Sentiment and Text Analysis of TripAdvisor Reviews

# Introduction

As a consumer it is sometimes difficult to navigate review websites to find meaningful reviews that help you determine if you want to make a purchase or not. Some reviews may be positive in nature but do not provide a lot of insight into why the reviewer had a positive experience.

By classifying reviews as positive or negative we can find features within those reviews that may help businesses drive new approaches to their products.

# Literature Review

In order to prepare for this project, I have reviewed some material regarding sentiment analysis and the caveats that go with opinion mining via review sites, forums and other online sources such as social media. The papers I mentioned below cover similar themes regarding some of the errors involved in drawing conclusions from human generated reviews. One paper in particular discusses the interesting point that sometimes review text does not match a review rating score (human nature to want to be more positive therefore score is more positive than textual review).

**Sentiment Analysis using Product Review Data by Xing Fang and Justin Zhan**:

http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

This paper covers some of the caveats of opinion mining using online data. For example, sometimes paid reviewers will give “fake” reviews and skew the opinions or results of any type of analysis. Also because people can freely post reviews as they like you cannot always guarantee that they are meaningful.

**Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? By Isa Maks and PiekVossen**

http://www.aclweb.org/anthology/R13-1054

This paper discusses weaknesses in using star ratings in opinion mining. They point out discrepancies between the review ratings and the actual review text itself (using some examples where the text was actually more negative than the actual rating score itself). This paper also discusses the effect on a reader when they read the review text versus when they see the ratings.

**Sentiment Analysis for Hotel Reviews by Walter Kasper and Mihaela Vela**

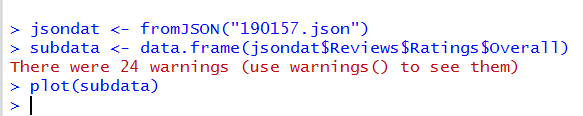
This paper discusses a point that is of particular interest to the data within this project. They cover the topic of considering ratings of a 3-star hotel and a 4-star hotel. Just because the higher end hotel received a lower rating on one attribute does not necessarily mean that the 3-star hotel is inherently better. The textual comments of the reviews would be of more interest to a hotel manager in that case to determine what the customer believed to be objectionable or positive

# Dataset

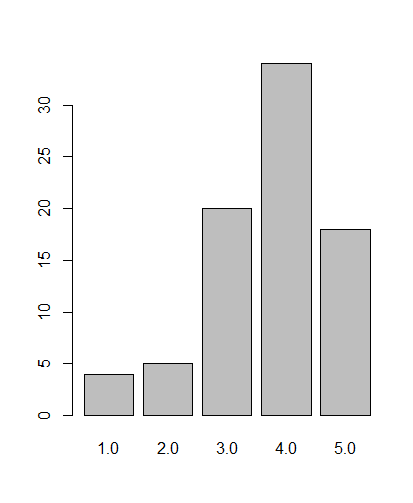
The dataset for this project will be the TripAdvisor reviews curated by Hongning Wang, Chi Wang, ChengXiangZhai and Jiawei Han for their paper; Learning Online Discussion Structures by Conditional Random Fields. The 34th Annual International ACM SIGIR Conference (SIGIR&#39;2011), P435-444, 2011.

It contains multiple trip reviews grouped by hotel. The hotels are scored on Service, Cleanliness, Overall Value and Location. The review contains the reviewer id, review content and the date the review was posted. The hotels are also rated more in depth for other items that I have decided not to include in this project due to timing; Check-in Rating, Business Services Rating, Overall Room Rating, Sleep Quality of the Room Rating. I have also decided to take a subset of the original 12000 hotels because my system did not have the resources to process that many JSON files and I ran into some issues with RStudio when the dataset was too large.

Loading 1 Hotel’s reviews into R just to take a preliminary look at the type of data available in the set;



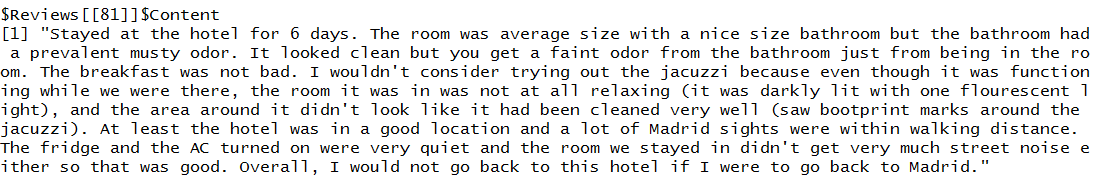
Showing preliminary look at how “Overall” ratings vary for 1 hotel – the “Overall” ranking is a value from 1-5 and the website does allow negative numbers (lowest being -1). I would like to use this rating as the basis for labelling data.



