A business based on trust: the contribution of text mining as a big data technique to identify consumer insights to support the business of airbnb

Uma empresa baseada na confiança: a contribuição da mineração de texto como uma grande técnica de dados para identificar as percepções dos consumidores para apoiar o negócio da airbnb

DOI:10.34115/basrv5n2-024

Recebimento dos originais: 30/02/2021
Aceitação para publicação: 30/03/2021

Cláudia Pereira Ignácio Zuppo
Especialista, MBA Executivo em Big Data e Business Analytics, FGV
Instituição: Navega Pesquisa & Estratégia
Endereço: Avenida Doutor Francisco Ranieri no.182 cj.34 Lauzane Paulista, São Paulo/SP CEP 02435-060
E-mail: claudia.zuppo@navegapesquisa.com.br

Gustavo Corrêa Mirapalheta
Doutor, Adm.de Empresas, EAESP/FGV
Instituição: EAESP/FGV
Endereço: Rua Itapeva 474, 9º andar, Bela Vista, São Paulo/SP, CEP 01332-000
E-mail: gustavo.mirapalheta@fgv.br

João Luiz Chela
Pós Doutor, Computação, Unifesp, S.J. dos Campos/SP
Instituição: EAESP/FGV
Endereço: Rua Itapeva 474, 9º andar, Bela Vista, São Paulo/SP, CEP 01332-000
E-mail: joao.chela@fgv.br

ABSTRACT

This study uses a quantitative technique (Text Mining) coupled with a qualitative (subjective) approach to identify Airbnb's key supporting business factors. This coupled approach demonstrates the feasibility of using a Big Data technique side-by-side with a "humanized" approach, Letouzé (2014), when the author says that Big Data exists "by the people and for the people". To do so, it is used a database of nearly one million guests reviews of New York City properties made from 2008 through 2018. The adopted analytical approaches in Text Mining used the knowledge explored by Silge and Robinson, in their book "Text Mining with R" (2017), which includes the examination of Word Frequency, Bigrams, Sentiment Analysis using NRC Lexicon, Term Frequency Times Inverse Document Frequency (tf_idf), Zipf´s Law and Bigrams Network. All these analyses where developed from an historical perspective, since the Airbnb's reviews database divides its data in 5 points in time: 2009-2010, 2011-2012, 2013-2014, 2015-2016 and 2017-2018 (this last year up to June). Main results confirmed that Trust is a key issue for the development and growth of Airbnb business. Word Frequency and Bigrams prove that Airbnb success goes way beyond comfortable and well located, properties, but instead they are mainly a first-step towards the development of human relations based on

Trust. Learnings provide actionable insights both for Airbnb and for the segment of travel as a whole in their quest of conquering new customers in the future.

Keywords: Airbnb, Text Mining, Sentiment Analysis, Bigrams, Tf-Idf.

RESUMO

Este estudo utiliza uma técnica quantitativa (Text Mining) associada a uma abordagem qualitativa (subjectiva) para identificar os principais factores empresariais de apoio da Airbnb. Esta abordagem acoplada demonstra a viabilidade de utilizar uma técnica de Big Data lado a lado com uma abordagem "humanizada", Letouzé (2014), quando o autor diz que Big Data existe "pelo povo e para o povo". Para o fazer, é utilizada uma base de dados de quase um milhão de revisões de propriedades da cidade de Nova Iorque feitas de 2008 a 2018. As abordagens analíticas adoptadas em Text Mining utilizaram os conhecimentos explorados por Silge e Robinson, no seu livro "Text Mining with R" (2017), que inclui o exame de Frequência de Palavras, Bigrams, Análise de Sentimento utilizando Léxico NRC, Frequência de Documentos Inversa de Tempos de Frequência de Termos (tf_idf), Lei Zipf e Rede de Bigrams. Todas estas análises foram desenvolvidas numa perspectiva histórica, uma vez que a base de dados de revisões da Airbnb divide os seus dados em 5 pontos no tempo: 2009-2010, 2011-2012, 2013-2014, 2015-2016 e 2017-2018 (este último ano até Junho). Os principais resultados confirmaram que a confiança é uma questão-chave para o desenvolvimento e crescimento do negócio da Airbnb. Frequência de palavras e Bigrams provam que o sucesso da Airbnb vai muito além de propriedades confortáveis e bem localizadas, mas, em vez disso, são principalmente um primeiro passo para o desenvolvimento de relações humanas baseadas na Trust. As aprendizagens proporcionam conhecimentos accionáveis tanto para a Airbnb como para o segmento de viagens como um todo na sua busca de conquista de novos clientes no futuro.

Palavras-Chave: Airbnb, Text Mining, Sentiment Analysis, Bigrams, Tf-Idf.

1 INTRODUCTION

Why Airbnb?

The main trigger for this work is the need in explaining why Airbnb is considered a unique kind of business. In 2018 Airbnb celebrates its 10th Anniversary. Back to 2008, the company founded by Brian Chesky and Joe Gebbia, two young designers, was nothing but a couple of air mattresses offered to visitors of a Design Convention in San Francisco, California, who had no place to spend the night, due to the high demand of guests and low offer of hotel rooms available in the city at that time. Ten years later, this couple of air mattresses, which would value no more than U$ 50,00 each in 2008, turned out to be a U$ 38-billion business, according to Forbes, in May 2018, with 4,850 global listings offered in 191 countries.

A second reason to support the choice of Airbnb as a case study here is its remarkable contribution to the current Era of Sharing and Accessibility business.

According to Brad Stone, in his book "The Upstarts: How Uber, Airbnb and the Killer Companies of the New Silicon Valley are Changing the World" (2017), Airbnb is today the biggest hotel company in the world, but without having one single room in its possessions. Stone also highlights that companies as Airbnb and Uber represent the third phase of the internet history (the era post-Google and post-Facebook), because they "allowed the digital realm to expand into the physical world". If, for so many years, people could interact on digital platforms, with initiatives as Airbnb, people started personal interaction, due to digital facilities. And interaction is key to Airbnb. According to the official Airbnb site, one of the upsides of staying in an Airbnb, rather than in a typical hotel, is that guests can "live like a local". Besisdes that, the Airbnb's initiative has provided earnings improvements to ordinary citizens: according to Airbnb official information, 43% of hosting income generated by the business is used to pay for regular household expenses, not to mention that 6% of hosts used their Airbnb income to start a new business.

Why Text Mining and Airbnb?

By reading a lot of content from Airbnb creators, mainly Brian Chesky, it is possible to understand the Airbnb's customers reviews Text Mining helping potential. According to Chesky (2016), one of the key factors to run a successful business is to have mentors, and for that, you have to be shameless. He mentions that most people tend to be afraid of asking for others people's feedback. But Feedback is exactly what Airbnb reviews are all about: more and more guests, day after day providing their opinions on their stays and experiences, and, by doing so, influencing more and more future guests.

Another significant input from Chesky about a business is the importance of using the right information sharing methodology. For him, the best way is to start something with the perfect experience of just one person. In Airbnb´s case, one single guest. After making this single person live such perfect experience, the next challenge is to try to scale it, that is, try to make more and more people produce the same perfect experience the very first person has lived. And this is exactly what Airbnb reviews do. After a "perfect experience" during a stay, a guest writes it down so that everyone else can access it, and then the guests-to-be will try to repeat it. Stone (2017) describes this "perfect experience" with the example of the first guest Brian Chesky and Joe Gebbia received in their house, as the first guest from Airbnb in history. Their first guest was a guy called Amol Surve, a design student from India, who stayed for five days in an inflated airbed in their house, during a Design International Conference. What is inspiring about that, which in fact proves how Trust (and therefore the guests´ reviews), is a key issue of Airbnb, is that, while staying in their house, Amol Surve saw his own picture in one of the slides of a presentation Chesky was preparing about his home-sharing servisse in a designer's innovation event. The fact of having the first guest of Airbnb as the key-slide in that presentation, was a huge signal that guests´ reviews were going to be a key factor to the business success.

