Text Mining and Sentiment Analysis in Hotel Booking reviews

Objectives

A professor of management science have experienced displeasing hotel staying in New York City.

- 4.2 ratings out of 5
- $130/night with free breakfast

For choosing a new hotel for her upcoming trip

- analyze customer review
- 200 reviews
- Average 170 words in a length

Problems and Solutions

- Identifying how good or bad customer rating is to understand a hotel’s review
- Making structured data from unstructured data
- Finding a typical pattern in data
- Fitting data into machine learning model for providing some recommendations
- Applying text mining to derive insights and relationships from textual data
- Using sentiment analysis to understand public opinion

Business solutions

- to understand the hotel's standing for making decision on hotel booking

Data Analysis

Document Indexing:

The simplest indexing methods depend on a list of stop word-high frequency words that did not add value (such as “is”, “the”, “there”). Document and queries are represented as vectors:

$D_j = (w_{1'j}, w_{2'j}, ,,,,,,,,,w_{t'j})$

$Q_j = (w_{1'q}, w_{2'q},;;;,,,,w_{t'q})$

Data Analysis

-Python programming

-Raw data into a clean data

-Text normalization process

- converting all text to lowercase
- removing numbers
- removing punctuations and accent marks
- removing white space
- removing stop words
- stemming words

Data Analysis

a review before text normalization (raw data)

```
"I stayed at the the BW Downtown on two separate business trips in Nov. 2011. The access to downtown is
perfect--a 10 minute walk to the office, although the shuttle is availble free of charge. The rooms are
a bit larger than average, and the beds and bedding were very comfortable. One thing that struck me is t
hat the place is very quiet--the hallways and stairs were recarpeted recently with heavy underlay and yo
u don't notice people walking up and down the halls. There is a pool and hottub that I did not take adva
ntage of. The complimentary breakfast every morning is great--eggs, bacon, sausage, yoghurt, toast, juic
e, coffee, etc. In the evening there are complimentary snacks and beverages as well. The high speed inte
rnet worked fine; much better than at larger places I've stayed at.\nThere are many restaurants and pubs
 nearby and the Galleria is about a 20 min drive. River Oaks is quite close for shopping as well and the
 theatre district is very close. The Texas Art Supply is about 5 minutes away on Montrose, and on Westhe
imer about 10 min away there are a bunch of eclectic antique stores that are worth checking out on the w
eekend.\nAccess to the airport is very good-- about a 30 minute drive on I-45 with no tolls.\nI will def
inately stay here again. The staff are very friendly and they really do make sure that your stay is comf
ortable."
```

a review after text normalization (clean data)

```
>>> data['tweettext'][2]
'stay bw downtown two separ busi trip nov access downtown perfecta minut walk offic although shuttl avai
lbl free charg room bit larger averag bed bed comfort one thing struck place quietth hallway stair recar
pet recent heavi underlay dont notic peopl walk hall pool hottub take advantag complimentari breakfast e
veri morn greategg bacon sausag yoghurt toast juic coffe etc even complimentari snack beverag well high
speed internet work fine much better larger place ive stay mani restaur pub nearbi galleria min drive ri
ver oak quit close shop well theatr district close texa art suppli minut away montros westheim min away
bunch eclect antiqu store worth check weekend access airport good minut drive toll defin stay staff frie
ndli realli make sure stay comfort'
>>>
```

