG_TOPIC_4	-2.910670	0.133287	-21.838	< 2e-16	***
xLDA_NEG_TOPIC_5	-1.355691	0.090870	-14.919	< 2e-16	***
xLDA_NEG_TOPIC_6	-1.631468	0.128342	-12.712	< 2e-16	***
xLDA_NEG_TOPIC_7	-1.673328	0.121731	-13.746	< 2e-16	***
xLDA_NEG_TOPIC_8	-1.904469	0.152812	-12.463	< 2e-16	***
xLDA_NEG_TOPIC_9	-2.167873	0.119857	-18.087	< 2e-16	***
xLDA_NEG_TOPIC_10	-0.670138	0.132875	-5.043	4.57e-07	***
xLDA_NEG_TOPIC_11	-3.080630	0.129120	-23.859	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.221298	0.123219	-18.027	< 2e-16	***
xLDA_NEG_TOPIC_13	-1.276056	0.123824	-10.305	< 2e-16	***
xLDA_NEG_TOPIC_14	-0.689357	0.128912	-5.347	8.92e-08	***
xLDA_NEG_TOPIC_15	-0.992852	0.104955	-9.460	< 2e-16	***
xLDA_NEG_TOPIC_16	-0.623354	0.126839	-4.915	8.90e-07	***
xLDA_NEG_TOPIC_17	-3.065943	0.159768	-19.190	< 2e-16	***
xLDA_NEG_TOPIC_18	-1.098585	0.153132	-7.174	7.28e-13	***
xLDA_NEG_TOPIC_19	-1.466297	0.145167	-10.101	< 2e-16	***
xLDA_NEG_TOPIC_20	-1.351717	0.131041	-10.315	< 2e-16	***
Log(scale)	0.437597	0.008098	54.036	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.549

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -1.722e+04 on 42 Df
Wald-statistic: 3.128e+05 on 40 Df, p-value: < 2.22e-16

6. Leisure trip; Guest type: Couple, Family with older children, Group ; Bed: Single, Double, Triple, Queen, Unknown; 20 < Stay duration <= 29

Observations:

Total	Uncensored	Right-censored
16714	12936	3778

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.852678	0.059695	148.298	< 2e-16	***
xLDA_POS_TOPIC_1	0.446522	0.098963	4.512	6.42e-06	***
xLDA_POS_TOPIC_2	1.051937	0.104427	10.073	< 2e-16	***
xLDA_POS_TOPIC_3	0.926436	0.094834	9.769	< 2e-16	***
xLDA_POS_TOPIC_4	0.927936	0.098763	9.396	< 2e-16	***
xLDA_POS_TOPIC_5	1.822159	0.122855	14.832	< 2e-16	***
xLDA_POS_TOPIC_6	0.747919	0.102853	7.272	3.55e-13	***
xLDA_POS_TOPIC_7	1.541498	0.117800	13.086	< 2e-16	***
xLDA_POS_TOPIC_8	1.187988	0.102554	11.584	< 2e-16	***
xLDA_POS_TOPIC_9	0.910367	0.090921	10.013	< 2e-16	***
xLDA_POS_TOPIC_10	0.707395	0.106118	6.666	2.63e-11	***
xLDA_POS_TOPIC_11	-0.015522	0.122362	-0.127	0.899	
xLDA_POS_TOPIC_12	1.068649	0.118942	8.985	< 2e-16	***
xLDA_POS_TOPIC_13	0.693936	0.096826	7.167	7.67e-13	***
xLDA_POS_TOPIC_14	1.552828	0.106622	14.564	< 2e-16	***
xLDA_POS_TOPIC_15	1.022460	0.088113	11.604	< 2e-16	***
xLDA_POS_TOPIC_16	2.366996	0.101846	23.241	< 2e-16	***
xLDA_POS_TOPIC_17	1.080250	0.090968	11.875	< 2e-16	***
xLDA_POS_TOPIC_18	0.668482	0.098439	6.791	1.12e-11	***
xLDA_POS_TOPIC_19	0.868071	0.121865	7.123	1.05e-12	***
xLDA_POS_TOPIC_20	1.387335	0.110526	12.552	< 2e-16	***
xLDA_NEG_TOPIC_1	-2.396697	0.098026	-24.450	< 2e-16	***
xLDA_NEG_TOPIC_2	-0.771569	0.139639	-5.525	3.29e-08	***
xLDA_NEG_TOPIC_3	1.661840	0.085987	19.327	< 2e-16	***
xLDA_NEG_TOPIC_4	-2.647013	0.109600	-24.152	< 2e-16	***
xLDA_NEG_TOPIC_5	-1.166076	0.074830	-15.583	< 2e-16	***
xLDA_NEG_TOPIC_6	-1.412607	0.097725	-14.455	< 2e-16	***
xLDA_NEG_TOPIC_7	-1.333924	0.090863	-14.681	< 2e-16	***
xLDA_NEG_TOPIC_8	-2.025645	0.121403	-16.685	< 2e-16	***
xLDA_NEG_TOPIC_9	-2.307810	0.096558	-23.901	< 2e-16	***
xLDA_NEG_TOPIC_10	-0.956465	0.104621	-9.142	< 2e-16	***
xLDA_NEG_TOPIC_11	-2.696366	0.103614	-26.023	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.350226	0.103500	-22.708	< 2e-16	***
xLDA_NEG_TOPIC_13	-1.454586	0.099480	-14.622	< 2e-16	***
xLDA_NEG_TOPIC_14	-1.034019	0.100078	-10.332	< 2e-16	***
xLDA_NEG_TOPIC_15	-0.841583	0.090225	-9.328	< 2e-16	***
xLDA_NEG_TOPIC_16	-0.814329	0.095263	-8.548	< 2e-16	***
xLDA_NEG_TOPIC_17	-2.854407	0.100911	-28.286	< 2e-16	***
xLDA_NEG_TOPIC_18	-1.004613	0.127000	-7.910	2.57e-15	***
xLDA_NEG_TOPIC_19	-1.458356	0.108151	-13.484	< 2e-16	***
xLDA_NEG_TOPIC_20	-1.324508	0.100090	-13.233	< 2e-16	***
Log(scale)	0.434745	0.006442	67.491	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.545

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -2.698e+04 on 42 Df
Wald-statistic: 4.884e+05 on 40 Df, p-value: < 2.22e-16

7. Leisure trip; Guest type: Couple, Family with older children, Group ; Bed: Single, Double, Triple, Queen, Unknown; 29 < Stay duration

Observations:

Total	Uncensored	Right-censored
422	328	94

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	9.13800	0.42906	21.298	< 2e-16	***
xLDA_POS_TOPIC_1	1.14989	0.64656	1.778	0.075326	.
xLDA_POS_TOPIC_2	-0.02918	0.80563	-0.036	0.971110	
xLDA_POS_TOPIC_3	0.53784	0.76537	0.703	0.482232	
xLDA_POS_TOPIC_4	0.23331	0.68090	0.343	0.731858	
xLDA_POS_TOPIC_5	2.62365	0.80467	3.261	0.001112	**
xLDA_POS_TOPIC_6	0.80283	0.70644	1.136	0.255771	
xLDA_POS_TOPIC_7	2.10451	0.78407	2.684	0.007273	**
xLDA_POS_TOPIC_8	1.78986	0.66385	2.696	0.007014	**
xLDA_POS_TOPIC_9	1.68818	0.64894	2.601	0.009283	**
xLDA_POS_TOPIC_10	1.45560	0.73655	1.976	0.048129	*
xLDA_POS_TOPIC_11	-1.03444	0.80818	-1.280	0.200561	
xLDA_POS_TOPIC_12	0.77292	0.78310	0.987	0.323642	
xLDA_POS_TOPIC_13	0.23883	0.66425	0.360	0.719183	
xLDA_POS_TOPIC_14	1.92176	1.01826	1.887	0.059120	.
xLDA_POS_TOPIC_15	0.41479	0.62867	0.660	0.509387	
xLDA_POS_TOPIC_16	2.60046	0.71620	3.631	0.000282	***
xLDA_POS_TOPIC_17	0.68411	0.70907	0.965	0.334643	
xLDA_POS_TOPIC_18	1.45992	0.67088	2.176	0.029547	*
xLDA_POS_TOPIC_19	-0.26511	0.81089	-0.327	0.743715	
xLDA_POS_TOPIC_20	0.53366	0.82725	0.645	0.518864	
xLDA_NEG_TOPIC_1	-3.76045	0.66445	-5.659	1.52e-08	***
xLDA_NEG_TOPIC_2	-1.22992	0.83526	-1.473	0.140883	
xLDA_NEG_TOPIC_3	1.94169	0.74471	2.607	0.009125	**
xLDA_NEG_TOPIC_4	-3.87210	0.62136	-6.232	4.62e-10	***
xLDA_NEG_TOPIC_5	-0.78620	0.51258	-1.534	0.125080	
xLDA_NEG_TOPIC_6	-1.57916	0.68066	-2.320	0.020338	*
xLDA_NEG_TOPIC_7	-0.91934	0.71420	-1.287	0.198012	
xLDA_NEG_TOPIC_8	-3.42219	0.72080	-4.748	2.06e-06	***
xLDA_NEG_TOPIC_9	-2.65484	0.75631	-3.510	0.000448	***
xLDA_NEG_TOPIC_10	-0.41851	0.71184	-0.588	0.556583	
xLDA_NEG_TOPIC_11	-3.94532	0.80217	-4.918	8.73e-07	***
xLDA_NEG_TOPIC_12	-3.38720	0.77419	-4.375	1.21e-05	***
xLDA_NEG_TOPIC_13	-1.61189	0.85049	-1.895	0.058060	.
xLDA_NEG_TOPIC_14	-1.58127	0.99454	-1.590	0.111845	
xLDA_NEG_TOPIC_15	-1.75462	0.72263	-2.428	0.015177	*
xLDA_NEG_TOPIC_16	-0.77652	0.63967	-1.214	0.224773	
xLDA_NEG_TOPIC_17	-3.43378	0.55532	-6.183	6.27e-10	***
xLDA_NEG_TOPIC_18	1.30149	0.96407	1.350	0.177017	
xLDA_NEG_TOPIC_19	-1.89690	0.58895	-3.221	0.001278	**
xLDA_NEG_TOPIC_20	-1.40136	0.68594	-2.043	0.041056	*
Log(scale)	0.50133	0.04030	12.441	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.651

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -696.7 on 42 Df
Wald-statistic: 1.022e+04 on 40 Df, p-value: < 2.22e-16

8. Leisure trip; Guest type: Family with young children, Travelers with friends, Solo traveler, With a pet ; Bed: Single, Double, Triple, Queen, Unknown;
Quarter of the year: 1

Observations:

Total	Uncensored	Right-censored
1934	1474	460

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.68385	0.16722	51.930	< 2e-16	***
xLDA_POS_TOPIC_1	0.33831	0.32407	1.044	0.296525	
xLDA_POS_TOPIC_2	1.41123	0.30378	4.646	3.39e-06	***
xLDA_POS_TOPIC_3	1.39440	0.31543	4.421	9.84e-06	***
xLDA_POS_TOPIC_4	1.07246	0.31246	3.432	0.000599	***
xLDA_POS_TOPIC_5	1.38656	0.35806	3.872	0.000108	***
xLDA_POS_TOPIC_6	1.09555	0.29854	3.670	0.000243	***
xLDA_POS_TOPIC_7	2.48479	0.34526	7.197	6.16e-13	***
xLDA_POS_TOPIC_8	0.64139	0.29228	2.194	0.028204	*
xLDA_POS_TOPIC_9	0.77746	0.25131	3.094	0.001977	**
xLDA_POS_TOPIC_10	1.26747	0.32486	3.902	9.56e-05	***
xLDA_POS_TOPIC_11	0.15357	0.34162	0.450	0.653052	
xLDA_POS_TOPIC_12	0.73074	0.34856	2.096	0.036041	*
xLDA_POS_TOPIC_13	1.07996	0.29690	3.637	0.000275	***
xLDA_POS_TOPIC_14	1.71861	0.28569	6.016	1.79e-09	***
xLDA_POS_TOPIC_15	1.29657	0.25349	5.115	3.14e-07	***
xLDA_POS_TOPIC_16	2.67118	0.34008	7.855	4.01e-15	***
xLDA_POS_TOPIC_17	1.41454	0.29966	4.721	2.35e-06	***
xLDA_POS_TOPIC_18	0.72845	0.27196	2.679	0.007395	**
xLDA_POS_TOPIC_19	1.24890	0.31988	3.904	9.45e-05	***
xLDA_POS_TOPIC_20	0.88291	0.30001	2.943	0.003252	**
xLDA_NEG_TOPIC_1	-1.83318	0.31052	-5.904	3.56e-09	***
xLDA_NEG_TOPIC_2	-0.52082	0.39060	-1.333	0.182411	
xLDA_NEG_TOPIC_3	1.77626	0.24663	7.202	5.92e-13	***
xLDA_NEG_TOPIC_4	-2.83039	0.33833	-8.366	< 2e-16	***
xLDA_NEG_TOPIC_5	-0.90751	0.22720	-3.994	6.49e-05	***
xLDA_NEG_TOPIC_6	-1.18350	0.32343	-3.659	0.000253	***
xLDA_NEG_TOPIC_7	-0.96416	0.25588	-3.768	0.000165	***
xLDA_NEG_TOPIC_8	-2.24270	0.33223	-6.750	1.47e-11	***
xLDA_NEG_TOPIC_9	-2.05363	0.29301	-7.009	2.40e-12	***
xLDA_NEG_TOPIC_10	-0.68949	0.32717	-2.107	0.035079	*
xLDA_NEG_TOPIC_11	-2.76818	0.26718	-10.361	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.19204	0.33229	-6.597	4.20e-11	***
xLDA_NEG_TOPIC_13	-1.84505	0.33349	-5.533	3.16e-08	***
xLDA_NEG_TOPIC_14	-1.16354	0.30000	-3.878	0.000105	***
xLDA_NEG_TOPIC_15	-0.59027	0.30169	-1.957	0.050401	.
xLDA_NEG_TOPIC_16	-1.00873	0.30141	-3.347	0.000818	***
xLDA_NEG_TOPIC_17	-1.86646	0.31764	-5.876	4.20e-09	***
xLDA_NEG_TOPIC_18	-1.39043	0.39377	-3.531	0.000414	***
xLDA_NEG_TOPIC_19	-1.08102	0.28262	-3.825	0.000131	***
xLDA_NEG_TOPIC_20	-1.23440	0.29771	-4.146	3.38e-05	***
Log(scale)	0.44778	0.01913	23.408	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.565

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -3116 on 42 Df
Wald-statistic: 5.542e+04 on 40 Df, p-value: < 2.22e-16

9. Leisure trip; Guest type: Family with young children, Travelers with friends, Solo traveler, With a pet ; Bed: Single, Double, Triple, Queen, Unknown;
Quarter of the year: 2, 3, 4

Observations:

Total	Uncensored	Right-censored
6475	5208	1267

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.54400	0.09408	90.812	< 2e-16	***
xLDA_POS_TOPIC_1	0.63650	0.16183	3.933	8.38e-05	***
xLDA_POS_TOPIC_2	1.29200	0.18887	6.841	7.88e-12	***
xLDA_POS_TOPIC_3	1.11482	0.17075	6.529	6.62e-11	***
xLDA_POS_TOPIC_4	1.16034	0.16239	7.145	8.97e-13	***
xLDA_POS_TOPIC_5	1.97415	0.23841	8.280	< 2e-16	***
xLDA_POS_TOPIC_6	1.01386	0.15403	6.582	4.64e-11	***
xLDA_POS_TOPIC_7	1.77819	0.19223	9.250	< 2e-16	***
xLDA_POS_TOPIC_8	1.16370	0.17282	6.734	1.65e-11	***
xLDA_POS_TOPIC_9	1.42667	0.14896	9.578	< 2e-16	***
xLDA_POS_TOPIC_10	0.95583	0.18130	5.272	1.35e-07	***
xLDA_POS_TOPIC_11	-0.15820	0.18414	-0.859	0.39027	
xLDA_POS_TOPIC_12	0.97818	0.20420	4.790	1.67e-06	***
xLDA_POS_TOPIC_13	0.71441	0.16679	4.283	1.84e-05	***
xLDA_POS_TOPIC_14	1.79659	0.17531	10.248	< 2e-16	***
xLDA_POS_TOPIC_15	1.07201	0.14253	7.521	5.42e-14	***
xLDA_POS_TOPIC_16	2.34740	0.19422	12.086	< 2e-16	***
xLDA_POS_TOPIC_17	1.34434	0.17150	7.839	4.55e-15	***
xLDA_POS_TOPIC_18	1.20217	0.14525	8.276	< 2e-16	***
xLDA_POS_TOPIC_19	1.43087	0.18647	7.674	1.67e-14	***
xLDA_POS_TOPIC_20	1.37210	0.19266	7.122	1.06e-12	***
xLDA_NEG_TOPIC_1	-2.39012	0.16030	-14.910	< 2e-16	***
xLDA_NEG_TOPIC_2	-1.06301	0.20755	-5.122	3.03e-07	***
xLDA_NEG_TOPIC_3	1.98393	0.15070	13.164	< 2e-16	***
xLDA_NEG_TOPIC_4	-2.79310	0.16629	-16.797	< 2e-16	***
xLDA_NEG_TOPIC_5	-1.19400	0.12567	-9.501	< 2e-16	***
xLDA_NEG_TOPIC_6	-1.57363	0.20338	-7.737	1.02e-14	***
xLDA_NEG_TOPIC_7	-1.38796	0.15572	-8.913	< 2e-16	***
xLDA_NEG_TOPIC_8	-2.69556	0.18694	-14.419	< 2e-16	***
xLDA_NEG_TOPIC_9	-2.52487	0.17421	-14.494	< 2e-16	***
xLDA_NEG_TOPIC_10	-0.94073	0.19043	-4.940	7.81e-07	***
xLDA_NEG_TOPIC_11	-2.79167	0.16108	-17.331	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.33487	0.17117	-13.640	< 2e-16	***
xLDA_NEG_TOPIC_13	-1.22620	0.18692	-6.560	5.38e-11	***
xLDA_NEG_TOPIC_14	-1.12869	0.20193	-5.590	2.28e-08	***
xLDA_NEG_TOPIC_15	-1.05569	0.14368	-7.348	2.02e-13	***
xLDA_NEG_TOPIC_16	-0.46779	0.16708	-2.800	0.00511	**
xLDA_NEG_TOPIC_17	-2.99179	0.16748	-17.864	< 2e-16	***
xLDA_NEG_TOPIC_18	-1.02041	0.17986	-5.673	1.40e-08	***
xLDA_NEG_TOPIC_19	-1.56207	0.17702	-8.824	< 2e-16	***
xLDA_NEG_TOPIC_20	-1.16180	0.17805	-6.525	6.79e-11	***
Log(scale)	0.49215	0.01011	48.685	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.636

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -1.099e+04 on 42 Df
Wald-statistic: 1.629e+05 on 40 Df, p-value: < 2.22e-16

10. Leisure trip; Bed: Twin

Observations:

Total	Uncensored	Right-censored
4203	3447	756

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
x(Intercept)	8.6001	0.1096	78.435	< 2e-16	***
xLDA_POS_TOPIC_1	0.8817	0.1872	4.708	2.50e-06	***
xLDA_POS_TOPIC_2	1.3508	0.1872	7.216	5.34e-13	***
xLDA_POS_TOPIC_3	0.8617	0.1897	4.543	5.56e-06	***
xLDA_POS_TOPIC_4	0.9702	0.2080	4.665	3.08e-06	***
xLDA_POS_TOPIC_5	2.0069	0.2502	8.021	1.05e-15	***
xLDA_POS_TOPIC_6	0.8503	0.2054	4.139	3.49e-05	***
xLDA_POS_TOPIC_7	1.2407	0.2434	5.096	3.46e-07	***
xLDA_POS_TOPIC_8	1.6392	0.2013	8.142	3.89e-16	***
xLDA_POS_TOPIC_9	0.9236	0.1791	5.158	2.50e-07	***
xLDA_POS_TOPIC_10	1.0114	0.2299	4.399	1.09e-05	***
xLDA_POS_TOPIC_11	-0.2746	0.2296	-1.196	0.231643	
xLDA_POS_TOPIC_12	0.6013	0.2477	2.428	0.015201	*
xLDA_POS_TOPIC_13	0.5740	0.1910	3.006	0.002650	**
xLDA_POS_TOPIC_14	1.7174	0.2316	7.415	1.22e-13	***
xLDA_POS_TOPIC_15	1.2994	0.1827	7.111	1.15e-12	***
xLDA_POS_TOPIC_16	2.3840	0.2248	10.606	< 2e-16	***
xLDA_POS_TOPIC_17	1.2851	0.1783	7.207	5.73e-13	***
xLDA_POS_TOPIC_18	0.8262	0.1866	4.427	9.53e-06	***
xLDA_POS_TOPIC_19	1.0683	0.2858	3.738	0.000185	***
xLDA_POS_TOPIC_20	1.3562	0.2371	5.721	1.06e-08	***
xLDA_NEG_TOPIC_1	-2.7099	0.2025	-13.383	< 2e-16	***
xLDA_NEG_TOPIC_2	-0.8321	0.2547	-3.267	0.001087	**
xLDA_NEG_TOPIC_3	1.9474	0.1962	9.925	< 2e-16	***
xLDA_NEG_TOPIC_4	-2.5747	0.2365	-10.885	< 2e-16	***
xLDA_NEG_TOPIC_5	-1.5079	0.1204	-12.521	< 2e-16	***
xLDA_NEG_TOPIC_6	-1.5668	0.1987	-7.887	3.11e-15	***
xLDA_NEG_TOPIC_7	-1.5281	0.2060	-7.417	1.20e-13	***
xLDA_NEG_TOPIC_8	-2.6856	0.2421	-11.095	< 2e-16	***
xLDA_NEG_TOPIC_9	-2.4434	0.1941	-12.588	< 2e-16	***
xLDA_NEG_TOPIC_10	-1.1668	0.2344	-4.979	6.39e-07	***
xLDA_NEG_TOPIC_11	-2.8875	0.2326	-12.411	< 2e-16	***
xLDA_NEG_TOPIC_12	-2.0432	0.2432	-8.401	< 2e-16	***
xLDA_NEG_TOPIC_13	-1.3192	0.2092	-6.305	2.88e-10	***
xLDA_NEG_TOPIC_14	-0.6793	0.2095	-3.242	0.001185	**
xLDA_NEG_TOPIC_15	-1.0311	0.1755	-5.875	4.23e-09	***
xLDA_NEG_TOPIC_16	-0.5619	0.2156	-2.606	0.009159	**
xLDA_NEG_TOPIC_17	-3.4075	0.2239	-15.220	< 2e-16	***
xLDA_NEG_TOPIC_18	-1.4426	0.2587	-5.576	2.46e-08	***
xLDA_NEG_TOPIC_19	-1.6542	0.2251	-7.347	2.03e-13	***
xLDA_NEG_TOPIC_20	-1.2435	0.2350	-5.292	1.21e-07	***
Log(scale)	0.4831	0.0124	38.967	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 1.621

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -7189 on 42 Df
Wald-statistic: 1.066e+05 on 40 Df, p-value: < 2.22e-16

Bibliography

Archak, N., Ghose, A., & Ipeirotis, P. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, 1485-1509.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet Rating of Product Reviews. (pp. 461-472). Berlin Heidelberg: Springer-Verlag .

Balikas, G., Amini, M., & Clausel, M. (2016). On a Topic Model for Sentences. *SIGIR '16 Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 921-924). Pisa, Italy: ACM.