2 OBJECTIVE AND STUDY OBJECT

2.1 THE OBJECTIVE

This study's objective is to identify key Airbnb's business success factors from Airbnb New York 2009 – 2018 guest's reviews database.

2.2 QUANTITATIVE APPROACH

This study used in its analytical approach the techniques described in Silge and Robinson (2017). The database is composed of almost one million reviews, split into five periods of time, (2009-2010, 2011-2012, 2013-2014, 2015-2016, 2017-2018, up to July). Key text mining techniques used were Word Frequency, Bigrams, Sentiment Analysis using NRC Lexicon, Term Frequency Times Inverse Document Frequency (tf_idf), Zipf´s Law and Bigram Networks. Another example that employs analytical methodologies is CHENG and JIN (2018). They explore a database of 170.000 Airbnb's reviews from Sydney, Australia. In their work, they identified as key business success factors: location, amenities and the role of the host.

2.3 QUALITATIVE APPROACH

The qualitative side of this study aims to support Letouzé (2014), whom said that Big Data is something made by the people and for the people. When read in the context of Airbnb's business model, this can be understood as "Airbnb reviews are made by the community of guests, interfering in the decisions of future guests, and, therefore, for the people". To achieve this, opinions were, extracted, quoted and throughfully analyzed.

3 METHODOLODY

3.1 DATABASE

The database used here (http://insideairbnb.com/get-the-data.html) shows 980,754 reviews, produced from Airbnb guests in New York City, from January 2009 to July 2018. A small sample of this dataset is in Table 1, which also presents its structure.

Table 1: Variables referring to a Review

	Variable	**Description**	**Example Data**
1	Listing_id	Listing's identification	2515
2	Id	Review's identification	1083
3	Date	Date that review was published	3/25/2009
4	Reviewer_id	Reviewer's identification	9759
5	Reviewer_name	Reviewer's name	Cem
6	Comments	The actual review contente	I just got back from a trip to NYC during which I stayed at Bill's...

Table 2 and Graphic 1 show the written reviews speed increase in this ten years period.

Table 2: Amount of Airbnb Reviews in New York City – Evolution 2009 - 2018

Year	**Number or Reviews**	**Growth of number of reviews % vs previous year**
2009	234	-
2010	1468	627%
2011	5098	347%
2012	11534	226%
2013	24260	210%
2014	50,265	207%
2015	108189	215%
2016	206841	191%
2017	342268	165%
2018 (Jan to July5)	230597	67%
2018 estimate full year	452516	196%

Graphic 1: Number of Reviews Airbnb NY – Evolution 2009 to 2018

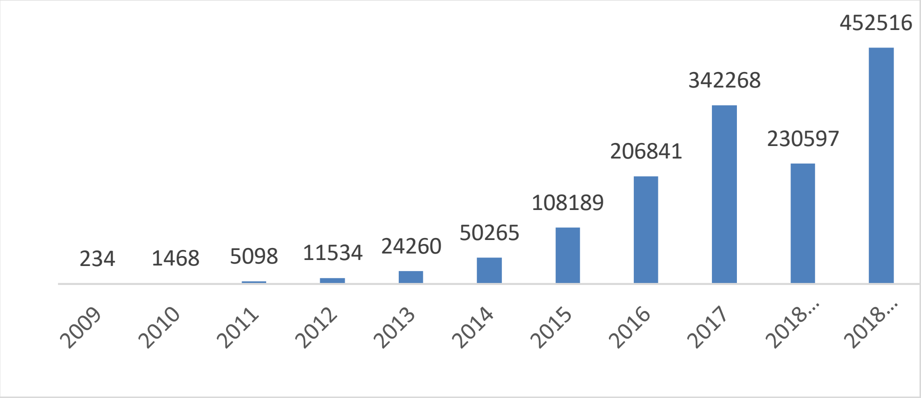
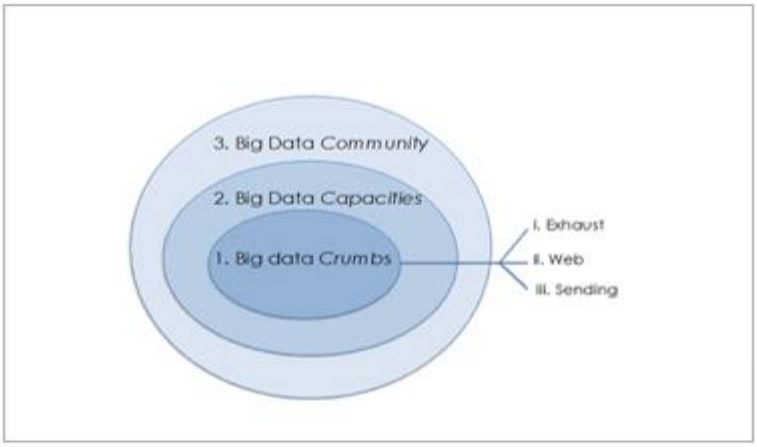


Table 2 and Graphic 1 provide a clear example of Big Data’s “3 Cs” Letouzé (2014) as opposed to Big Data’s classic “4 Vs” (Volume, Variety, Velocity, Veracity). While each of the 4 Vs refer to Big Data’s descriptive characteristics, the 3Cs approach add to the understanding of their uses and contributions to the society and the businesses, bringing a more updated view.

By defining the 3Cs as “Crumbs”, “Capacities” and “Community”, one can adapt this approach to Airbnb’s NY reviews database by first matching “Crumbs” with “breadcrumbs” (the reviews themselves), produced by Airbnb guests when visiting NY. Opinions about locations, cleanliness, comfort, and hosts hospitality levels, among other variables, are not routine and ordinary information anymore. As ordinary breadcrumbs, reviews are opinions created about travellers´ stays and therefore a signal of the experiences guests have “spread” along the way.

Next, one can match “Community” with the whole world of “crumbs” production. In fact, this is exactly what happens here: people from everywhere, staying in NY, and producing their “breadcrumbs” (reviews) along their stay. Finally, the third “C” (Capacities), can be seen here as tools, methods, softwares and other techniques that provide patterns and insights. Figure 1 below shows the systematization of Letouzé´s 3Cs theory on Airbnb’s NY reviews database.

Figure 1: The 3Cs of Big Data as an Ecosystem, Letouzé (2014)



4 ANALYSIS

The reviews database split in five time periods provided a better managerial perspective. Since the year 2008 had only one review, it was taken out of the study. Therefore, the analysis made in this study will show a five point information evolution timeframe (Table 3):

Table 3: Structure of 5-point period throughout this study

Years Analysed – Reviews Airbnb NYC				
Period 1	*Period 2*	*Period 3*	*Period 4*	*Period 5*
2009-2010	2011-2012	2013-2014	2015-2016	2017-2018
Jan-Dec	Jan-Dec	Jan-Dec	Jan-Dec	Jan-Jul 5

Silge and Robinson (2017) outline the R package tidytext importance and Wickham (2014) show the Tidy data structure importance in simplifying a Text Mining analysis (Table 4). In it, each variable is a column, each observation is a row and each observational unit is a table.

