TEXT MINING HOTEL ONLINE REVIEWS

Group Members:

Frances Wang (Data collecting, Cleaning & Preprocessing, Vader for Review Distinction, Classification, Final Report)

Joshua Hong (Clustering, Topic Modeling & Analysis, Emolex for Sentiment Analysis, Membership Level Analysis)

Cornell University
INFO 6350

OUTLINE

1. **Problems and hypothesis.**
    a. What problem are you working on? Why is it interesting and important?
    b. What have other people said about it? What do you expect to find?
2. **Corpus, data, and methods.**
    a. What data have you used? Where did it come from? How did you collect it?
    b. What major methods will you use to analyze it? Why are those methods the appropriate ones?
3. **Results.**
    a. What did you find? How did you find it? How should we read your figures?
4. **Discussion and conclusions.**
    a. What does it all mean? Do your results support your hypothesis? Why or why not?
    b. What are the limitations of your study and how might those limitations be addressed in future work?
5. **Code and Data Attached.**

PROBLEMS & HYPOTHESIS

1. **Problems:**
    a. Do bad reviews tend to be shorter in number of words?
    b. Do good and bad reviews show clear sentiment distribution?
    c. Can we use review content to effectively predict bad reviews?
    d. Are certain topics more related to positive/negative reviews?
    e. Do different rating's reviews associate more with certain emotions (based on emolex) ?
    f. Based on the membership system of Marriott, do higher level members have more positive ratings and emotions (based on emolex)?
2. **Hypotheses:**
    a. There is no significant difference in the word length of bad and good reviews.
    b. Good and bad reviews show clear sentiment distribution.
    c. It is reliable to use review content (sentiment, length, word use) to predict bad reviews.
    d. Certain topics are more related to positive/negative reviews.
    e. Different rating's reviews associate more with certain emotions (based on emolex).
    f. Based on the membership system of Marriott, higher level members have more positive ratings and emotions (based on emolex).
3. **Why interesting and important:**
    - Understand trends: to understand what consumers are talking about, such as things they like or things they do not like.
    - Understand customer rating behavior better by aligning it with their text review.
    - Provide insights to improve service/products from users feedbacks.

CORPUS, DATA, METHODS

1. Data:

We collected online review data from Marriott.com. It includes the username, their membership level in Marriott, the date of the review, review content, and the overall rating score, totaling to **10 hotels** with **11417 reviews**. It was scraped from Marriott.com using Python Selenium and Webscrapper.

2. Methods:
    a. Initial data cleaning (remove nulls, unify the format, exclude unrelated variables, create new variable is_bad_review – we define as rating < 3)
    b. NLTK toolkit for review cleaning and preprocessing (stop words, lemmatize, etc.)
    c. SciPy library for statistical tests (t-test, etc.)
    d. Sentiment Analysis using Vader and Emolex
    e. Vectorization using CountVectorizer and TfIdfVectorizer
    f. Scikit-learn library for Random Forest Classification to predict bad reviews
    g. KElbowVisualizer to find optimal number of clusters and KMeans Clustering
    h. LatentDirichletAllocation for Topic Modeling
    i. Matplotlib for visualization
3. Initial Data Display

	member_level	review	is_bad_review	rating
0	Member	This hotel is very dirty.	1	1
1	Member	The driveway/entrance is uninspiring. Checkin ...	0	3
2	Platinum	Not impressed Continue to undergo construction...	0	3
3	Member	Amazing every aspect ... the food is an issue ...	0	5
4	Member	Very nice staff, beautiful hotel and view. Ver...	0	5
5	Member	Done. My stay at Marriott was good. I apprecia...	0	3
6	NaN	Requested shower caps but as room was not assi...	0	4
7	Gold	We stayed at the Atlanta Marriott Marquis duri...	0	5
8	Gold	Nice downtown location. Lovely hotel. Good ser...	0	3
9	Member	Enjoyed the staying at the Marriott. Most of t...	0	4

