Theme Extraction of Hourlie Marketplace Reviews

Introduction

5-star reviews are numerical reviews and it is considerably difficult to figure out as to what the users actually care about when they are assigning a numerical value to the thing they are reviewing. Text reviews, however, provide the opportunity to actually understand what exactly is it the users want from the service they are paying for.

For this analysis, I wanted to classify the reviews of the users into three different categories- Price of the Service, Quality of the product and service and Time of Delivery- and analyse the variation in these reviews in combination with the sentiment score and with other factors like the cost of the service, the expected time of delivery and location. This analysis will allow Buck Me Up a general idea as to what aspect of the service does the majority of the users care about and would thus allow them to better cater to the needs of the users by ranking the hourlies over different themes and would allow users to find a better service.

Project Workflow

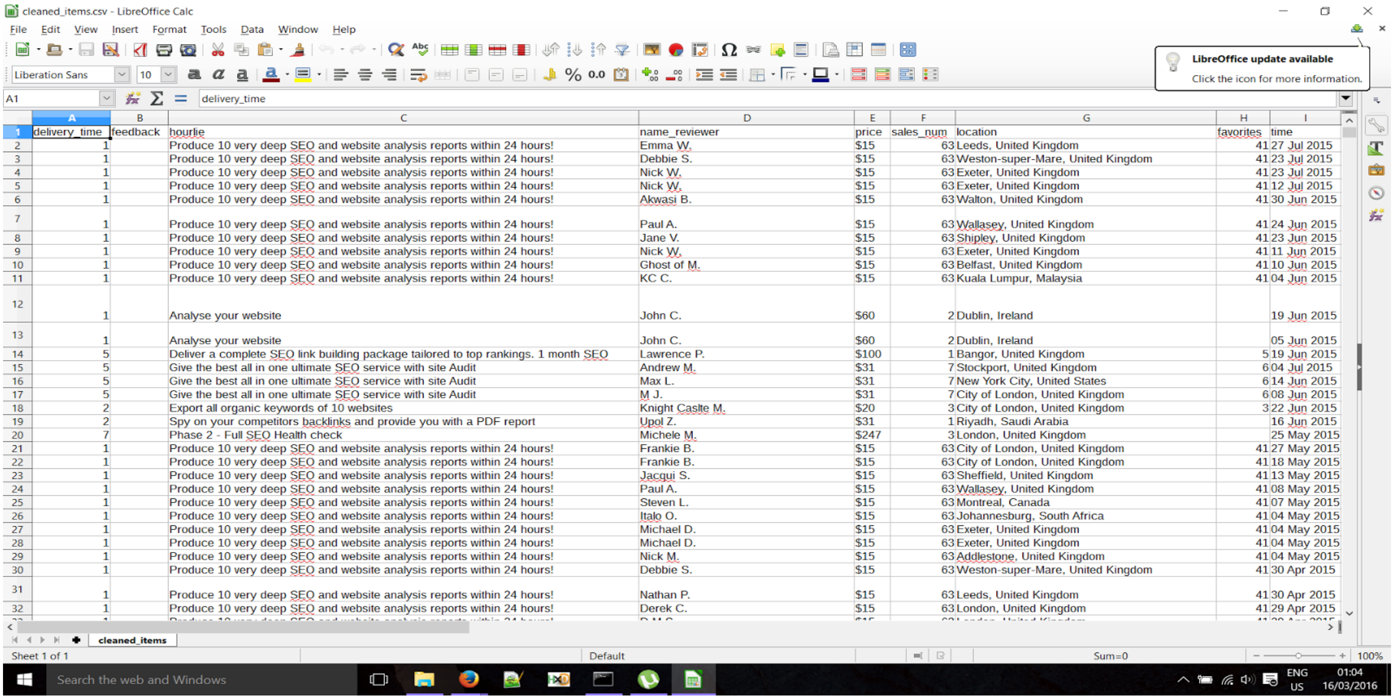
Data Gathering

Extracting and collecting the data was the first major and significant task as this whole project depended upon the availability and the quality of an appropriate dataset. The data I used for the analysis and classification was scraped from peopleperhour.com SEO Analysis sub theme. This was done as I felt that their reviews data was in the exact form I was looking for; reviews for hourlies based on single jobs which provided the necessary variation in the dataset.

