### Introduction
- Hotel ratings help users understand the reputation of a hotel.
- Ratings provided by users may vary even if the users have a similar opinion of the hotel. This affects the overall rating of a hotel.
- Each hotel has thousands of reviews and it is difficult for users to go through all reviews to understand the pros and cons of the hotel.
- A number of hotel reviews exist in the form of tweets and other social media. Such reviews do not have a rating value.
- Hotel ratings change over a period of time, based on a variety of factors and understanding the rating trend helps the hotel management to plan accordingly.

### Goal
- To predict the rating for each review of a hotel on TripAdvisor and then compute the overall rating of a hotel. This will standardize the rating process based on user reviews.
- Visualize rating trends of the hotel over time.
- Provide review highlights to help users learn about the hotel without much effort.

### Sample Review
"Good stay"
(3.5) Reviewed March 19, 2016, by mobile
We booked this hotel because there was a deal on Hotwire. We were pleased with the staff. The staff was helpful and our room was comfortable. We didn't have children with us and at first I was concerned as the families surrounding us would become annoying but we didn't find it at all. I would recommend this hotel, especially for those traveling with kids.

### Architecture
- Data Extraction From TripAdvisor
- Data Cleaning
- Train Model Using SVM
- Compute Overall Hotel Rating
- Predict Rating of Each Review
- Test Model

### Implementation
- All reviews of "Circus Circus Hotel" were extracted from the TripAdvisor website by Web Crawling.
- Data pre-processing and data cleaning was performed. This step involved the following:
  - Stop words like "the", "and", etc. were removed.
  - Removed markups and punctuation.
  - Stemming was performed.
  - Tokenization was performed.
  - Spellings of incorrectly spelt words were corrected.
  - Words with very low frequency and very high frequency were excluded.
- Used n-gram technique.
- Split the original dataset into training (70%) and testing (30%) datasets.
- Trained the model using SVM.
- Tested the model and predicted ratings of each review.
- Computed overall rating of the hotel from the individual rating values predicted.
- Visualized rating trends.
- Generated a word cloud to understand what's most talked about.

### Results
- The rating prediction accuracy was found to be 91.27%.
- The overall hotel rating predicted by the model is 3.68 and the actual overall rating on the TripAdvisor website is 3.5.
- The hotel's average rating shows a positive trend overall since January 2015.

### Conclusion & Future Work
- The model is able to predict the rating score for a review with an accuracy of 91.27% using SVM.
- The dataset needs to have reviews distributed across all rating classes to train the model effectively.
- The overall hotel rating predicted by the model (3.68) is very close to the actual rating (3.5) on the TripAdvisor website.
- Names of other hotels mentioned in reviews can be extracted to understand the hotel's competitors.
- Reviews in languages other than English can also be included.

### References
1. Aju Thalappilly Scaria, Rose Marie Philip, Sagar V Mehta: Predicting Star Ratings of Movie Review Comments
2. Vikram Elango, Govindrajan Narayanan: Sentiment Analysis for Hotel Reviews
4. Sasank Channapragada, Ruchika Shivashwamy: Prediction of rating based on review text of Yelp reviews