



CASE 2

GROUP 6

SUN, FANGZHENG

LIU, HUANHAN

XU, QIAN

BAGUL, AASHISH

The background is a deep blue gradient with a subtle pattern of white stars and dots. Overlaid on this are several faint, white circular and semi-circular lines of varying thicknesses. Some of these lines have small arrowheads pointing in different directions, suggesting a sense of motion or rotation. A prominent circular scale is visible on the left side, with numerical markings from 140 to 260 in increments of 10. The overall aesthetic is futuristic and technological.

WHAT IS OPINION MINING??

THE PROBLEM

- Why should Businesses care about reviews???
- Consumers do not provide explicit ratings too
- So there is a need for an automated system to classify reviews as positive and negative and correct them as well.

EXAMPLES OF DUMB REVIEWS



Tony B.
Glendale, CA
👤 61 friends
★ 10 reviews



12/29/2013

I give this restaurant a 4* just because I don't believe a restaurant should be 5*'s. This is a great place to grab a bite to eat. The breakfast is top quality for a country style restaurant and they have a new coffee vendor, that is pretty darn good. The staff is enjoyable but I come for the great breakfast options. I'm never let down.

Was this review ...?



Useful 1



Funny



Cool



Jim S.



Jim S.
Boston, MA
👤 6 friends
★ 3 reviews



Share review



Compliment



Send message



Follow Jim S.



4/10/2012

1st problem: Ugly girls. You would think an upscale bar like this would have hot chicks, well you're wrong. Overweight girls, no makeup on, butter faces, you name it. On the plus side, I've had a couple one night stands from this place. On the downside, they were both really ugly and I could barely even perform. 2nd problem: Bathroom policy. The freaking bathroom is outside the bar and then it's hard to get back into the bar afterwards. They gave me a bathroom card, but I lost it while making out with one of the ugly chicks described above. The doorman remembered me, but he wouldn't let me in without the bathroom card. Shouldn't he have been happy that I was helping to remove some of the ugly chicks from the bar? DOUCHE BAG. Even if you do get

DATA COLLECTION

- We converted the yelp data from JSON format to a CSV file.
- We used The yelp Dataset and Filtered the yelp's 1,000,000 reviews data to about 10000 “most useful” one.
- At last our visualization is done by Word Cloud

Data Collection

Data Storage

Data Storage

Data Analysis

Data Visualization

DATA EXPLORATION

- Text Blob Technique:-

Text Blob is a Python library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

DATE EXPLORATION

Logistic Regression Models Technique

- Bag of Words Model

In this model, a text (such as a sentence or a document) is represented as the **bag** (multiset) of its **words**, disregarding grammar and even **word** order but keeping multiplicity

- IDF/TF Model

It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

THE EXAMPLE RESULT OF TEXT BLOB

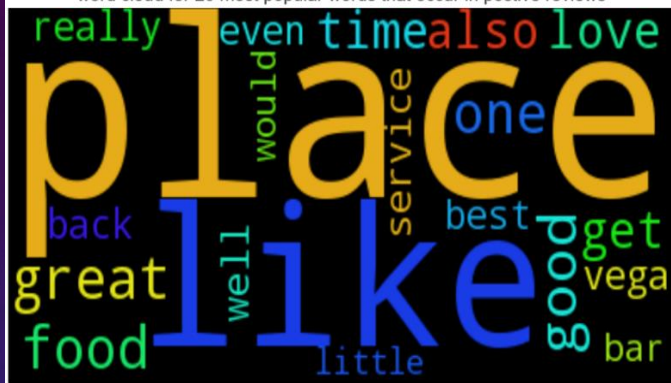
	cool	funny	stars	text	useful	textblob_sentiment_score	textblob_predicted_stars
0	7	6	5	So....You say you want authentico?? My family ...	11	0.091254	1
1	7	11	5	You're getting old. Your college days are over...	10	0.087778	1
2	29	58	5	5 totally skanky stars! and i say that in the ...	36	0.077722	1
3	8	6	5	When you get tired of DJ's and electronic musi...	17	0.120377	5
4	9	7	5	This would be me on a daily basis.\r\n\r\nhttp...	10	0.403750	5
5	6	0	5	In terms of outdoor public swimming pools, thi...	11	0.087387	1
6	6	10	1	Ugh. My ladies and I rolled in sometime past m...	11	0.026954	1
7	16	19	1	Dear person who decided that we would have our...	15	0.079305	1
8	5	5	1	So we were looking for something different in ...	11	-0.070517	1
9	9	8	1	The last couple of weeks had been painful. A p...	11	-0.045178	1
10	4	6	5	\$12 haircut. SERIOUS.\r\n\r\nWho the hell care...	10	0.255000	5
11	27	19	5	OK, people. Time to hunker down and get real s...	50	0.177121	5
12	24	21	5	Das Autokino in Kornwestheim erinnert mich an ...	27	0.600000	5
13	17	7	5	A wonderful cause, and I've always liked the i...	15	0.319527	5
14	6	7	5	Verona is a welcome addition to the valley. I...	10	0.215274	5
15	17	16	5	Best... Steak... EVER!!!\r\n\r\nThat's it. Thos...	18	0.296087	5
16	14	15	5	Once in a while, miracles happen to me. It's o...	15	0.072856	1
17	12	28	5	If this place were crack cocaine*, I'd snort i...	15	0.400000	5
18	8	7	5	Ever since I was a food judge at the Culinary ...	11	0.114461	5