I decided to use python and it's scrapy package to query the website and extract the reviews.

The major issue that I faced with gathering the data was with the automation of the scraping the reviews data from the website. The issue was that the peopleperhour domain was automatically redirecting our queries to the website to a different web domain which created a lot of issues initially. The way I solved it was by hardcoding all the links in the code for scraping.

Illustration


cleaned_items.csv - LibreOffice Calc
File Edit View Insert Format Tools Data Window Help
Liberation Sans
10
LibreOffice update available
Click the icon for more information.
A1
delivery_time
cleaned_items
Sheet 1 of 1
Default
Sum=0
100%
Search the web and Windows
ENG
US
01:04
16/03/2016

	A	B	C	D	E	F	G	H	I
1	delivery_time	feedback	hourlie	name_reviewer	price	sales_num	location	favorites	time
2	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Emma W.	$15	63	Leeds, United Kingdom	41	27 Jul 2015
3	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Debbie S.	$15	63	Weston-super-Mare, United Kingdom	41	23 Jul 2015
4	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	23 Jul 2015
5	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	12 Jul 2015
6	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Akwasi B.	$15	63	Walton, United Kingdom	41	30 Jun 2015
7	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Paul A.	$15	63	Wallasey, United Kingdom	41	24 Jun 2015
8	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Jane V.	$15	63	Shipley, United Kingdom	41	23 Jun 2015
9	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick W.	$15	63	Exeter, United Kingdom	41	11 Jun 2015
10	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Ghost of M.	$15	63	Belfast, United Kingdom	41	10 Jun 2015
11	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	KC C.	$15	63	Kuala Lumpur, Malaysia	41	04 Jun 2015
12	1		Analyse your website	John C.	$60	2	Dublin, Ireland		19 Jun 2015
13	1		Analyse your website	John C.	$60	2	Dublin, Ireland		05 Jun 2015
14	5		Deliver a complete SEO link building package tailored to top rankings. 1 month SEO	Lawrence P.	$100	1	Bangor, United Kingdom	5	19 Jun 2015
15	5		Give the best all in one ultimate SEO service with site Audit	Andrew M.	$31	7	Stockport, United Kingdom	6	04 Jul 2015
16	5		Give the best all in one ultimate SEO service with site Audit	Max L.	$31	7	New York City, United States	6	14 Jun 2015
17	5		Give the best all in one ultimate SEO service with site Audit	M J.	$31	7	City of London, United Kingdom	6	08 Jun 2015
18	2		Export all organic keywords of 10 websites	Knight Caslte M.	$20	3	City of London, United Kingdom	3	22 Jun 2015
19	2		Spy on your competitors backlinks and provide you with a PDF report	Upol Z.	$31	1	Riyadh, Saudi Arabia		16 Jun 2015
20	7		Phase 2 - Full SEO Health check	Michele M.	$247	3	London, United Kingdom		25 May 2015
21	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Frankie B.	$15	63	City of London, United Kingdom	41	27 May 2015
22	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Frankie B.	$15	63	City of London, United Kingdom	41	18 May 2015
23	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Jacqui S.	$15	63	Sheffield, United Kingdom	41	13 May 2015
24	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Paul A.	$15	63	Wallasey, United Kingdom	41	08 May 2015
25	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Steven L.	$15	63	Montreal, Canada	41	07 May 2015
26	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Italo O.	$15	63	Johannesburg, South Africa	41	04 May 2015
27	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Michael D.	$15	63	Exeter, United Kingdom	41	04 May 2015
28	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Michael D.	$15	63	Exeter, United Kingdom	41	04 May 2015
29	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nick M.	$15	63	Addlestone, United Kingdom	41	04 May 2015
30	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Debbie S.	$15	63	Weston-super-Mare, United Kingdom	41	30 Apr 2015
31	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Nathan P.	$15	63	Leeds, United Kingdom	41	30 Apr 2015
32	1		Produce 10 very deep SEO and website analysis reports within 24 hours!	Derek C.	$15	63	London, United Kingdom	41	29 Apr 2015

Data Cleaning

There were 2 major steps involved in the data cleaning process.