RESULTS

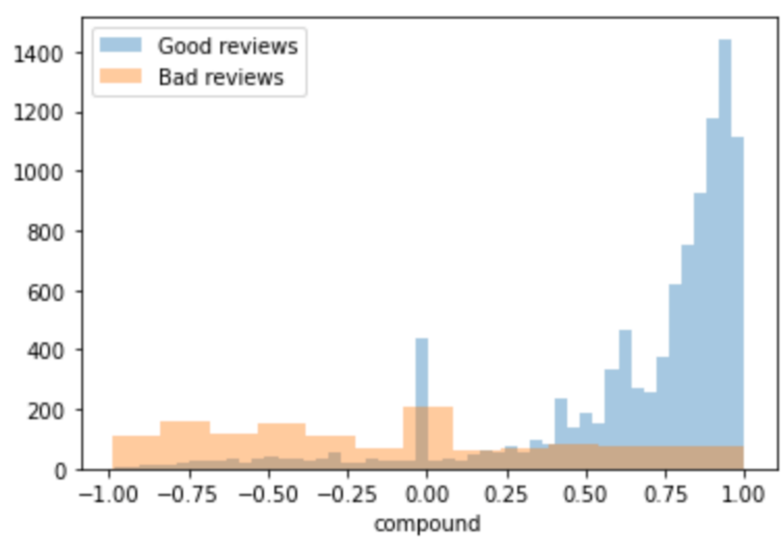
Hypothesis a : There is no significant difference in the word length of bad and good reviews.

Based on our t-test, P-value << 0. 01, we should reject the hypothesis that there is no significant difference between word count of bad and good reviews. Surprisingly, bad reviews have longer reviews than good reviews in general. It is possible to believe that people who left bad reviews tend to give more details in their context to make their reviews look reliable.

```
t-statistic: 20.112297208627133
p-value:     1.9433400873468887e-88
Average length of bad reviews:
56.310344827586206
Average length of good reviews:
32.94698100069631
```

Hypothesis b: Good and bad reviews show clear sentiment distribution.

We used Vader to build the compound sentiment score and got a clear distribution. It shows that for good reviews, the sentiment score is much higher than bad reviews. However, for bad reviews, it is towards the negative compound sentiment score with an even distribution in the positive part as well. This graph shows that we can use Vader sentiment score as a reliable predictor to good/bad reviews to some degree.



Hypothesis c: It is reliable to use review content (sentiment, length, word use) to predict bad reviews.

Using random forest classifier, we obtained a test accuracy score of 89.93%, and the ranking of feature importance. So, we believe this method of prediction is reliable. It is within our assumptions that the sentiment scores are the most reliable predictors for good/bad reviews, followed by the number of characters and words. Certain words show stronger predicting abilities too and we assume it could be because certain words are more related to good/bad reviews, which leads to our next hypothesis.

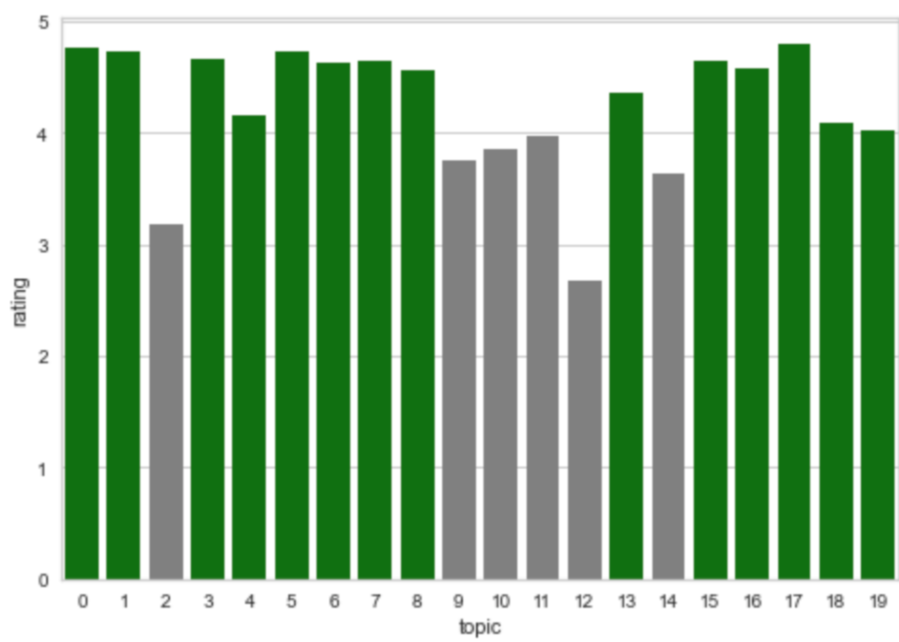
	feature	importance
4	pos	0.0718779
5	compound	0.0710882
2	neg	0.0595928
3	neu	0.0293585
0	nb_chars	0.0162052
1	nb_words	0.0159708
714	word_great	0.0104593
466	word_dirty	0.00987193
146	word_bad	0.00923313
781	word_hotel	0.00875304
1352	word_room	0.00820754
1507	word_stay	0.00799395