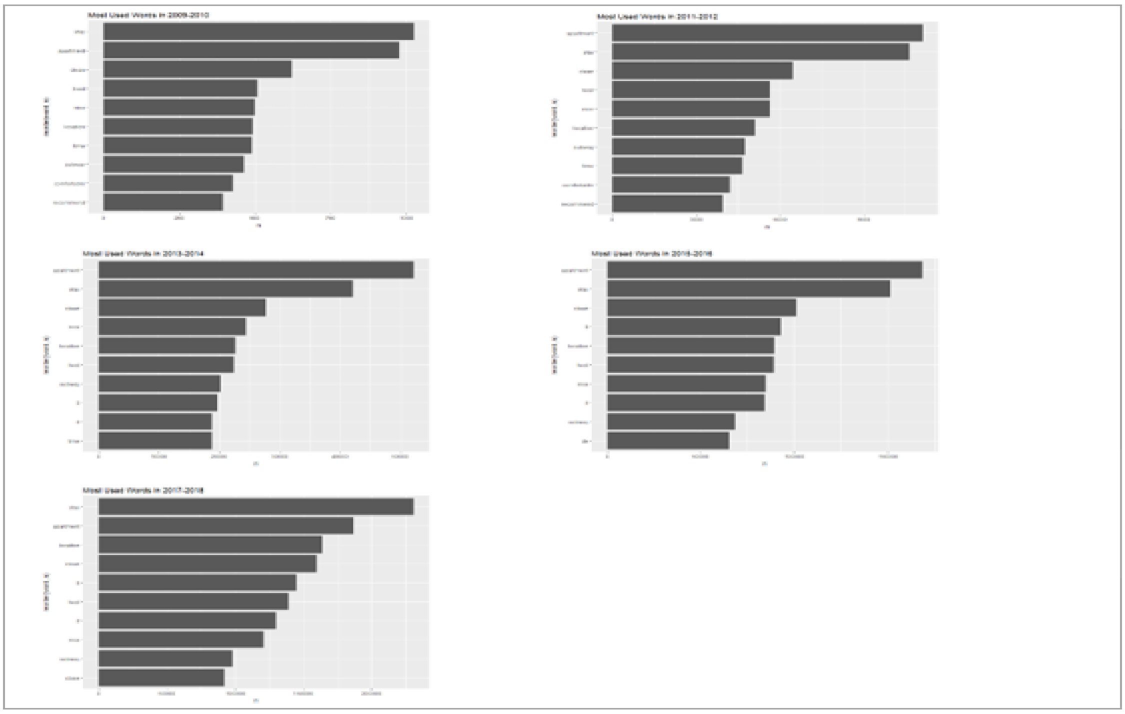
Table 4: A small sample of the dataset here used – reviews Airbnb NYC

Listing ID	ID	Date	Reviewer ID	Reviewer Name	Reviews/Comments
2515	859	08/03/09	8455	Roland	Such a wonderful place and very close to the metro station on the line 123. She was awesome with us, it was a pleasure to stay in her apt.
5441	903	12/03/09	8512	Stephen	I had an excelent experience with Kate. The place was comfortable and she is very easy to talk to. She can be a great help with getting around NY and what are the good restaurants, nightlife, etc.
2515	1083	25/03/09	9759	Cem	I just got back from a trip to NY during which I stayed at Stephanie´s apartment. The bedroom and private bathroom that comes with it were both kept very clean and are spacious so that I could unpack and put my clothes in the closet. I did not have to live out of my rucksack for ten days which made me comfortable.
5441	1231	02/04/09	9703	Jimena	Kate was the best host we could wish for, she´s been living in NY for a long time so she has lots of recommendations. The breakfast was spectacular, so we could skip straight to dinner. She´s such an interesting person to talk to, we did not want to leave. The location was excellent, best of both worlds, Hell´s Kitchen with inexpensive ethnic restaurants, dive bars and Times Square at equal distance.
5172	1261	07/04/09	7778	Bill	Wonderful location – warm and friendly.

4.1 ANALYSIS – WORD FREQUENCIES

In Text Mining, one of the first analysis done is the identification of the most common words in the dataset. Here there is an analysis split by timeframe as the database itself.

Graphic 2: Set of Ten Years of Track Reviews – Airbnb NY – 2009 to 2018



The 10 years Airbnb NY reviews history shows that “stay” and “apartment” are by far the two most common words in any year. The third most common word is “clean” in most periods, with the exception of 2017-2018 where it is replaced in the ranking third position by “location”.

Right after the word “clean”, the presence of the word “host” in the first four years (2009 to 2012), but not anymore in the following years (from 2013 to 2018) opens the hypothesis that, in Airbnb’s beginning the concept of “super host” could have been more important than the location itself, being later replaced by “location”.

If this trend continues in the future, this could indicate that business had to prove itself by having super hosts first, and then offer other facilities. This makes sense if one remembers that Airbnb is a business based on Trust. Nothing could be more clarifying than having guests providing good references on hosts to express their trust in a listing, and this positively interfering in others guests future choices, ZHANG and YAN (2018) and Table 5 (below).

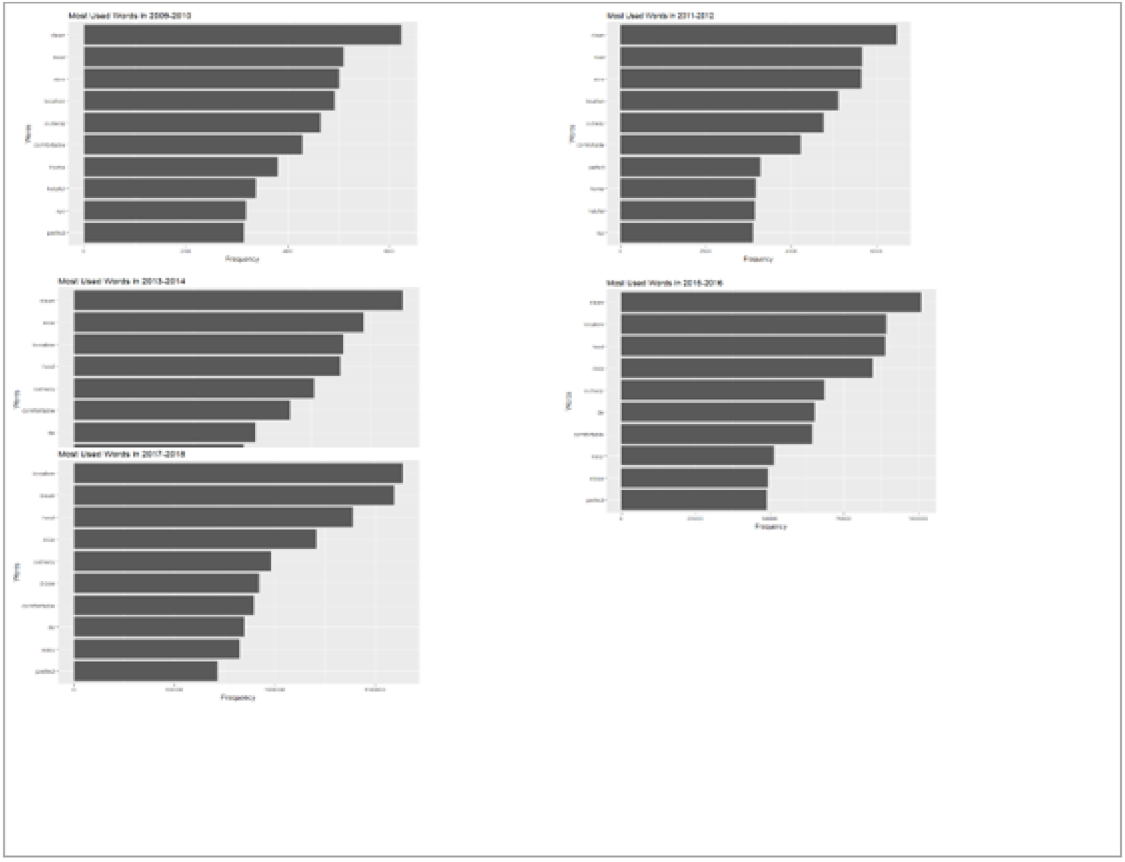
Table 5: Examples of Comments on Trust based on Reviews about Hosts

Listing ID	Date	Reviewer Name	Reviews/Comments
5803	03/06/09	Cam	Great place, great garden, friendly cats and **nice host.**
9357	04/01/10	Niall	We stayed for a week at New Year´s Eve and really enjoyed Laurelle's place. Very centrally located, nice and warm with a

			really good shower (very important). Laurelle was a **very nice host.** Would gladly stay again
271694	03/12/2013	Simona	First of all James is **a great host**, very friendly and easygoing. The studio is very nice, in a super safe area, close to everything. I had a great time and I would definitely recommend it to anyone!
51485	11/12/2013	Mike	Great room, very comfortable for two, and great location. **Great host** as well. .
805218	11/06/2016	Micky	I only stayed for a week with Leah and Reed but they made me feel so welcome and really made the move to New York just so much easier. The room was huge, everything was clean, they made space for my things in the bathroom and kitchen. They were **great hosts!**