I. Cleaning and Conversion

The first step was to clean the reviews by removing the HTML tags and to convert them in a form which can be directly used for analysis without requiring any further processing.

A major problem that I faced was that after initially removing the HTML tags was that I was getting unrecognized character errors while trhying to read the reviews. This was happening due to the fact that when the reviews were downloaded, they were not encoded in UTF8 which was causing the errors.

II. Keywords Extraction

The second step was to extract the keywords from a review. This was done using Python's NLTK package and it's Parts of Speech Tagger.

Initially this tagger produced keyowrds with a lot of pronouns and conjunctions which I feel were just acting as noise for the real data.

Illustration

Excellent
job
quick
SEO
delivered
will
work
good
Thank
report
Great
service
recommend
fast
Highly
helpful
time
Jon
thorough
detailed
definitely
Martin
use
information
efficient
comprehensive
Fantastic
communication
well
Brilliant
happy
quality
turnaround
professional
future
lot
working
prompt
provided
clear
keyword
improve
always
Many
pleased
really
Steven
value
useful
analysis
much
promised
looking
help
recommended
need
done
advice
Nikhil
know
site
exactly
money
ranking
website
link
super
best
quickly
completed
result
forward
friendly
needed
within

This tagger was then changed to only extract keywords which were verbs or adjectives or adverbs or nouns and ignored the prepositions and conjunctions.

Illustration

helpful'
fast'
delivered'
Great
'Jon'
good'
'use'
report'
Excellent'
will'
job'
SEO'
Thank'
work'
is'
Thanks'
quick'
Martin'
detailed'
Highly'
very'
again'
really'
comprehensive'
service'
was'
definitely'
recommend'
would'
have'
be'
did'
well'
not'
provided'
time'
Fantastic'
website'
more'
done'
Many'
much'
advice'
communication'
analysis'
recommended'
information'
thorough'
efficient'
prompt'

Building a Classifier

I needed to build a classifier to actually classify the data into the 3 categories. The underlying classifier I used was a Naive Bayes Classifier from the textblob package in python. I used 80% of the extracted data as the training dataset and 20% as the test dataset. The accuracy of the classifier was about 60%.

I feel that the classifier currently is not accurate enough and I feel that this is one aspect which should be improved upon as I feel there is a lot of room for improvement for the classifier.

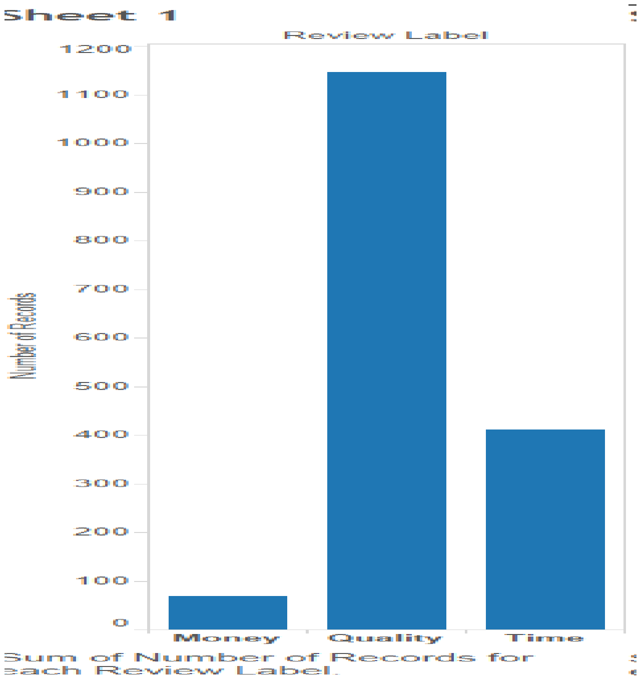
Data Analysis

All the data analysis was done using Tableau.

Theme Selection

The first thing I did while analysing the dataset was look at the overall numbers of theme selection. The result I got was the fact that majority of the people cared about the quality of the product or service they were receiving instead of the timing of the delivery or the price of the service.

One possible explanation of this result is that the source of the dataset is that of SEO Analysis where people care more about the final result of the service than they do about any other thing.