Data Analysis

a review of matrix of token

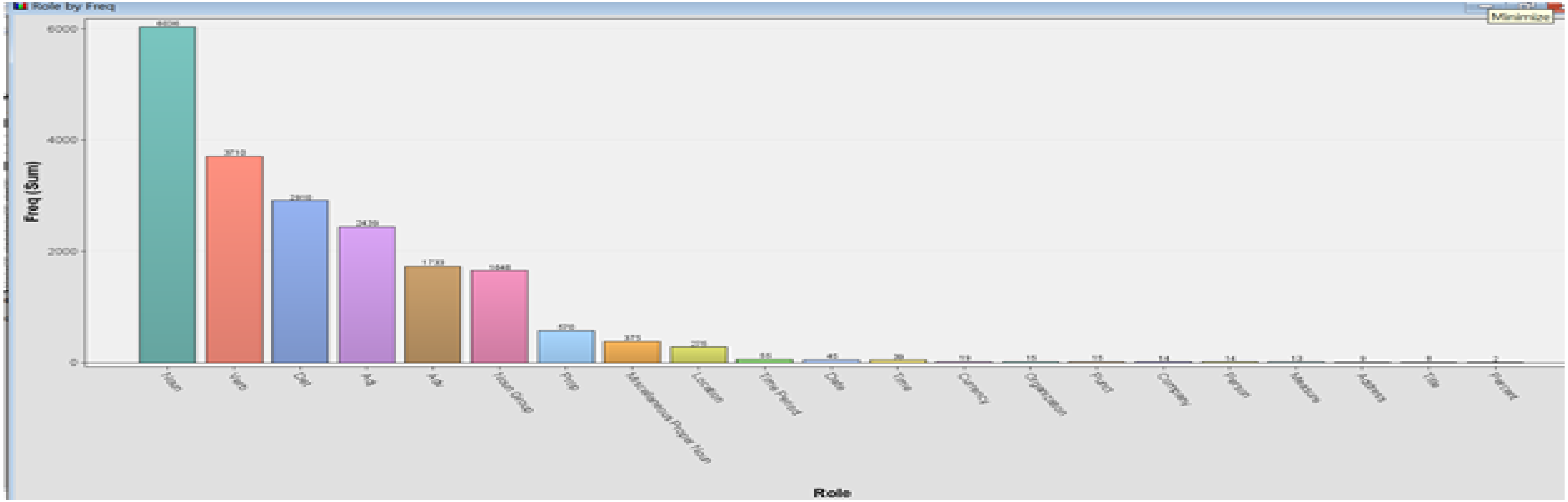
```
>>> data.rename(columns={'text': 'tweettext'}, inplace= True)
>>> LDA_value= data['tweettext'].values.astype('U')
>>> doc_term_matrix = vectorizer.fit_transform(LDA_value)
>>> doc_term_matrix.shape
(200, 642)
>>>
```

Data Analysis

-Alternative method

-SAS Enterprise Miner

- Deep dive into understanding the data set, the data has different roles and the top roles consists of ”Noun, Verb, Det, Adj” and so on.



Data Analysis

- Applying spell check function

EMWS1.TextFilter_spellDS

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary	Key	Parent ID
1	4.0	air	1.0	air	Verb	Noun	0.0	Y	2299.0	33.0
2	4.0	air	2.0	air	Adj	Noun	0.0	Y	4452.0	33.0
3	4.0	äôs	1.0	äôs	Prop	PROP_MISC	0.0	N	3207.0	59.0
4	3.0	mention	1.0	mention	Noun	Verb	0.0	Y	2794.0	64.0
5	6.0	plaza	2.0	plaza	Prop	Noun	0.0	Y	2449.0	78.0
6	3.0	sherry nethrland	1.0	sherry netherland	PROP_MISC	PROP_MISC	6.0	N	2472.0	104.0
7	17.0	new york	1.0	new york	NOUN_GROUP	LOCATION	0.0	N	1451.0	109.0
8	4.0	corridor	1.0	corridor	Noun	Prop	0.0	Y	2776.0	166.0
9	8.0	hard	2.0	hard	Adv	Adj	0.0	Y	289.0	229.0
10	5.0	pleasure	1.0	pleasure	Verb	Noun	0.0	Y	2502.0	239.0
11	9.0	professional	1.0	professional	Noun	Adj	0.0	Y	3955.0	258.0
12	17.0	trip	3.0	trip	Prop	Noun	0.0	Y	2665.0	275.0
13	4.0	miss	1.0	miss	TITLE	Verb	0.0	Y	2402.0	297.0
14	4.0	miss	1.0	miss	Noun	Verb	0.0	Y	3039.0	297.0
15	26.0	restaurante	1.0	restaurant	Noun	Noun	3.0	N	3865.0	298.0
16	21.0	definately	2.0	definitely	Noun	Adv	10.0	N	4205.0	330.0
17	13.0	last	2.0	last	Verb	Adj	0.0	Y	913.0	376.0
18	10.0	return	3.0	return	Noun	Verb	0.0	N	3137.0	378.0
19	3.0	citycenter	1.0	city center	PROP_MISC	ORGANIZATION	10.0	N	3492.0	402.0
20	32.0	suite	2.0	suite	Prop	Noun	0.0	Y	2825.0	423.0
21	7.0	wifi	1.0	wifi	Prop	Noun	0.0	N	2829.0	428.0
22	7.0	wi-fi	3.0	wifi	Noun	Noun	12.0	N	4271.0	428.0
23	7.0	wifi	2.0	wifi	PROP_MISC	Noun	0.0	N	3466.0	428.0
24	6.0	bedroom	5.0	bedroom	Adj	Noun	0.0	Y	3300.0	437.0
25	6.0	bed room	1.0	bedroom	NOUN_GROUP	Noun	14.0	N	1952.0	437.0
26	3.0	west	1.0	west	Prop	Noun	0.0	Y	2887.0	469.0