Graphic 3 (below) shows a different approach in word frequency, by eliminating stop words (which will disregard “stay” and “apartment”) and “noise” (words like “ã” or “recommend”) from the reviews. In this scenario, the rankings are:

Graphic 3: Word frequency disregarding stop words



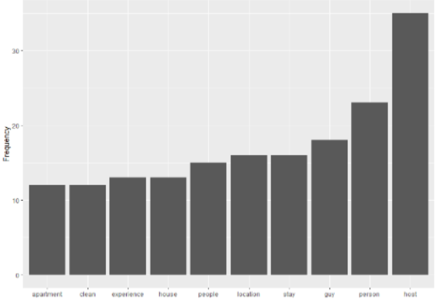
In this ranking version appears the word “clean” (except in the last period). While the word “clean” has a clear interpretation (the apartment cleanliness), the word “nice” can refer to either the place itself, the host, the location, the stay, or even all of them, as can be seen in Table 6:

Table 6: Random Reviews that include the word “nice”

Listing ID	Date	Reviewer Name	Reviews/Comments with the word “nice”	Word “nice” is related to...
6848	14/06/09	Mike	Allen and Irena were simply wonderful hosts and ambassadors to NY and Brooklyn. Just a really **nice place** and really **nice people**.	Apartment and people
3330	04/01/10	David	I haven´t met Julia but **her place is very nice** and comfortable. The location is great – just a 15 min train ride to Manhattan. Within walking distance, there are many shops, restaurants and supermarkets.	Apartment
978615	07/10/13	Tanya	It was important to us to find an affordable place fast, because we had to stay one more night in Astoria, so we called Rosa. She answered us very fast, the communication was excelente and that helped us a lot! **The location is really nice**, so you have everything you need in about 5 minutes walk.	Location
3404668	19/04/15	Alejandra	What a pleasant experience! This is the second time I use Airbnb as a guest and I have to say my wasn't that good, but Carl is just a gentleman! He helped us with our suitcase when we checked in and he was very acomodating about the checking in time. The **place is very nice and clean.**	Apartment
2019775	16/07/17	Kimberley	Thank you so much to Brad and Natasha for having us to stay, you made us feel very welcome straight away, even **inviting us for dinner with their friends on several occasions. Which is nice**, as me and my partner were travelling so it meant we got to meet a few more people.	Situation

Table 6 above raises one of the key-questions in this study: how dependent on Trust, and, therefore, on hosts, the business of Airbnb is. To clarify this issue, the N-grams analysis is very useful, since it allows the identification of consecutive word sequences. Here, it is calculated how often the word “nice” precedes words like “host”, “stay”, “location”, “house” or “apartment”, in such a way that one can build a relationship model between them (Graphic 4):

Graphic 4: Ranking of bigrams between the words most related to “nice” – 2009 to 2018



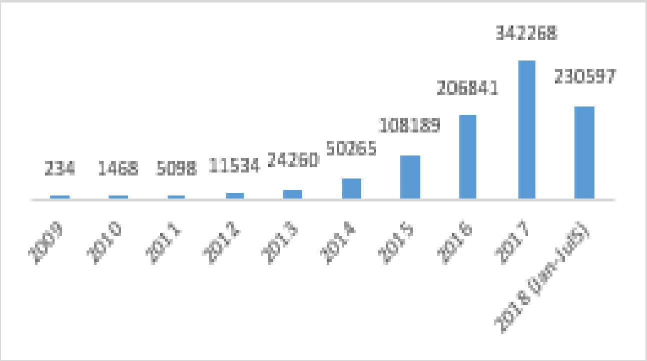
Graphic 4 above proves one of the hypothesis here created: the business of Airbnb is founded on trust, more than anything else. And that explains why the Top3 relations of the word “nice” are “host”, “person” and “guy”, which means quite the same thing (= the host), as the samples randomly extracted and shown below:

Table 7: Examples of the bigrams of the word “nice” extracted from dataset

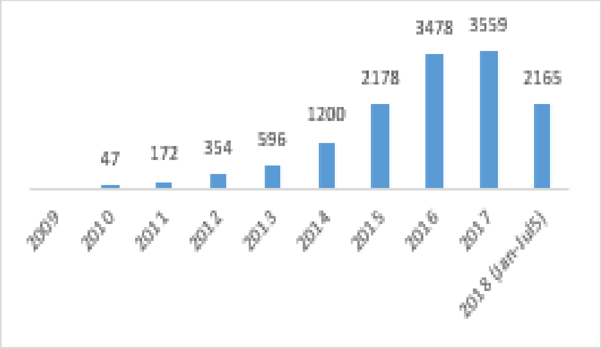
Listing ID	Date	Reviewer Name	Reviews/Comments with the word “nice”	Word “nice” is related to...
5803	03/06/2009	Cam	Great place, great garden, and **nice host.**	host
9668	12/10/2009	Carrie	**Nice host**. Honest and easy going. Would board with again.	host
9783	27/03/2010	Daniela	Sameer is a perfect host, incredibly helpful, very **nice person**, shares everything and thinks about what you might need before you ask for it. Liked the atmosphere. Highly recommended	person
24143	10/01/2011	Beth	Staying at Seth's was great! Seth is a super **nice guy** and has a fantastic apartment, complete with the most comfortable bed, EVER. Apartment is right around the corner from the L train. Highly recommended! Thanks!	guy
3310706	17/08/2015	Kylie	Robert was a really **nice guy** over all but the place wasn´t exactly clean when we arrived, but bareable. Some nights he was doing renovations that carried on to all hours through the night.	guy

Table 7 analysis makes clear that Airbnb creators, Chesky and Gebbia were right when they said that Airbnb’s business is a business based on Trust. Chesky believed that providing guests with an incredibly successful first experience with Airbnb was the guarantee that they would return (Stone, 2017). By comparing the amount of reviews that mention “first time” throughout the years, it can be seen a very similar behavior. In it, the amount of reviews grows but also grows the word “first time” mentioning. This proves that, even after 10 years in business, attracting new guests and providing them with a successful first experience is a key factor.

Graphic 5: Number or reviews per year



Graphic 6: Number of reviews with “first time in Airbnb”



Another example is in Table 8. In it, one can read some reviews mentioning the importance of “first time in Airbnb” as success milestone. Besides that, these reviews were reproduced here because they represent the “breadcrumbs” guests spread along their journeys. Bringing them are a signal of the power datasets have in bringing a “human profile” to Big Data (Table 8):

Table 8: Reviews referring to “first time in Airbnb”