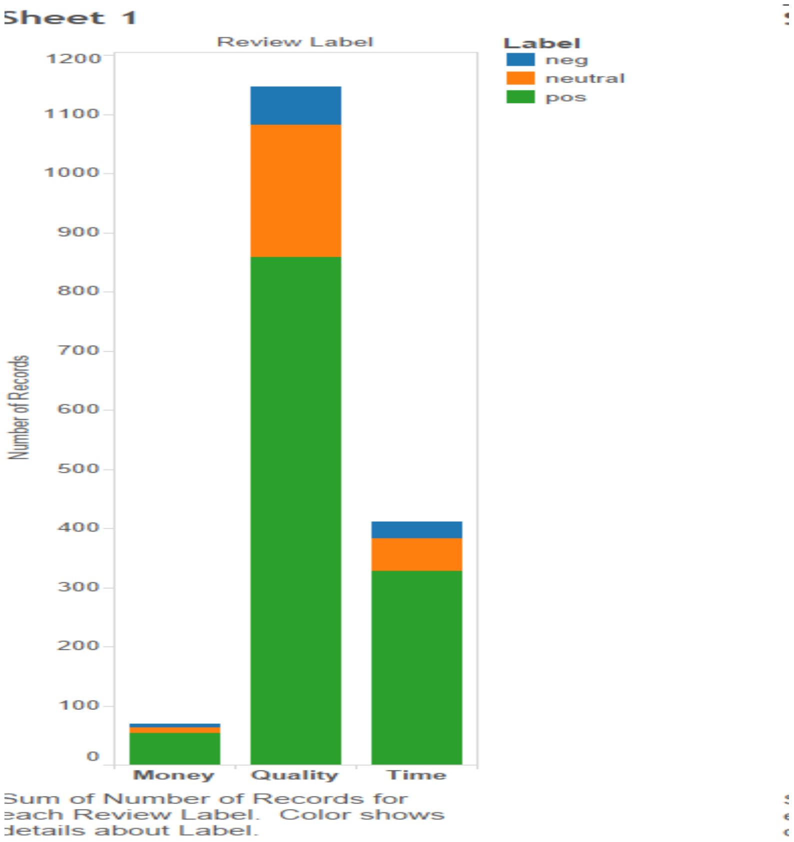


Variation with Sentiment Score

The overall number of positive reviews dominated the overall number of negative and neutral reviews. One plausible explanation that I feel is that people only tend to post reviews if they have something good to say and don't post negative reviews until something was completely wrong with the service.

The reviews tagged with the theme of Time had the highest relative proportion of negative reviews. This suggests that people tend to get more annoyed and complain more when the job isn’t done within the expected time.

While, the reviews tagged with the theme of Money had the highest relative proportion of positive reviews. This suggests that people tend to only write money-related reviews if there is something special about the cost of the service.



Sheet 1

Sum of Number of Records for each Review Label. Color shows details about Label.

Variation with Cost

The majority of the money-related reviews for lower cost jobs. The positive money related reviews were for the lower cost jobs and negative money related reviews were for the higher cost jobs.

While the majority of the time-based reviews and quality-based reviews were for jobs costing 30-60 dollars. One possible explanation is that people tend to hire more people in this cost range as it provides a balance in expected quality and cost.