Data Analysis

- Included a dictionary file

BC2.filter

	term	termrole	parent	parentrole
1	central park south	LOCATION	central park	LOCATION
2	central-park	Prop	central park	LOCATION
3	central-parks	PROP_MISC	central park	LOCATION
4	new york city	LOCATION	new york	LOCATION
5	nyc	LOCATION	new york	LOCATION
6	san	Prop	carlos	Prop
7	san carlos	LOCATION	carlos	Prop
8	sherry	Prop	sherry	Noun
9	sherry netherland hotel	PROP_MISC	sherry	Noun
10	sherry-netherland	PROP_MISC	sherry	Noun
11	york	Prop	new york	LOCATION

Data Analysis

Words that are being filtered out for the “ignore parts of speech” as per below shown like the “Abbr, Aux, Conj, Det, Intej, Num, Part, Pref, Prep and Prop”.

Terms

TERM	FREQ	# DOCS	KEEP ▲	WEIGHT	ROLE	ATTRIBUTE
the	1553	178		0.0	Det	Alpha
be	986	177		0.0	Verb	Alpha
a	657	166		0.0	Det	Alpha
have	231	109		0.0	Verb	Alpha
very	177	102		0.0	Adv	Alpha
not	208	97		0.0	Adv	Alpha
my	140	91		0.0	Det	Alpha
this	138	87		0.0	Det	Alpha
our	130	68		0.0	Det	Alpha
do	121	64		0.0	Verb	Alpha
get	84	53		0.0	Verb	Alpha
make	68	52		0.0	Verb	Alpha
here	62	45		0.0	Adv	Alpha
go	59	45		0.0	Verb	Alpha
an	55	45		0.0	Det	Alpha
also	51	38		0.0	Adv	Alpha
again	42	37		0.0	Adv	Alpha
no	45	37		0.0	Adv	Alpha
just	45	36		0.0	Adv	Alpha