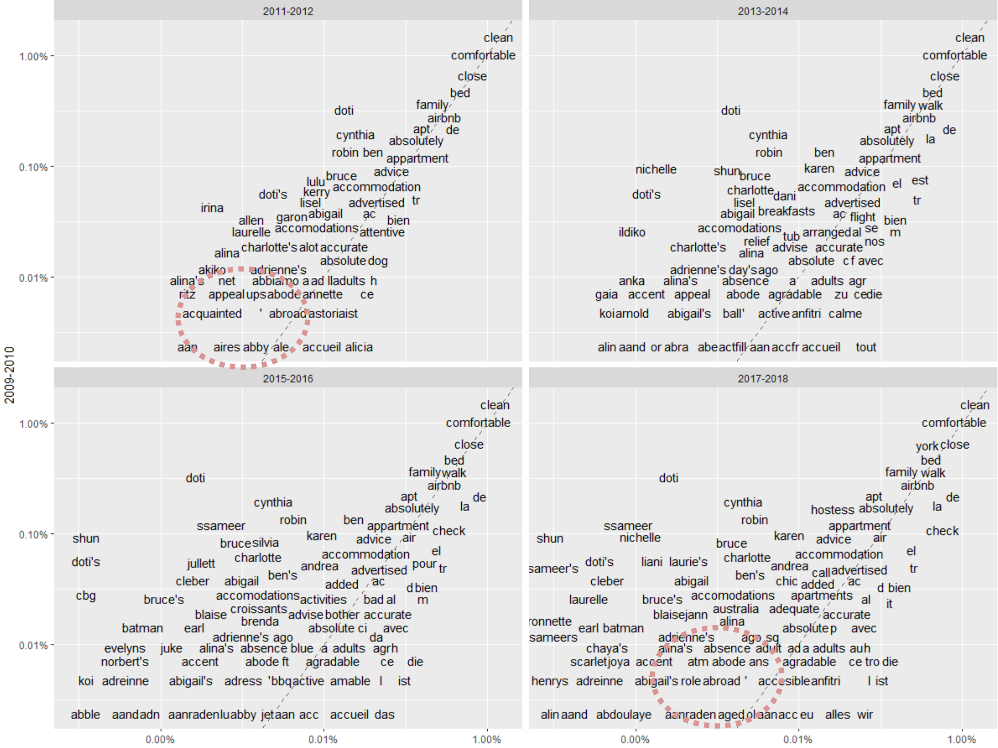
Listing ID	Date	Reviewer Name	Reviews/Comments
957642	03/04/14	Rafram	It was **my first time in airbnb, and I had some fears** about how it would be like. But it was so nice, and mostly comfortable and so cheap, comparing to other places. Thank you!
10132452	28/12/15	Yasin	This was **my first time in Airbnb and my first experience helped me build confidence in Airbnb**. Kevin's place was nice and as described. He was welcoming, helpful and friendly. He also kept the breakfast food for everybody. The neighborhood is good. We came back home late at night and nothing happened.
9806373	25/03/16	Katherine	**My first time in Airbnb and was a little nervous** about what it would be, but Grace was the best hostess! So sweet, friendly and very attentive. Greeted at home the first day and showed us the place, is as shown in the pictures, very nice, comfortable and warm for cold winter (it was like being at home). The day of our arrival my sister was happy for her birthday and Grace gave us a surprise to receive a cake, balloons and more! A beautiful detail. She did our trip the best, 100% recommended!!

4.2 COMPARING THE WORD FREQUENCIES BETWEEN PERIODS

Another analysis binds the periods together, Silge and Robinson (2017). Here there is an extension of the authors’s idea by comparing the first period studied to the following ones.

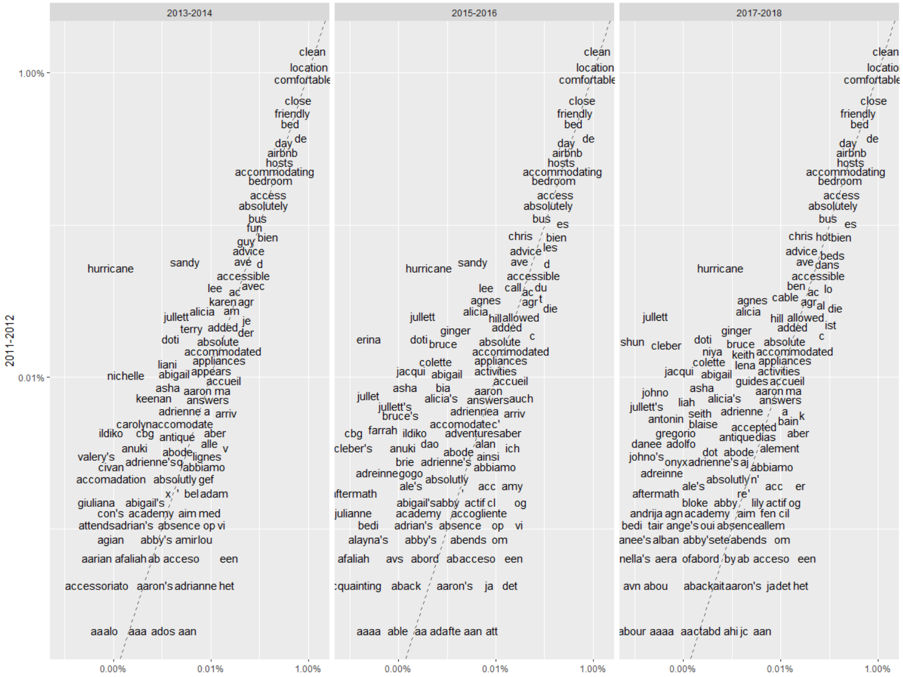
Words close to the 45^{o} line have similar frequencies in both periods analyzed (Graphic 7). Therefore, if one compares 2009-2010 to 2011-2012, he/she will notice the words “clean”, “comfortable”, “close”, “bed” and “family” as the most common ones in both periods. On the contrary, words that are far from the line are words found more in one period than in another.

Graphic 7: Comparison of the most common words 2009-2010 to the following periods



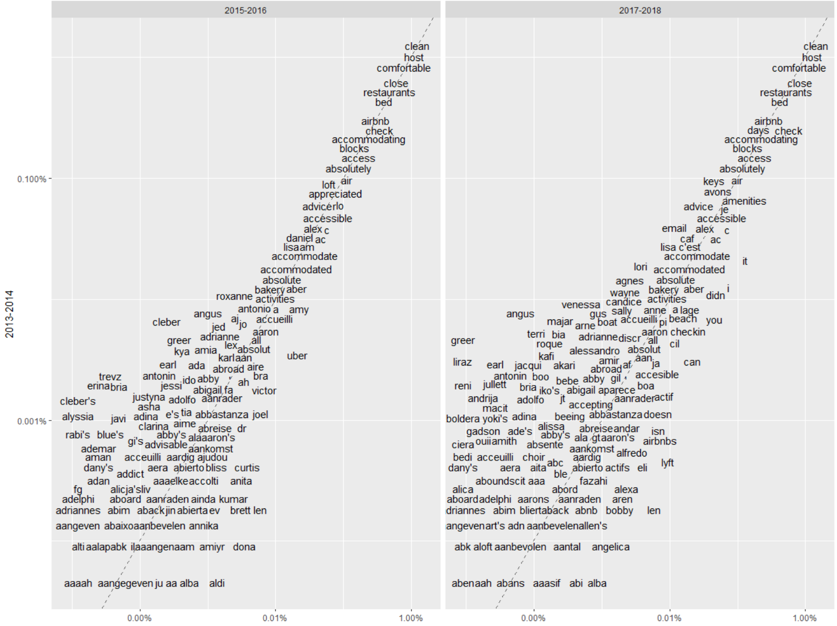
In Graphic 7 there is an empy space at low frequencies in the 2009-2010 versus 2011-2012 period, Silge and Robinson (2017). As time goes by, such empty space close to the low frequencies tends to disappear. Words that were in the beginning closer to the 45^{o} line, latter became a lot more “spread”. This indicates that in the years 2009-2010 and 2011-2012 there were more similar words than in 2009-2010 and 2017-2018. Not only that, but there are fewer data points in 2009-2010 vs 2017-2018, with the transitioning happening slowly.

Graphic 8: Comparison of the most common words 2011-2012 to the following periods



As the period of comparisons get closer, this effect become less and less marked, if compared to the first analysis 2009-2010 and each of the following years, as shown by Graph 9 below.

Graphic 9: Comparison of the most common words 2013-2014 to the following periods

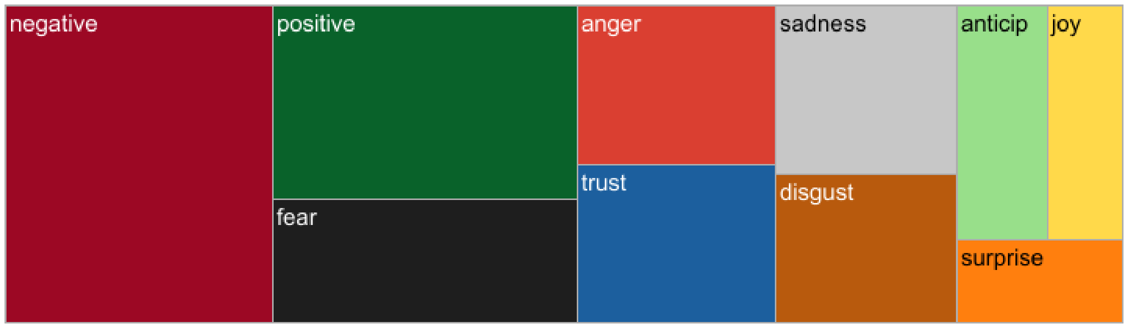


4.3 SENTIMENT ANALYSIS

With the help of sentiment lexicons, one can carry out the sentiment analysis in the tidy tool ecosystem, more specifically with the tidytext package, Silge and Robinson

(2017). The lexicon used here is the NRC (from Saif Mohammad and Peter Turney). It contains more than 6,000 words in English. Each word receives a sentiment score either positive or negative, as well as an emotion classification. The emotions classifications are anger, anticipation, disgust, fear, joy, sadness, surprise and trust (Graphic 10 and Table 9).