Variation with Star Reviews

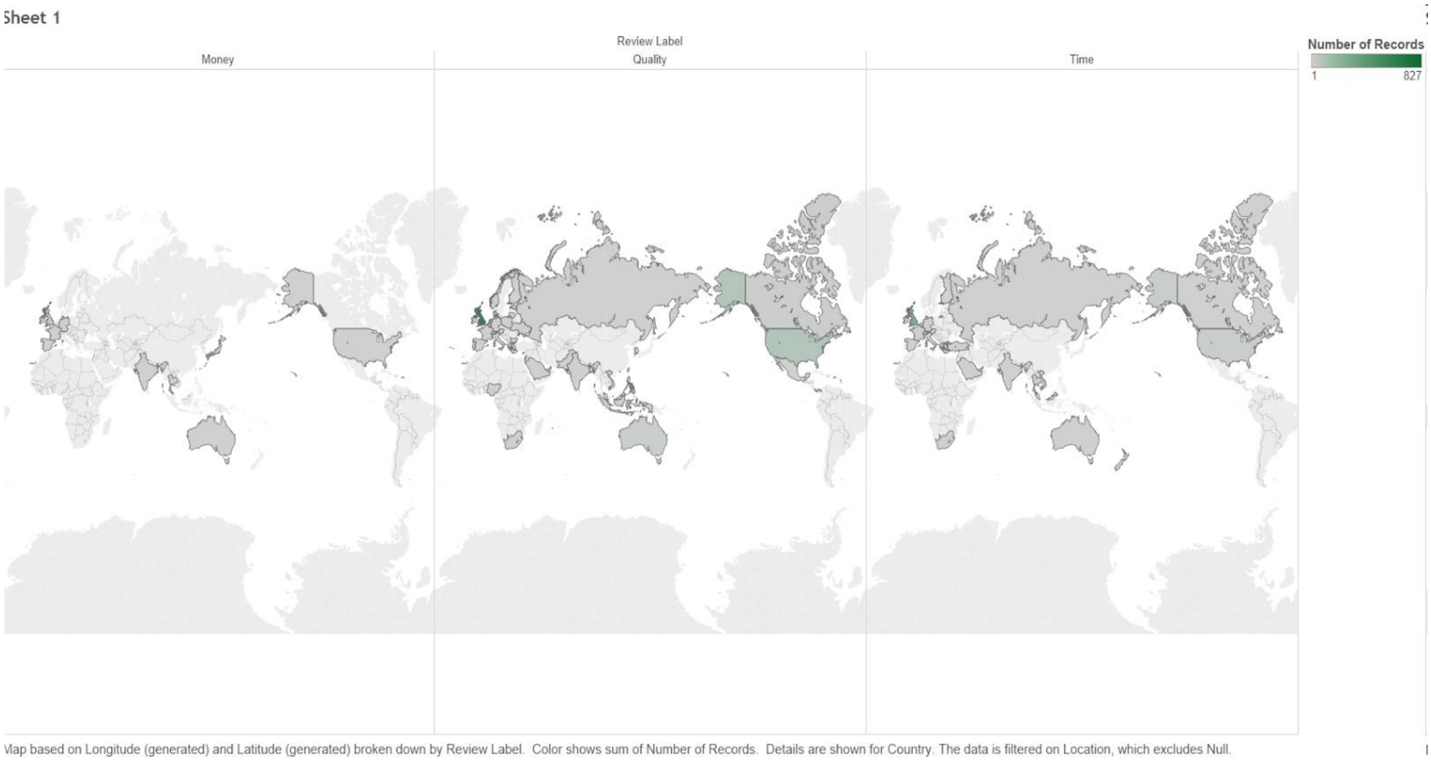
This particular analysis yielded the most surprising results. None of the negative money related reviews got a rating below 3 stars; the negative quality based reviews was evenly distributed across all 5 stars. One plausible explanation is that the people don't really properly and accurately express themselves when they are using a quantitative rating system.

Variation with Estimated Delivery Time

Most of the time-related reviews were for jobs that had 1 or 2 days as the delivery time. This suggests that people only comment about the speed of the service if the service is really fast. There was no other significant trend with reviews classified in the other 2 categories.

Variation with Location of the user

There was no significant variation with location and all locations seemed to follow the global trends. One possible issue with this is that the majority of the reviews were from the United Kingdom and thus it is hard to make any logical inferences about this.



Map based on Longitude (generated) and Latitude (generated) broken down by Review Label. Color shows sum of Number of Records. Details are shown for Country. The data is filtered on Location, which excludes Null.

Next Steps and Future Improvements

- Improve the accuracy of the classifier
- To make more logical inferences, we need more data which is not biased to a single location(United Kingdom in this case)
- Improving the review metrics in Buck Me Up to provide a better user experience.
- Buck Me Up can use my classifier in their website for automatic review classification.

Recommendations

The most important advice I will give is that you should always have a project plan and always know what your aim of the analysis is as it helps you drive towards your goal and helps you reach your goal in a more systematic way.