Data Analysis

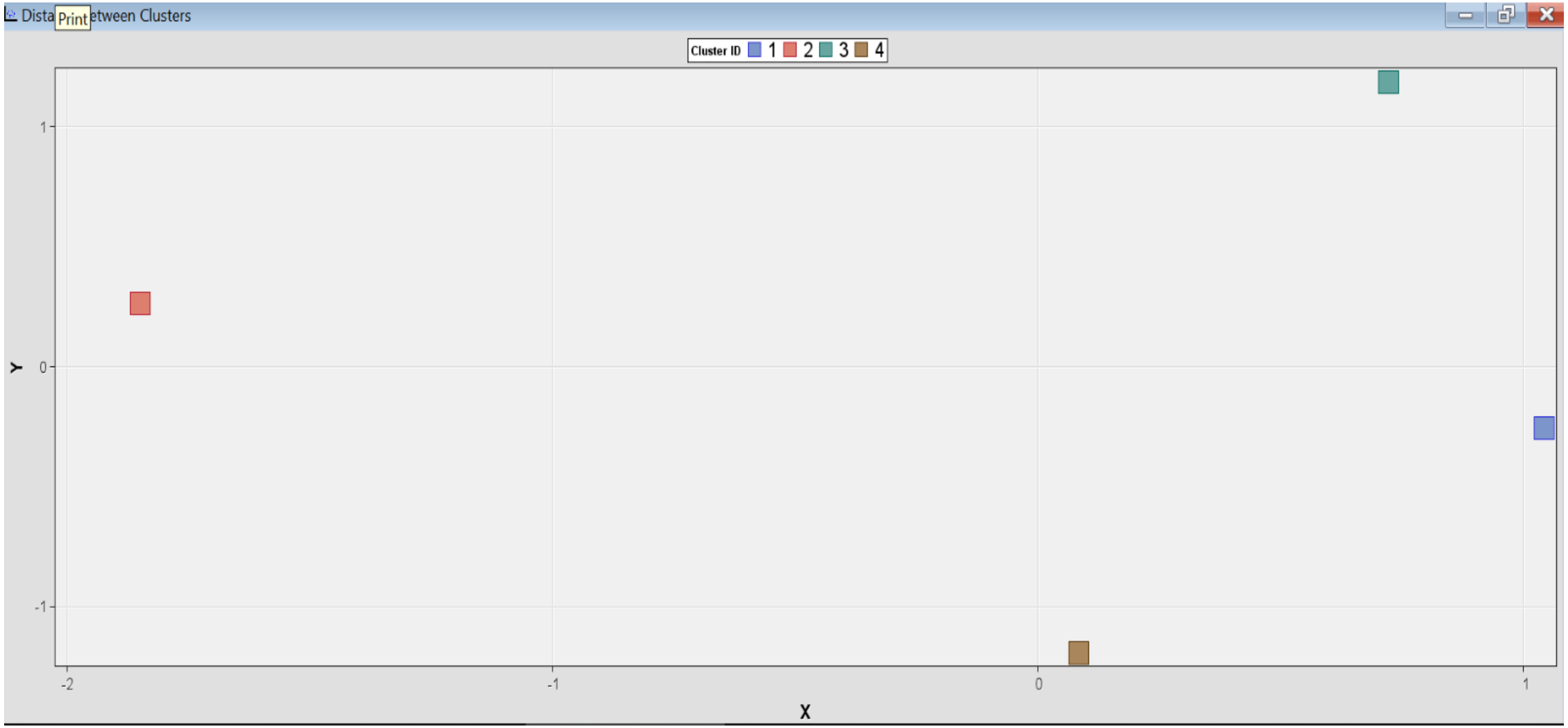
Words that are being filtered out for the “ignore parts of speech” as per below shown like the “Abbr, Aux, Conj, Det, Intej, Num, Part, Pref, Prep and Prop”.

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage ▼	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coordinate 7	Coordinate 8	Coordinate 9	Coordinate 10	Coordinate 11	Coordinate 12
4	+'new york' +sherry +suite +next service free +location +always ...	23	51%	0.4243...	-0.25299	0.0749...	0.0162...	0.0689...	-0.30954	0.1058...	-0.01071	0.0532...	-0.00179	0.0176...	-0.0639
1	+recommend +decide +good nice +houston +center desk +look ...	9	20%	0.5799...	-0.00953	0.01539	-0.05088	0.0232...	-0.06628	-0.08114	0.1426...	-0.18462	0.2017...	0.0321...	0.0017.
3	+westin +memorial +shower +business +houston +area +decide +great ...	9	20%	0.4493...	-0.17845	-0.1968	-0.29953	0.06346	0.07331	-0.02883	-0.0886	-0.06851	0.08212	-0.04527	0.1160.
2	+hilton +cater +expectation +line +food +cover +detail +manager ...	4	9%	0.3604...	0.1134...	0.4725...	-0.2181	0.0473...	0.1764...	0.03018	-0.07645	0.2224...	0.0056...	0.1904...	0.0710.