Graphic 10: Treemap showing Sentiments and Emotions from NRC Lexicon



Source:NRC Lexicon Treemap

Table 9: Number of words in each Sentimentand each Emotion in the NRC Lexicon

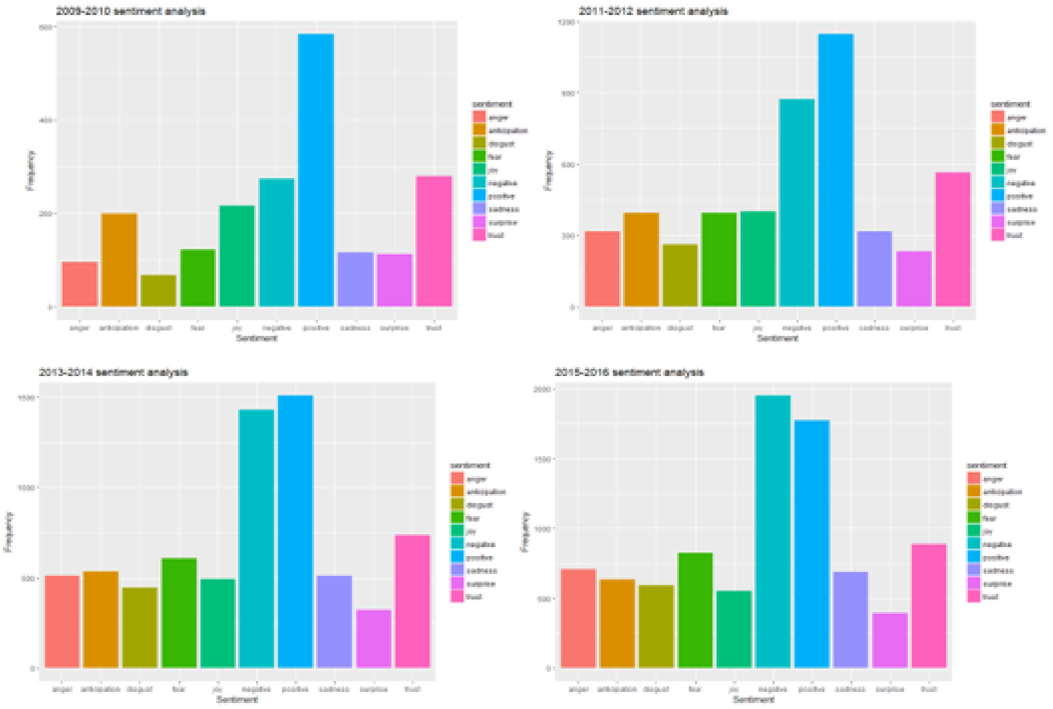
Sentiments	
Positive	2,317
Negative	3,338
Emotions	
Fear	1,483
Anger	1,250
Trust	*1,234*
Sadness	1,195
Disgust	1,060
Anticipation	842
Joy	691
Surprise	535

Source:NRC Lexicon

Based on unigrams, Airbnb’s 10-years reviews database sentiment analysis does not account for qualifiers before or after a word (Graphics 11). In the first two periods, the amount of positive sentiment is significantly higher than the negative sentiment. However, as time goes by, the amount of negative sentiment surpasses the positive ones.

The sentiment of Trust remains important throughout the period. The sentiment of surprise has the opposite behavior, being the lowest throughout the 10 years timeframe.

Graphic 11: Ten Years of Sentiment Analysis using NRC - Airbnb NY – 2009 to 2018



Below (Table 10) are the words that belong to the “Trust” Emotion (NRC Emotion Lexicon, https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm) and are in the Airbnb’s reviews dataset.

Table 10: Words under the classification of “Trust Emotion” from NRC Lexicon and found at Airbnb NYC database:

1.	Appreciated	8.	Hope	15.	Safe
2.	Beautiful	9.	Helpful	16.	Shared
3.	Comfortable	10.	Lovely	17.	Super
4.	Clean	11.	Love/loved	18.	Sweet
5.	Friendly	12.	Pleasure/pleasant	19.	Top
6.	Free	13.	Perfect/perfectly		
7.	Happy	14.	Pretty		

Again, the number of words in the Trust category proves once more the importance of it for Airbnb’s business model success. Another Airbnb’s dataset (from Boston, USA) shows the same relationship, TANG and McNICHOLAS (2017). The authors made a clustered binary data analysis, identifying three reviews clusters, two for positive and one for negative comments.

4.4 TF_IDF

Tf-idf (term-frequency times inverse document frequency) analysis “decreases the weight for commonly used words and increases the weight for words that are not used very much”, Silge and Robinson (2017).

This analysis shows both kinds of scores related to tf_idf: Table 11 shows words with low score tf_idf while Table 12 shows words with high tf_idf. As already expected, words with low tf_idf, that is, words very commonly used, are stop words. Table 11 lists the same low tf-idf words as Silge and Robinson report as low tf-idf words on Jane Austen’s novels from early XIX century. As a side commentary this finding shows the english language structure stability during the last 200 years period.

Table 11: Words with Low tf_idf score

	period	word
1	2017-2018	the
2	2017-2018	and
3	2015-2016	the
4	2015-2016	and
5	2017-2018	a
6	2017-2018	to
7	2017-2018	was
8	2015-2016	a
9	2015-2016	to
10	2017-2018	is

As opposed to Table 11 above, Table 12 shows words with a high tf_idf.

Table 12: Words with High tf_idf score

	period	word	n	tf	idf	tf_idf
1	2009-2010	marieka	19	0.000159	0.916	0.000146
2	2017-2018	automated	10020	0.000358	0.223	0.0000798
3	2015-2016	automated	6422	0.000311	0.223	0.0000694
4	2009-2010	marieka's	8	0.0000669	0.916	0.0000613
5	2009-2010	danette's	4	0.0000335	1.61	0.0000539
6	2009-2010	yaya	7	0.0000586	0.916	0.0000537
7	2013-2014	automated	1282	0.000222	0.223	0.0000496
8	2011-2012	hurricane	244	0.000188	0.223	0.0000419
9	2009-2010	orwell	3	0.0000251	1.61	0.0000404
10	2009-2010	seija	3	0.0000251	1.61	0.0000404

Graphic 12 shows the same kind of words as table 12, but separating them by year.