Hypothesis d: Certain topics are more related to positive/negative reviews.

We used LDA to build topics. Originally, we tried 5 topics, but there are a lot of overlapping between topics, so we increased it to 20 topics. We also used K-means clustering to get an ideal cluster of 5, but the boundaries are not very clear, so we didn't follow this method. The topics are showed below. We believe we can increase the effectiveness of this part, but due to time constraint we didn't continue altering number of topics and create a clear topic groups by measuring the topic distance or coherence.

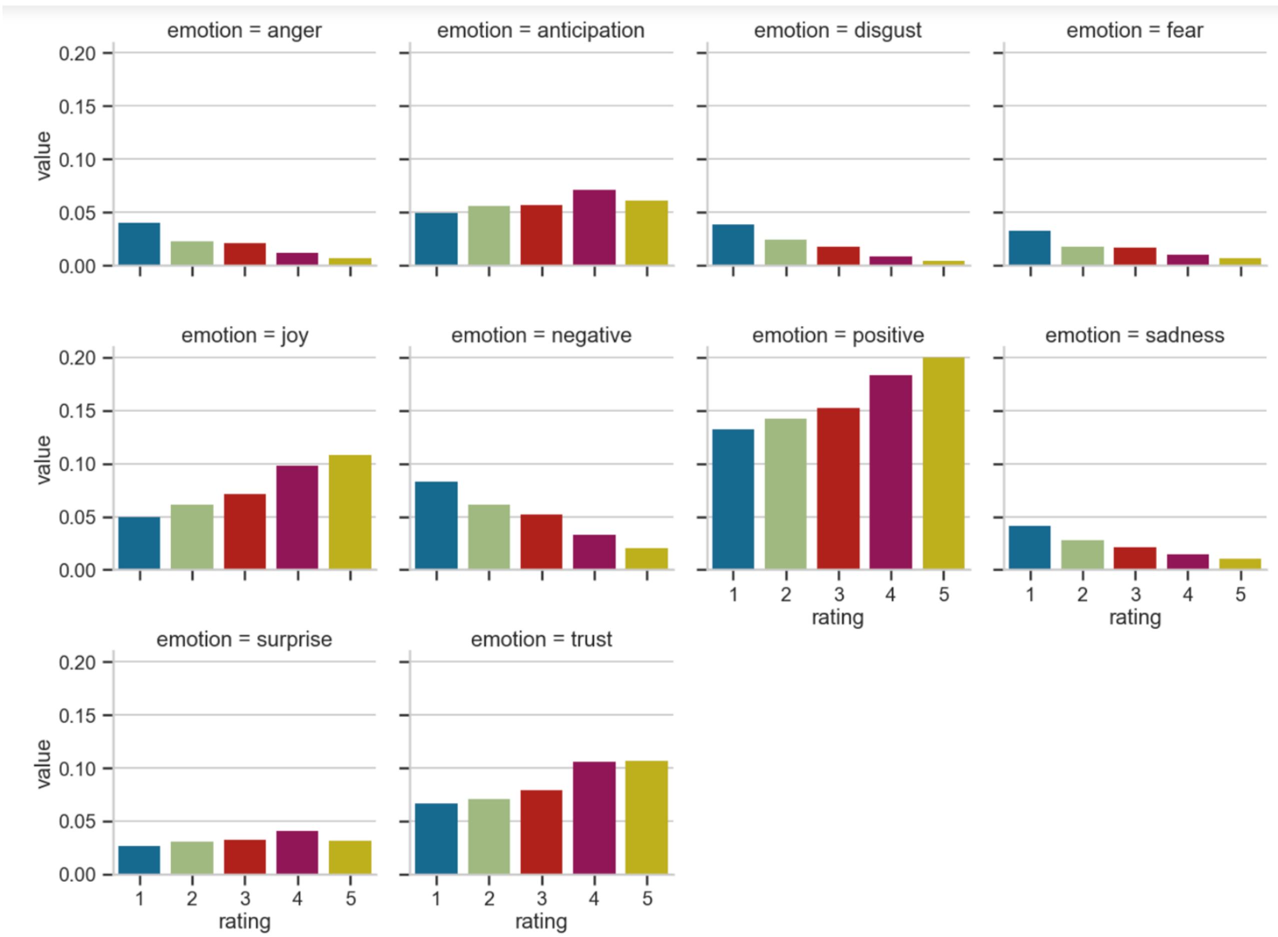
```
Topic 0: staff view hotel amaze amazing wonderful stay bar great make weekend room service beautiful welcome
Topic 1: staff stay hotel friendly helpful clean great comfortable enjoy convention room super really nice center
Topic 2: park hotel valet check charge car pay parking internet free room day cost desk make
Topic 3: recommend staff marriott great hotel highly experience stay excellent marquis property location event atlant
a attend
Topic 4: make like feel stay home hotel room felt away people staff marriott leave treat sleep
Topic 5: great location hotel walk stay place staff downtown distance restaurant business close awesome room clean
Topic 6: locate easy access location hotel near conveniently mission valley convenient diego san area good marriott
Topic 7: stay atlanta marriott hotel marquis time enjoy definitely trip visit look business place marriot wonderful
Topic 8: hotel room nice great clean staff friendly food view restaurant bar comfortable center beautiful fitness
Topic 9: pool park nice hotel parking breakfast expensive area worth little bit pay money night lot
Topic 10: lounge concierge hotel need marriott la jolla breakfast del mar stay area room upgrade close
Topic 11: time wait stay elevator food breakfast great long room restaurant day hotel order check minute
Topic 12: room check desk night hotel tell ask day issue clean bed say work floor hour
Topic 13: staff hotel friendly room clean stay food good professional nice beautiful overall accommodate helpful prop
erty
Topic 14: room bed hotel comfortable bathroom floor small need clean like space use shower microwave guest
Topic 15: excellent location close pool walk restaurant beach area bike hotel shop nice property right place
Topic 16: hotel boston area cambridge ithaca room common right marriott walk check visit lobby restaurant stop
Topic 17: san diego coronado marriott island view resort great bay gaslamp stay hotel relax room place
Topic 18: service good room great excellent hotel food staff location overall customer clean price stay experience
Topic 19: marriott stay property hotel year service best ve member experience time travel rate staff great
```