Data Analysis

EM Cluster distance



Data Analysis

Hierarchical clustering

Hierarchy Data

Hierarchy Level	Cluster ID	Parent	Descriptive Terms	Frequency	Graph Description
1	1	.			451
2	2	1	+trip restaurants +favorite +location +'new york' excellent free +bed +business service york +clean +always +comfortable +little		332: +trip
2	10	1	nice included decided +bar +back +definitely drinks stayed +place helpful +bathroom +time +nice +good +houston		1210: nice
3	3	2	+hotel +staff +room +stay york +'new york' rooms service +location +clean +good +breakfast +great free +nice		333: +hotel
4	4	3	+city +trip restaurants +bed +business +lobby hotels +'new york' +clean +good +great free +stay york +hotel		274: +city
4	7	3	wedding +big +daughter +feel +hilton +reception colors event appointed +food +manager +warm details +floor few		67: wedding
5	5	4	perfect +back +always +next +park +place +sherry dining parking sherry +room +bed +stay york service		115: perfect
5	6	4	+breakfast +comfortable +houston small +airport +carlos +restaurant +street +great restaurants +little +lobby hotels +good +suite		166: +breakfast

Data Analysis

Hierarchical clustering – 4 clusters

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage ▼	Coordinate 1	Coordinate 2	Co ate
6	+breakfast +comfortable +houston small +airport +carlos +restaurant +street +great restaurants +little +lobby hotels +good +suite	16	36%	0.4381...	-0.1778	0.0
10	nice included decided +bar +back +definitely drinks stayed +place helpful +bathroom +time +nice +good +houston	12	27%	0.4394...	-0.12877	-0.(
5	perfect +back +always +next +park +place +sherry dining parking sherry +room +bed +stay york service	11	24%	0.5046...	-0.12599	0.0
7	wedding +big +daughter +feel +hilton +reception colors event appointed +food +manager +warm details +floor few	6	13%	0.4336...	-0.16787	0.1

Data Analysis

- Topics by the number of terms
- The top cluster by topics is “meeting/kitchen/hilton/help”.

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms ▼	# Docs
Multiple	22	0.096	0.021	+meeting,+hgi,+kitchen,+hilton,+help	549	5
Multiple	21	0.109	0.021	+year,a year,+accommodating,+business,+several	535	10
Multiple	13	0.120	0.021	san,san carlos,carlos,+bedroom,+new york	459	14
Multiple	17	0.125	0.021	+houston,drury,+la quinta,+clean,+east	459	11
Multiple	7	0.106	0.021	convenient,+traveler,+hilton,+property,+good	454	6
Multiple	24	0.111	0.021	+area,+access,easy,+bed,+comfortable	446	13
Multiple	20	0.120	0.020	+new york,+york,+sherry,+favorite,star	421	11
Multiple	15	0.093	0.021	+downtown,+pillow,everyday,kindly,+shuttle	419	4
Multiple	4	0.112	0.020	+sherry-netherland,+operator,+elevator operator,+elevator,+sherry	413	14
Multiple	14	0.114	0.021	+great,+mini,+block,free,flat	413	12
Multiple	16	0.109	0.021	+king,garage,+book,+park,level	395	8
Multiple	1	0.110	0.020	+westin,+memorial,card,city,+mall	394	6
Multiple	3	0.104	0.021	san,san carlos,carlos,assorted,cheese	372	11
Multiple	10	0.116	0.021	+want,+table,+wedding,+wed,+food	365	9
Multiple	11	0.125	0.020	+times square,+square,distance,walking distance,+walking	363	13
Multiple	23	0.095	0.021	+corner,+want,+side,+early,famous	363	3
Multiple	2	0.122	0.018	und,+sehr,º,+das,+ein	357	4
Multiple	25	0.095	0.020	äi,+point,+mistake,housekeeping,+reception	349	6
Multiple	18	0.103	0.020	+close,+question,event,+dress,personnel	341	3
Multiple	6	0.103	0.020	+hilton,+amaze,+family,saturday,+basket	339	7
Multiple	9	0.112	0.020	music,fire,brunch,+walk,+decide	328	6
Multiple	19	0.108	0.020	+review,+average,management,honest,+good review	274	5
Multiple	5	0.121	0.020	sonny,quilt,+bottle,+shuttle,+egg	250	8
Multiple	12	0.099	0.019	al,il,di,servizio,molto	170	2
Multiple	8	0.099	0.018	de,en,el,+la,como	162	2