Below contains a list of the attributes found within the dataset;

|  |  |
| --- | --- |
| **Attribute** | **Data Type** |
| Service Rating | Numerical |
| Cleanliness Rating | Numerical |
| Overall Score | Numerical |
| Location Score | Numerical |
| Review Title | Text |
| Author | Text |
| Content | Text |
| Date | Date |
| HotelID | Numerical |
| HotelName | Text |

Below is a sample of the review content:



# Approach

Below is an outline of the steps I have taken to complete this project.

**Step 1: Clean and Explore Data**

The dataset is quite large so I have extracted a subset of 1580 reviews for this project. These reviews are from a cross section of 8 unique hotels from the original TripAdvisor dataset. The JSON data files were loaded directly into R using the jsonlite R package which converts the data into an R dataframe. Then all of the chosen hotel data were combined into one dataset to explore further within R.

I have included a plot of the distribution of the Overall Rating scores in the Results section of this document.

As part of the data exploration step I have also included word clouds of the most used words in the entire dataset and also two word clouds that show the most used words in the positive text corpus and the negative text corpus - for comparison purposes.

**Step 2: Label Data based on the user's "Overall Rating" score**

The hotel review has the option of giving an "Overall Rating" of the hotel experience on a 5 point scale. These values were used to label training review text as either positive, negative or neutral. I made the assumption that a 4 or 5 point value was a positive review. A score of 3 was a neutral review and that anything under a 3 was likely a mostly negative or very negative review (only used "negative" classification however).

An extra column was added to the original data frame called "Sentiment" based on the above assumptions. I was then able to split this overall data set into training and testing in the further classification steps of this project.

**Step 3: Extract Review text for Analysis**

The review text portion of the data frame was turned into a text Corpus via R. The code for this section is included in my project github repository if you would like to see how this was done.

https://github.com/sdyck03/finalProject/blob/master/Final-Code/R

The above file has annotated R code that covers everything that was done for this project in R.

A sentiment score function was written to take a block of review text and clean, parse and process it in order to score it as positive or negative. For the purposes of the classification steps I have removed classifying any neutral data due to timing, but this could be something to add for future iterations of this project - should there be any.

**Step 4: Naive Bayes Classifier - Training and Testing**

A portion of the overall data set was used in a Naive Bayes classifier by combining the tokenized text and appending a sentiment value. As discussed in the previous section all of the R code can be viewed via the following link;

https://github.com/sdyck03/finalProject/blob/master/Final-Code/R

The results of the classification will be included in the next section of this document.

# Results

The review text was extracted from the original data set and converted into a Corpus. Cleaning steps were used before analyzing the frequently used words.



As well as removing stopwords, the following words were also removed; "room", "hotel", "stayed", "get", "got", "also" since they came up frequently but given the context of the data they did not tell much of a story (all of the reviews were about hotels and rooms so those two words were removed to find more meaningful results).

All of the words were used at minimum at least 10 times throughout the text Corpus.

The following word cloud shows the most commonly used words in the final overall text data that was used.



Some of the topics that users discussed a lot can be determined from these word visualizations; "location", "breakfast", "staff". The nearby attractions, staff and the food available during their stay at the hotel were discussed frequently in all of the reviews that were studied during this project.

The following visualization includes the most commonly used words in the labeled positive data set:



Some of these words are similar to the overall visualization because the data set was predominantly very positive.

The most frequently used words in the labeled negative data set:



This gives some insight into the topics that are discussed at least in some of the negative reviews that were studied during this project; staff, location, breakfast, floor and room size.

From some of the word frequency analysis I was able to draw some conclusions about what people were discussing in the reviews and I've outlined those results in the tables that follow.