We plot the average ratings of each topics and it shows difference, so certain topics are more related to positive reviews.



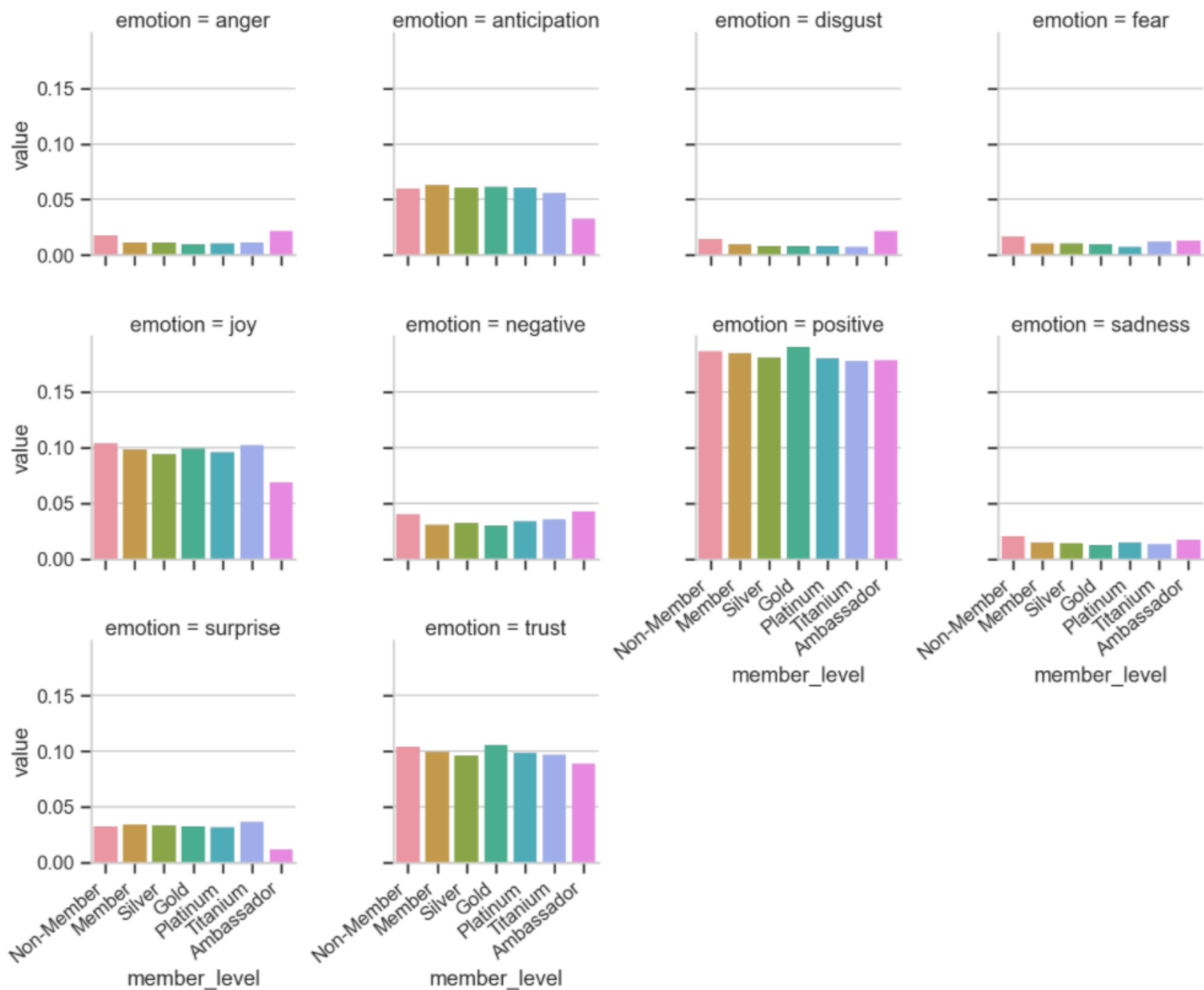
Hypothesis e: Different rating's reviews associate more with certain emotions (based on emolex).

We found clear difference of emotions between positive and negative reviews. The lower the ratings, the higher the emotions of anger, disgust, fear, and sadness. The emotion of anticipation and surprise actually doesn't differ a lot between different ratings.



Hypothesis f: Based on the membership system of Marriott, higher level members have more positive ratings and emotions (based on emolex).

It actually didn't show much average emotion difference between different levels of members. Surprisingly, the highest level, which is ambassador, tend to have higher negative emotions like anger and disgust, but has the lowest emotion value of anticipation. It could be either because the highest level of members tend to have higher expectations regarding the service, or because we have a very limited number of ambassadors in our profile, so the dataset is not large enough to be convincing.



DISCUSSION & CONCLUSION

We got several interesting insights from analyzing this dataset, and we believe there is room for improvement of our models. Some insights rejected our hypotheses, and we think it could be because of lack of efficient data or the use of methods. For some methods, we didn't use statistical methods to testify the significance and just drew the conclusion due to time constraint.

For future work, we have the following thoughts:

- Add more data to our dataset and diversify the profiles (since it's now just Marriott hotels and mostly in San Diego area).
- Increase the preprocessing efficiency by adding new stop words such as "hotel" because it is appearing in every topics of our models.
- Explore more on different types of clustering algorithms and NLP techniques.
- Dig deeper on our topic modeling methods by trying out different number of topics and calculate the distance or coherence between each topic.

Please see our code in attached ipynb file.

Thank you for this wonderful semester!