Result and Recommendation

Top ten words in terms of volume by python

```
>>> sort_text = tokens_data.sum()
>>> sort_text.sort_values(ascending = False).head(10)
hotel        358
room         263
stay         211
staff        150
great        121
nice          96
locat         93
clean         92
breakfast     92
good          84
dtype: int64
>>>
```

Result and Recommendation

Top words in terms of volume by SAS Enterprise Miner

Terms

Term	Role	Attribute	WEIGHT	Freq	# Docs	Keep	Rank for Variable NUMDOCS	+carlos,adequate,+breakfast,subway,+new york	und,der,°,+das,im	music,fire,brunch,+area,+several	card,+westin,+memorial,st.,regis	+wedding,+food,+hilton,+want,+reception	+sherry,+new york,+appoint,entrance,+harry cipriani	+king,garage,+bed,+book,level	+houston,+clock,+expect,+good,+close	size,mom,+real,+bedroom,+sheet	_termid
+ hotel	...Noun	Alpha	0.121333	86	36	Y	1	0.044	0.014	0.024	0.015	0.018	0.038	0.029	0.068	0.031	3531
+ room	...Noun	Alpha	0.133325	70	34	Y	2	0.042	-0.001	0.017	0.03	0.008	0.049	0.052	0.046	0.026	957
+ stay	...Verb	Alpha	0.15969	37	27	Y	3	0.049	-0.001	0.016	0.026	0.02	0.031	0.034	0.013	0.014	56
+ staff	...Noun	Alpha	0.162139	32	26	Y	4	0.031	-0.001	0.008	0.017	0.021	0.034	0.026	0.025	0.026	2872
+ new york	...Location	Entity	0.217408	64	23	Y	5	0.113	0.029	-0.013	-0.015	-0.004	0.143	0.011	0.035	0.034	109
service	...Noun	Alpha	0.234896	21	19	Y	6	0.047	-0.001	0.009	0.007	0.027	0.071	0.001	0.026	-0.014	3386
+ good	...Adj	Alpha	0.259611	28	18	Y	7	0.007	-0.001	0.014	0.041	-0.003	0.04	0.013	0.103	0.022	357
+ nice	...Adj	Alpha	0.289461	26	16	Y	8	0.083	-0.002	0.023	-0.002	0.033	0.004	0.067	0.026	0.023	1244
+ great	...Adj	Alpha	0.321604	21	15	Y	9	0.042	-0.001	-0.009	0.055	-0.006	0.004	0.01	0.064	0.06	845
+ location	...Noun	Alpha	0.319989	17	14	Y	10	0.013	0.001	-0.007	0.043	-0.008	0.063	-0.012	0.06	0.039	2328
+ breakfast	...Noun	Alpha	0.367908	19	13	Y	11	0.12	-0.001	-0.008	0.06	-0.001	0.004	-0.005	0.033	-0.011	1898
+ clean	...Adj	Alpha	0.346323	15	13	Y	11	0.051	-0.002	0.009	-0.011	0.06	0.008	0.012	0.054	0.015	2590
+ day	...Noun	Alpha	0.332739	14	13	Y	11	0.03	-0.003	-0.007	0.018	0.026	0.045	0.011	0.072	-0.022	2102
+ restaurant	...Noun	Alpha	0.339931	16	13	Y	11	0.034	0	0.023	0.054	0.004	0.038	0	0.044	-0.003	298