Graphic 12: Words with High tf_idf score

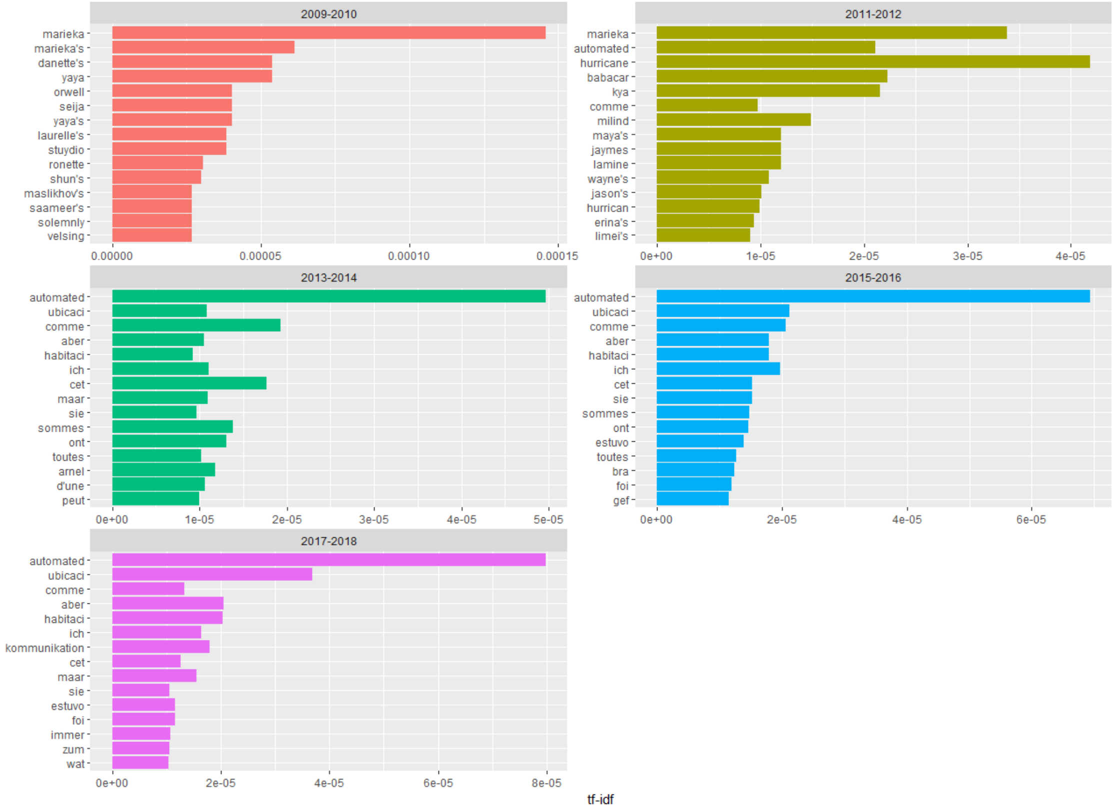


Table 13 shows a deeper look, presenting some reviews with high tf-idf score words:

Table 13 – Examples of reviews with High tf_idf scores

Listing ID	Date	Reviewer Name	Review	Word with hight tf_idf
443646	04/06/2012	Marsha	The reservation was canceled 123 days before arrival. This is an **automated** posting.	Automated
339386	30/07/2012	Maya	The reservation was canceled 2 days before arrival. This is an **automated** posting.	Automated
23501	15/04/2010	Michael	I really enjoyed my stay in NYC, and staying at **Marieka**'s was a big reason why. Her place is great, she made me feel very welcome, the room and bed are great, and it´s in a nice neighbourhood which is only a quick trip away from downtown. I highly recommend **Marieka's** place.	Marieka
26540	19/05/2010	Maria	The apartment is very very nice and the location is great. It is newly remodeled, clean and decorated with impeccable taste. I is ideal for a 1 or 2 people. **Danette** was very accommodating and helpful. Thanks so much!	Danette
212199	15/11/2011	Cristina	El apartamento es muy acogedor y limpio. Lo recomiendo por su **ubicación** en pleno corazon de Manhattan. Solo hay un pequeno inconveniente y es que está al lado de un tunnel y es un poco ruidoso. Gracias Danee.	ubicación
63657	05/07/2012	Stefan	Wir waren zu viert eine Woche in NY, die wirklich genau so aussieht, wie auf den Bildern.	aber (which

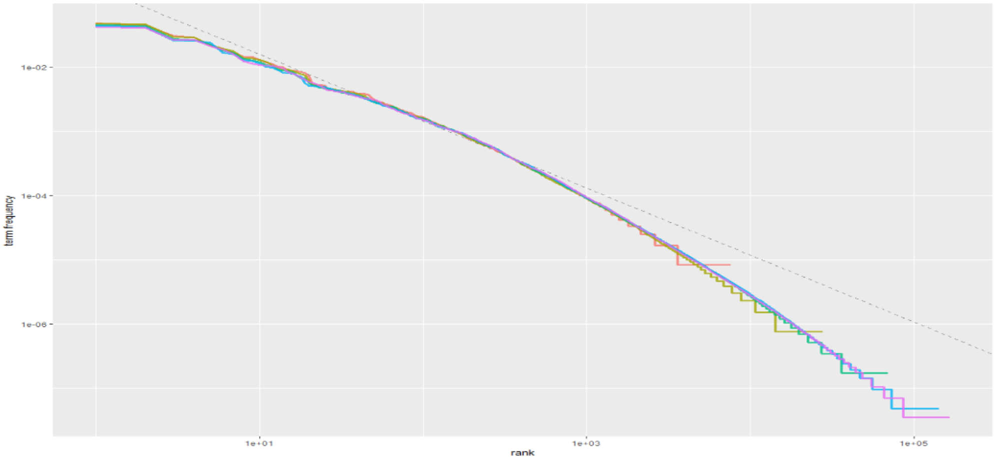
			Die Entfernung zur U-Bahn ist mit viel Gepack sicherlich eine Herausforderung, **aber** ohne wirklich gut zu machen.	means “but” in German)
242532	29/10/2012	Kristen	Jessica and her husband were wonderful hosts! Their house is absolutely stunning and the morning breakfasts were over the top. We stayed there the weekend **Hurricane** Sandy was set to arrive, and they said if our flight didn't make it out, we were more than welcome to return. What generosity!	hurricane

This closer look at the reviews on the words with high tf-idf scores, reveals that the words appear in very specific contexts: most of them refer either to an automated answer from the system, when a guest makes a reservation, but then cancels it, or to the names of hosts that are being thanked by guests for the good hospitality. Some other words with low frequency refer to reviews in other languages as Spanish and even German, which do not generalize.

4.5 ZIPF'S LAW

Moving forward in text mining analysis techniques, it is useful to highlight Zipf´s law. It says that the frequency that a word appears is inversely proportional to its rank. This relationship appears in all languages and all kinds of texts. Graphic 14 below shows the plot of the rank on the x-axis and term-frequency on the y-axis, on logarithmic scales. The plots are divided by timeframe and have each a different color.

Graphic 13: Zipf´s Law, the frequency of a word is inversely proportional to its rank

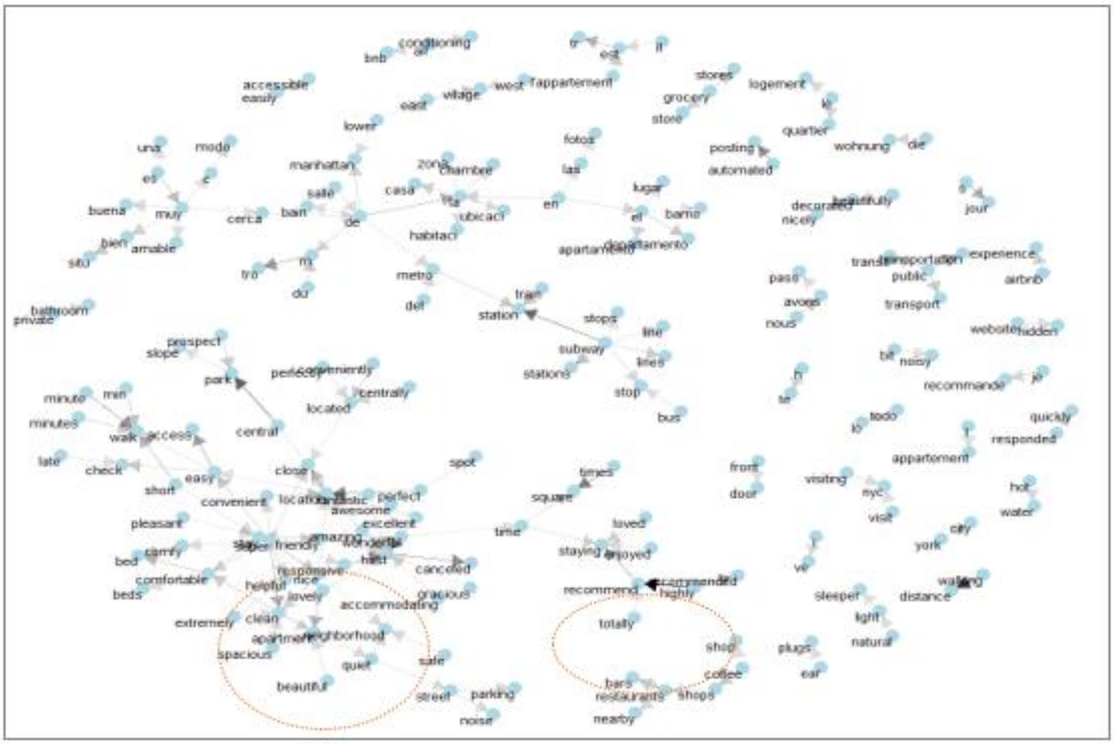


Graphic 13 shows a dotted line, which is the power law, that is, “the functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity” (http://en.wikipedia.org/wiki/Power_law), as suggested by Silge and Robinson (2014). In the present case the regression line equation is **y = -0.6225 + -1.1125x**. The comparison between Zipf´s Law and this dotted line shows a deviation at a high rank, showing that Airbnb database contains fewer rare words than predicted by a single power law. Meanwhile, the deviations at low rank are more unusual, as the database shows a lower percentage of the most common words than many language collections.