**Frequent Topics Covered Positive Reviews**

The following words were used more than 20 times in the positive reviews that were studied in this project. They cover multiple topics listed in the chart below;

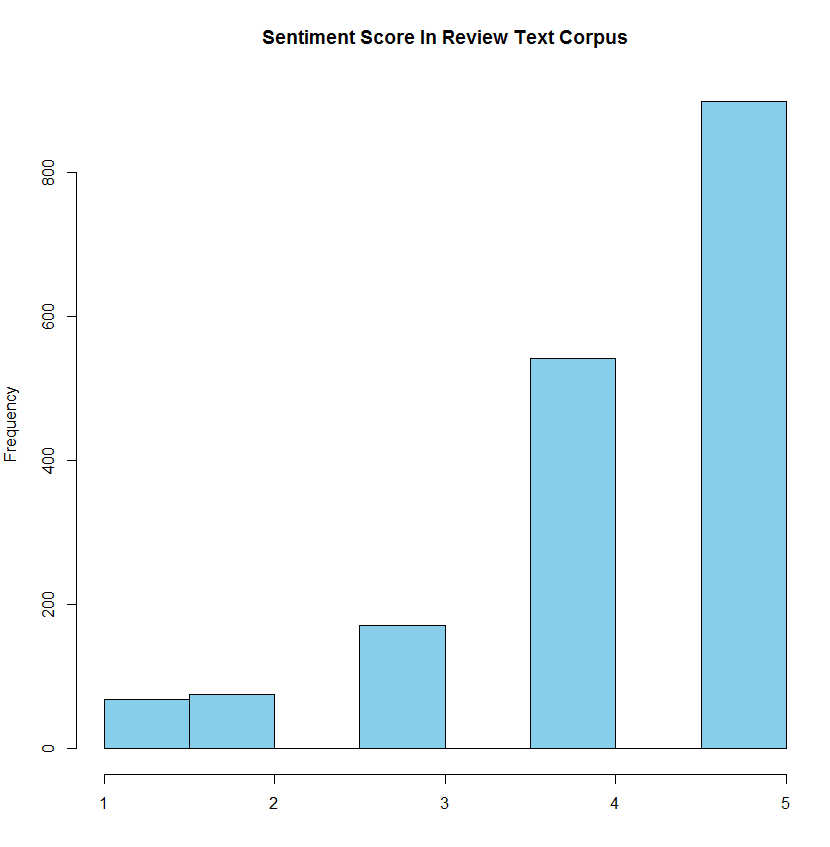
|  |  |
| --- | --- |
| **Topic** | **Frequent Words In Positive Reviews** |
| *Business Services Available* | Wi-fi, Internet |
| *Food and Drink* | Breakfast, Buffet, Champagne, Coffee, Dinner, Fruit, Meal, Menu, Restaurant |
| *Amenities* | Fitness, Attractions, Atmosphere, Bars, Areas, Location, Beach, Facilities, Market |
| *Room* | View, Bathroom, Bathrooms, Toilet, Toiletries, Sink, Sleep, Pillows, Shower, Bed, Balcony, Clean, Furniture, Décor, Sheets, Phone |
| *Staff* | Doormen, Housekeeping |
| *Costs* | Cheap, Cheaper, Price, euro, euros |

**Frequent Topics Covered Negative Reviews**

|  |  |
| --- | --- |
| **Topic** | **Frequent Words In Negative Reviews** |
| *Business Services Available* | Internet |
| *Food and Drink* | Food, Breakfast |
| *Amenities* | Lobby, Reception, Parking, Pool |
| *Room* | Bathroom, Desk, Dirty, Smell, Floor, Door, Noise, Experience, Booking, Shower, Toilet, Towels, View |
| *Staff* | Manager, Rude, Service, Valet, Staff |
| *Costs* | Expensive, Charge, Price |

We can determine that similar topics are covered in all reviews even if the content is regarded as having a different sentiment.