+ view	...Noun	Alpha	0.332739	14	13	Y	11	0.035	0.021	-0.001	0.073	0.003	0.026	-0.011	0.013	0.002	846
+ place	...Noun	Alpha	0.358752	14	12	Y	16	0.035	-0.002	0.032	0.03	0.009	0.057	0.019	-0.006	-0.001	1842
+ suite	...Noun	Alpha	0.370601	15	12	Y	16	0.093	0.019	-0.01	-0.01	-0.006	0.047	0.001	-0.006	0.064	423
+ time	...Noun	Alpha	0.401496	17	12	Y	16	0.006	-0.003	0.077	0.011	0.027	0.07	0.009	0.044	0.001	2466
+ bathroom	...Noun	Alpha	0.394583	14	11	Y	19	0.062	-0.001	-0.004	-0.009	0.019	0.004	0.059	-0.015	0.091	2064
+ city	...Noun	Alpha	0.394583	14	11	Y	19	0.009	0.001	-0.006	0.046	-0.015	0.048	-0.008	0.061	0.043	764
+ floor	...Noun	Alpha	0.392795	13	11	Y	19	0	-0.002	0.02	0.08	0.027	0.04	-0.014	-0.011	0.055	651
+ front	...Adj	Alpha	0.39488	15	11	Y	19	0.025	-0.001	-0.007	-0.014	0.037	0.023	0.064	0.025	0.093	803
+ stay	...Noun	Alpha	0.384764	14	11	Y	19	0.048	-0.002	0.033	0	0	0.027	0.06	0.043	-0.003	2791
desk	...Noun	Alpha	0.382222	13	11	Y	19	0.008	-0.001	-0.004	-0.014	0	0.023	0.055	0.066	0.079	2346
helpful	...Adj	Alpha	0.37757	12	11	Y	19	0.068	-0.002	0.003	0.037	0.002	0.018	0.011	0.019	0.023	1683
+ comfortable	...Adj	Alpha	0.395117	10	10	Y	26	0.035	-0.001	0.004	-0.011	0.031	0.011	0.007	0.058	0.033	1887
+ houston	...Location	Entity	0.425396	16	10	Y	26	-0.01	0	0.015	0.102	0.023	-0.014	-0.001	0.111	0.019	2253
+ little	...Adj	Alpha	0.420809	13	10	Y	26	0.02	0.001	0.064	0.011	0.022	-0.02	-0.011	0.074	0.044	2295
+ lobby	...Noun	Alpha	0.395117	10	10	Y	26	0.006	-0.002	0.017	0.059	-0.009	0.014	0.024	0.048	0.009	91
+ sherry	...Noun	Alpha	0.413763	17	10	Y	26	-0.014	0.038	-0.007	-0.005	0	0.196	-0.008	-0.002	0.014	3468
excellent	...Adj	Alpha	0.410235	13	10	Y	26	0.035	-0.001	-0.005	0.021	0.014	0.059	0.009	0.033	-0.009	2417
+ always	...Adv	Alpha	0.431535	10	9	Y	32	0.012	-0.001	-0.008	-0.009	0.047	0.068	0.035	-0.008	0.042	2125
+ love	...Verb	Alpha	0.431535	10	9	Y	32	0.013	-0.002	0.041	0.042	0.032	0.026	0.02	-0.03	0.04	641
+ next	...Adj	Alpha	0.431535	10	9	Y	32	0.043	-0.002	0.025	-0.007	0.049	0.071	0.007	0.004	-0.031	4375
+ area	...Noun	Alpha	0.551434	15	8	Y	35	0.011	0.001	0.121	0.01	-0.024	-0.023	-0.006	0.09	0.068	3110
+ bed	...Noun	Alpha	0.534599	15	8	Y	35	0.027	0	-0.003	0.019	0.005	0.002	0.167	0.017	0.046	1635
+ find	...Verb	Alpha	0.453736	8	8	Y	35	0.026	0.001	0.024	0.002	0.017	-0.005	-0.012	0.06	0.031	3668