THE RESULT

```
# word_count sum sorted by value  
counter.most_common(20)
```

```
[('like', 2563),  
 ('one', 2314),  
 ('place', 2125),  
 ('get', 2011),  
 ('good', 1937),  
 ('time', 1842),  
 ('food', 1675),  
 ('great', 1649),  
 ('also', 1540),  
 ('really', 1377),  
 ('back', 1356),  
 ('would', 1330),  
 ('well', 1326),  
 ('even', 1292),  
 ('love', 1097),  
 ('got', 1017),  
 ('service', 1005),  
 ('first', 1003),  
 ('vegas', 980),  
 ('best', 946)]
```

word cloud for 20 most popular words that occur in postive reviews



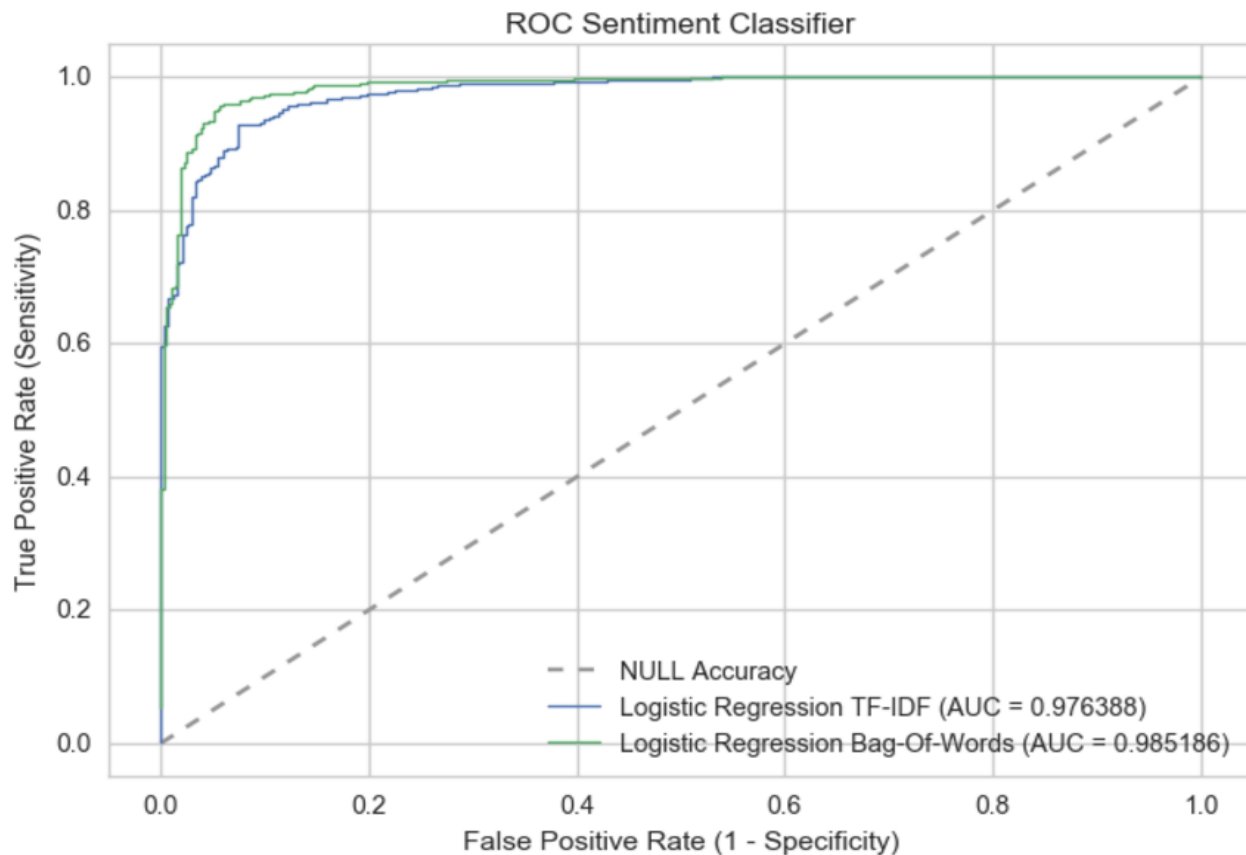
word cloud for 20 most popular words that occur in negative reviews



```
# word_count sum sorted by value  
counter.most_common(20)
```

```
[('would', 1669),  
 ('get', 1589),  
 ('one', 1410),  
 ('like', 1363),  
 ('time', 1344),  
 ('back', 1278),  
 ('place', 1168),  
 ('said', 1108),  
 ('even', 1070),  
 ('told', 1053),  
 ('never', 879),  
 ('service', 844),  
 ('people', 811),  
 ('could', 800),  
 ('know', 738),  
 ('got', 731),  
 ('asked', 721),  
 ('food', 715),  
 ('good', 691),  
 ('went', 641)]
```

COMPARISON OF LOGISTIC MODELS



CONCLUSION

- We did Sentimental Analysis on YELP Data we used Text Blob & Logistic Regression models i.e. Bag of Words, IDF/TF techniques for Prediction.
- We Successfully predicted and classified 10000 reviews and also predicted the number of stars for those reviews.
- Finally we compared the accuracy of the models and found Bag of model does best predictions.

LIMITATION

- The Data is Unstructured and even though our model can classify it but it cannot correctly classify big long story reviews.
- If the User reviews contains lots of words out of the Bag of words model.
- Computational Limit.

Thank You!!!!