4.6 BIGRAMS NETWORK

The good thing about this analysis is the possibility of checking the relationship among words simultaneously, instead of the top few separated from the whole context. By doing so, all data receives a "relational" approach. Graphic 14 shows a reviews bigram network from the last two years (2017-2018). The graphic has an upolished look, but shows the most intense bigrams with a darker arrow in areas related to Trust. This "Trust Territory" relates either positive qualifications to the host (responsive, nice, helpful, lovely, friendly, etc) or final recommendations done to future guests. Other strong connections are the importance of location and easy means of transportation.

Graphic 15 – Network of Bigrams – Airbnb Database 2017-2018



5 DISCUSSION

The first objective of this study was identifing Airbnb's business key Airbnb's business success factors. A second objective was contribuing to Letouzé´s approach on

Big Data, by adding a more humanized input to a Big Data technique. While the first objective was done by the use of Text Mining techniques, as Word Frequency, Sentiment Analsysis, tf_idf, Zipf´s Law and Bigrams, the second one was reached by the selection of some guests´ reviews from the database that could illustrate each of the findings of the first objective. The study findings are insightful and they do not represent an end on this research line. By learning how dependent on Trust Airbnb is, this paper is just a contribute to the network business success factors understanding.

As bases for the development of this study, an extended range of references have been used: from technical ones, as Silge and Robinson (2017) book “Text Mining with R”, and Letouzé´s (2014) learnings about Big Data to more business contributions, as Stone (2017) book “The Upstarts: How Uber, Airbnb and the Killer Companies of the New Silicon Valley are Changing the World”. Such association of technical to more economical and society input have led to some key learnings.

Regarding economical input, it is relevant to quote here the work of HEO and BLENGINI (2018) A macroneconomic perspective on Airbnb´s global presence. The objective of this academic article was to explain the number of rentals available in capitals throughout the world, by finding the correlation between the number of rentals and several macroeconomic factors. Among such factors analysed were the countries´ degree of technological development, economic size, the size of tourism and productivity level, among others. Some of the key findings were that Airbnb is more popular in countries where the population is technologically wise, and that there is a negative relationship between GDP (Gross Domestic Product) per worker and the number of Airbnb listings. This is considered a clarifying article, as it indicates how productive and abundant applying analytical techniques to datasets can be.

From the technical side, the analysis of word frequency, first run with stop words and later on without them, has indicated that the word “host” is a key factor to the guests´ level of trust with the service. By analysing bigrams that come with such word, the study showed that bigrams as “nice host” or “nice person”, among others, are again a significant sign of the presence of Trust. Sentiment Analysis and the identification of the words under the classification of Trust from the NRC Emotions Lexicon and found in the nearly one million reviews from guests have also represented how relevant Trust is to the business. As synonyms of Trust, these words can be more and more used by Airbnb professionals (and hosts!) to bring Trust to the business, as the analyzes and examples presented in this paper have shown.

REFERENCES

INSIDE AIRBNB, ADDING DATA TO THE DEBATE - 2018. Available on: http://insideairbnb.com/get-the-data.html. Accessed on: Sept 09, 2018.
TEAM, T., As A Rare Profitable Unicorn, Airbnb Appears To Be Worth At Least $ 38 Billion – 2018 – Available on: https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion. Accessed on: Aug 14, 2018.

THE 100 LARGEST COMPANIES IN THE WORLD BY MARKET VALUE IN 2018 – 2018. Available on: https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-value/. Accessed on: October 10, 2018.

UNITED STATES CENSUS BUREAU – 2018. Available on: https://factfinder.census.gov/faces/nav/jsf/pages/. Accessed on: October 10, 2018.

SALA DE IMPRENSA DO AIRBNB - 2018. Available on: https://press.atairbnb.com/br/about-us/. Accessed on: October 10, 2018.

O QUE DIZ A PRIMEIRA REGULAMENTAÇÃO PARA APPS COMO AIRBNB – 2018. Available on: https://www.nexojornal.com.br/expresso/2018/01/24/O-que-diz-a-primeira-regulamentacao-brasileira-para-apps-como-Airbnb. Accessed on Sept.08, 2018.

THE SHARING ECONOMY ISN´T ABOUT SHARING AT ALL – 2018 – Available on: https://hbr.org/2015/01/the-sharing-economy-isnt-about-sharing-at-all. Accessed on: Nov.21, 2018.

WHAT IS AIRBNB´S CORPORATE SLOGAN? WHA DOES IT MEAN? - 2018 – Available on: https://www.quora.com/What-is-Airbnbs-corporate-slogan-What-does-it-mean. Accessed on: Oct.25, 2018.

GEBBIA, J., How Airbnb designs for trust – 2018 – Available on: https://www.youtube.com/watch?v=16cM-RFid9U – How Airbnb designs for trust. Accessed on Oct.25, 2018.

AIRBNB COMMUNITY DATA – 2018 – Available on: https://www.airbnbcitizen.com/data - Accessed on Oct.25, 2018

INTERVIEW: EMMANUAL LETOUZÉ, DATA-POP ALLIANCE ON DEMOCRATIZING THE BENEFITS OF BIG DATA – 2018. Available on: https://www.kdnuggets.com/2015/04/interview-emmanuel-letouze-democratizing-benefits-big-data.html. Accessed on: Oct. 26, 2018.

STONE, B. The Upstarts: How Uber, Airbnb and the Killer Companies of the New Silicon Valley are Changing the World - 2017

BRIAN CHESKY'S TOP 10 RULES FOR SUCCESS (@BCHESKY). Available on: https://www.youtube.com/watch?v=t-qjHG_q3d8. Accessed on: Oct.26, 2018.
NRC WORD-EMOTION ASSOCIATION LEXICON – 2018 – Available on:

https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. Accessed on: Sept.09, 2018.

CHENG, M.; JIN, X. Key attributes of Airbnb experience: Text-mining and sentiment analysis – 2018 – https://scholarworks.umass.edu/ttra/2018/Academic_Papers_Oral/23/. Accessed on: 30 Dec.2018.

SUNG, B; KYUNGMIN, L; HANNA, L; CHULMO, K. In Airbnb we trust: Understanding consumers’ trust-attachment building mechanisms in the sharing economy - 2018 - https://www.sciencedirect.com/science/article/pii/S027843191830327X. Accessed on: 21 Nov.2018

HEO, C; BLENGINI, I. A macroeconomic perspective on Airbnb´s global presence - 2018 - https://www.sciencedirect.com/science/article/pii/S0278431918304882. Accessed on Dec.09, 2018.

ZHANG, L; YAN, Q, A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb, 2018 - https://www.sciencedirect.com/science/article/pii/S0167923618301593?via%3Dihub#!. Accessed on Dec.02, 2018.

TANG, Y; McNICHOLAS, P, Clustering Airbnb Reviews – 2017 - https://arxiv.org/abs/1705.03134 - Accessed on Dec.30, 2018