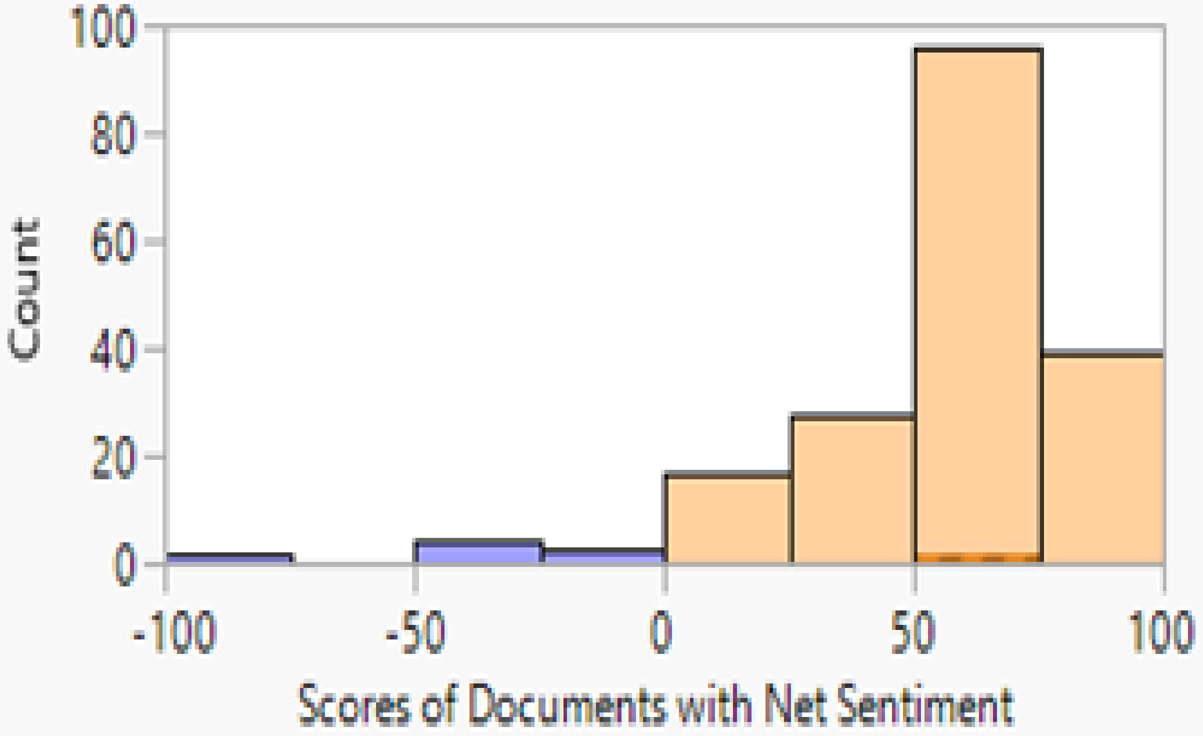
Result and Recommendation

Top words by JMP pro 16

Result and Recommendation

Sentiment Analysis by JMP pro 16

	N	Mean Score
All Scored Documents	185	52.7
Net Positive Documents	179	59.3
Net Negative Documents	6	-39.2
No Sentiment Documents	15	0.0



Result and Recommendation

Positive and negative word score

Document	Positive Sum	Positive Score Mean	Negative Sum	Negative Score Mean	Overall Score
1	80	80	0	0	80
2	120	30	-20	-20	20
3	240	60	0	0	60
4	210	70	0	0	70
5	170	85	0	0	85
6	90	90	0	0	90
7	318	40	-125	-63	19
8	80	80	0	0	80
9	170	85	0	0	85
10	155	52	0	0	52

Sentiment	Score	Count
great	80	112
good	60	61
nice	25	59
friendly	40	42
helpful	35	37
excellent	90	33
best	90	30
comfortable	40	28
wonderful	90	22
beautiful	80	21
perfect	90	21

Result and Recommendations

-In Python, JMP pro 16 and SAS Enterprise Miner, the most frequent and important terms are found to be hotel, room, staff, stay, breakfast, good, locat, clean

-Even Text cluster, and Hierarchical cluster’s terms have positive association with the hotel and frequent positive words are “favorite, excellent, clean, good, nice, comfortable, and perfect etc.

Result and Recommendations

Short-term recommendation

- to analyze top positive and negative words to understand hotel’s standing

Long-term recommendations

- To analyze the customer’s review from different sources (such as: from different website)
- to analyze large data by collecting from various sources

Conclusion

In this case, we have used text mining and sentiment analysis to analyze text for sentiment of 200 customers' review. Based on the Python program, SAS enterprise miner and JMP pro 16 results, we believed that the consumer reviews ratings would be aligned with the textual data.