Bazaarvoice. (2007). Opinion Poll Shows 8 out of 10 US Shoppers Put More Trust in Brands that Offer Customer Reviews.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Brody, S., & Elhadad, N. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 804-812). Los Angeles, California: Association for Computational Linguistics.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2008). A density-based method for adaptive LDA model Selection. *Neurocomputing 72*, 1775-1781.

Deveaud, R., Sanjuan, E., & Bellot, P. (2014). Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Document numérique* , 61-84.

Duan, W., Gu, B., & Whinston, B. (2008). Do online reviews matter? - An Empirical investigation of panel data. *Decision Support Systems*, 1007–1016.

Ellison, G., & Fudenberg, D. (1995). Word-of-mouth Communication and Social Learning. *Quarterly Journal of Economics*, 93-125.

eMarketer. (2007b, April 11). *Niche Sites Invigorate Online Travel*. Retrieved from http://www.eMarketer.com.

Embracing Consumer Buzz Creates Measurement Challenges for Marketers. (2006). *Compete Inc.*

Furnas, G., Deerwester, S., Dumais, S., T.K., L., Harshman, R., Streeter, L., & Lochbaum, K. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *SIGIR '88 Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 465-480). Grenoble, France: ACM.

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content. *Twelth International Workshop on the Web and Databases.*

Geigle, C. (2016). *Inference Methods for Latent Dirichlet Allocation.* University of Illinois at Urbana-Champaign, Department of Computer Science.

Gretzel, U., & Yoo, K. (2008). Use and Impact of Online Travel Reviews. *Information and Communication Technologies in Tourism*, (pp. 35-46). Innsbruck, Austria.

Heinrich, G. (2008). Parameter estimation for text analysis. *University of Leipzig*.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Stockhold, Sweden: Morgen Kaufmann Publishers Inc.

Jo, Y., & Oh, A. (2011). Aspect and Sentiment Unification Model. *Proceedings of the Forth International Conference on Web Search and Web Data Mining* (pp. 815-824). Hong Kong, China: ACM.

Kamath, R., Ochi, M., & Matsuo, Y. (2015). Understanding Rating Behaviour and Predicting Ratings by Identifying Representative Users. *Pacific Asia Conference on Language, Information and Computation*, (pp. 522-528). Shanghai, China.

Lin, C., & He, Y. (2009). Joint Sentiment/Topic Model for Sentiment Analysis. *CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375-384). Hong Kong, China: ACM.

Lin, C., He, Y., Everson, R., & Ruger, S. (2012). Weakly Supervised Joint Sentiment-Topic Detection from Tex. *IEEE Transactions on Knowledge and Data Engineering*, 1134-1145 .

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *Transactions on Information Theory*, 145-151.

Munk-Nielsen, A. (2016). Maximum Likelihood: The Tobit Model. University of Copenhagen, Department of Economics.

Porter, M. (1980). *An algorithm for suffix stripping.* Cambridge, U.K.: Computer Laboratory.

Project, P. I. (2006). Retrieved from www.PewInternet.org.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2019). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *Journal of Latex Class Files*.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online Peer and Editorial Recommendations, Trust, and Choice in Virtual Markets. *Journal of Interactive Marketing*, 15-37.

Teleconference: Forty Facts About The US Online Shopper. (2006). *Forrester Research*.

Titov, I., & McDonald, R. (2008). Modeling Online Reviews with Multi-grain Topic Model. *WWW '08 Proceedings of the 17th international conference on World Wide Web* (pp. 111-120). Beijing, China: ACM.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 24-36.

Tuominen, P. (2011). The influence of TripAdvisor consumer-generated travel reviews of hotel performance. *University of Hertfordshire Business School Working Paper, Presented at 19th Annual Frontiers in Service Conference*, 1-11.

Wang, H., & Ester, M. (2014). A Sentiment-aligned Topic Model for Product Aspect Rating Prediction. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, (pp. 1192-1202). Doha, Qatar.

Wang, H., Lu, Y., & Zhai, C. (2010). Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. *KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 783-792). Washington D.C., USA: ACM.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 180-182.

Zeileis, A., & Hornik, K. (2007). Generalized M-Fluctuation Test for Parameter Instability. *Statistica Neerlandica*, 488-508.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, 492-514.

Zhang, J., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 694-700.