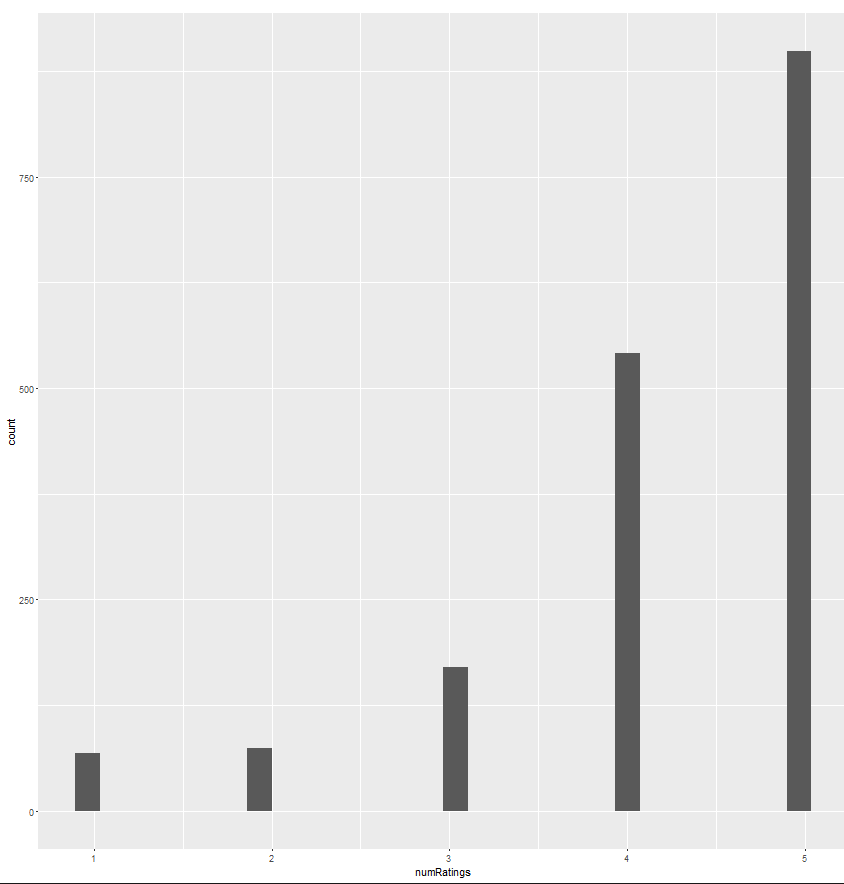
**A distribution of the labeled data based on "Overall Rating"**;



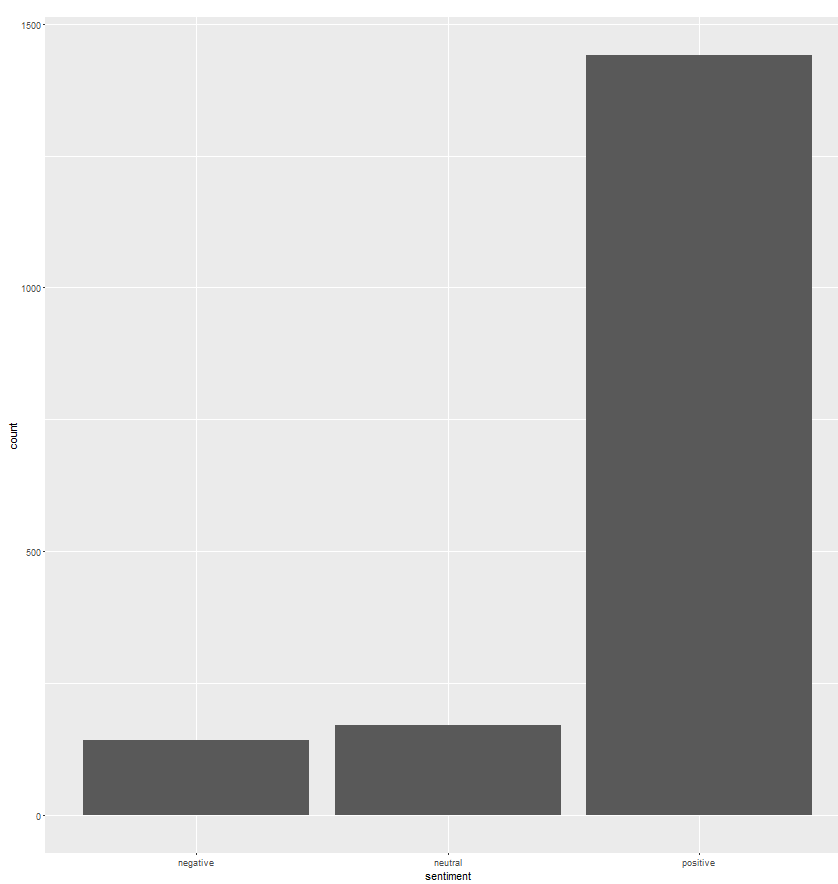
The above shows the distribution of the user's "Overall Rating" that was used to label the data set as either positive or negative as discussed in earlier sections of this document.

The following two plots indicate the overwhelming amount of positive data that was included in this dataset. I will make the assumption that the accuracy of my classification results will be off due to the skewed data set. An area for improvement for future iterations of this project would be to clean the data and obtain a more balanced training set.

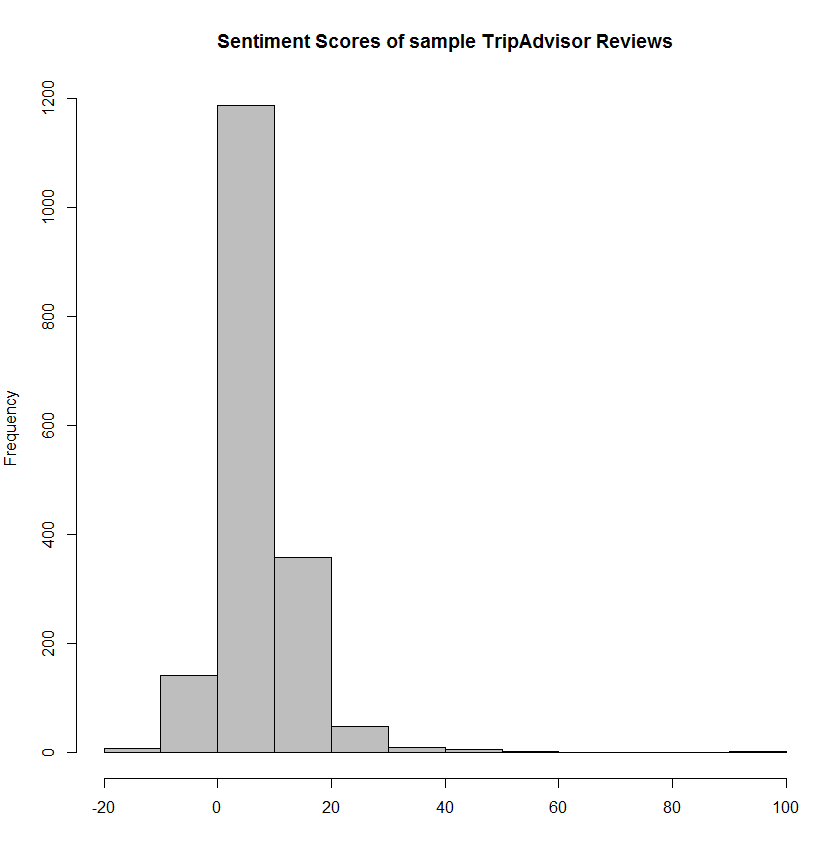
Below shows the Count of User's Overall Rating and as you can see by the two rightmost bars, the majority of the ratings were 4 or 5. The middle bar shows the count of 3's and the leftmost bars show the counts for 1's and 2's.



Below is the distribution of positive, negative and neutral data that was labeled based on the above ratings (Y = count, X = Sentiment of negative, neutral or positive).



The next plot shows a distribution of scores produced by the score sentiment function after processing the totaldata$content training data:



The above shows that most of the reviews showed a larger number of positive words found within the tokenized text. At most some of the review had approximately 100 hits for positive words and in some cases 20 or more negative words (lowest score sentiment sum of -20).

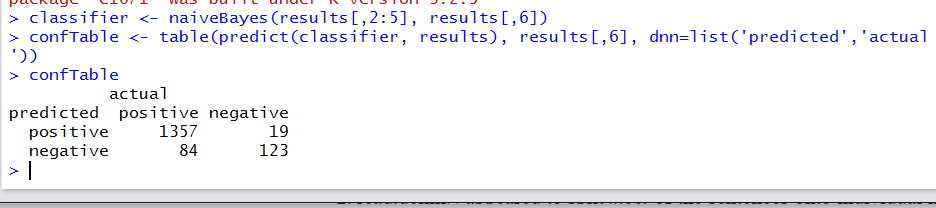
**Classification Results**

The labeled data was used to train a Naive Bayes classifier. The sentiment score function that was used in order to clean and process the data is included in the master code document that is found at the following link:

https://github.com/sdyck03/finalProject/blob/master/Final-Code/R

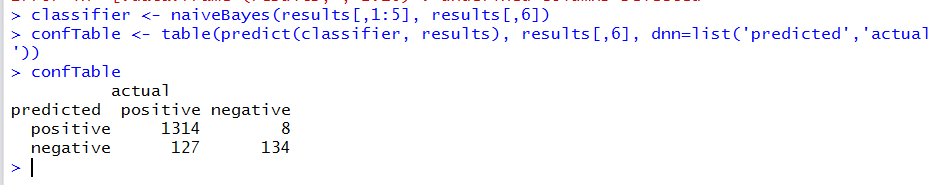
Below are the classification results using the labeled data from previous steps

Trial 1



From this test it seems that 19 of the 142 negative reviews are being classified as positive and 84 of the 1441 positive reviews are being classified as negative.

Trial 2 - More Data



It appears that it is classifying 8 of the 142 negative reviews as positive and classifying 127 of the 1441 positive reviews as negative. The accuracy for predicting negative reviews seems to have improved marginally but there is likely not enough data to make any proper conclusions about these results.

A future area for improvement for this project would be to continue to add more data to this "result" matrix and see if the accuracy improves. Also as discussed previously a more balanced training set would be more ideal.

# Conclusions

Based on the outcome of this project I can conclude a few items about online reviews;

1) The numerical rating scores do not always completely reflect the textual context of the review. In some cases a data point labeled as positive might have a more neutral rating score. And in some cases a negative rating score might be padded with more positive text. This could be an area to explore further.

2) The classification step of this project requires a lot more work to get more meaningful results due to the overwhelming positive data included in this data set. One thing I would do differently if I were to explore this project further is to see what results were obtained with a more balanced training data set.

From a business and marketing standpoint the area of Sentiment Analysis is quite useful but I would suggest based on some of these findings that Topic Modeling would be a next step to study within these reviews. A deeper understanding of the types of topics that are covered in reviews would help a business improve or course correct any issues that may be brought to light via online reviews. It should also be noted that due to human nature I believe more people try to say more positive things because they don't want to be seen in a negative light - therefore some of the review content still needs to be monitored outside of machine learning techniques to get a real feel for what the public is saying about the hotel experience.

# Next Steps/Areas for Improvement

Continuing to train the classification model further with additional reviews should improve the recall accuracy score achieved so far. However, based on some of the text analysis there is already a lot of information to go on to choose areas to explore further – such as topic modelling for example. Stretching this further into modelling topics within these reviews would help focus business intelligence even more as we would then be able to extract topical areas for business improvement or we could see in which areas we already excel in as a Hotel Manager or Hotel Corporation.

The term frequency analysis already shows some areas where reviewers have pointed out concerns and where they believe certain hotels are already excelling.

Another possible area for improvement within this project so far may be to find clusters of hotels where there are mostly negative reviews or mostly positive reviews and determine if there are any location correlations.

# References Used

Below is a summary of reference papers and websites that were used to assist with understanding the theory and code behind Sentiment Analysis using R as an analysis tool.

1. The score sentiment function was modeled after the one used in this blog post: <http://analyzecore.com/2014/05/11/twitter-sentiment-analysis-based-on-affective-lexicons-in-r/>

2. Another blog post with a similar sentiment score function: <http://andybromberg.com/sentiment-analysis/>

3. The positive and negative sentiment words used in the score sentiment function were used from this paper: Hongning Wang, Chi Wang, ChengXiangZhai and Jiawei Han for their paper; Learning Online Discussion Structures by Conditional Random Fields. The 34th Annual International ACM SIGIR Conference (SIGIR&#39;2011), P435-444, 2011.

4. Used to resolve an issue with jsonlite R package: <https://github.com/jeroenooms/jsonlite/issues/106>

5. Used to model method of analysis: <https://www.kaggle.com/amhchiu/whats-cooking/bag-of-ingredients-in-r/run/71436>

7. Used to model method of analysis and ideas for frequent word visualizations:

<http://www.r-bloggers.com/sentiment-analysis-on-donald-trump-using-r-and-tableau/>