This is a description of the features used in our outlier detection algorithms to calculate the TrueView scores.

This is meant to accompany the paper:

*TrueView: Harnessing the Power of Multiple Review Sites*

http://dx.doi.org/10.1145/2736277.2741655

- **id**: Hotel ID number
- **star_b**: Hotel rating on Booking.com normalized to a 1-5 scale
- **star_h**: Hotel rating on Hotels.com
- **star_t**: Hotel rating on TripAdvisor.com
- **star_int_b**: Hotel rating on Booking.com as integer
- **star_int_h**: Hotel rating on Hotels.com as integer
- **star_int_t**: Hotel rating on TripAdvisor.com as integer
- **num_reviews_b**: Number of review on Booking.com
- **num_reviews_h**: Number of reviews on Hotels.com
- **num_reviews_t**: Number of reviews on TripAdvisor.com
- **cleanliness_b**: Hotel’s cleanliness rating on Booking.com
• service_b: Hotel’s cleanliness rating on Booking.com
• comfort_b: Hotel’s comfort rating on Booking.com
• condition_b: Hotel’s condition rating on Booking.com
• neighborhood_b: Hotel’s neighborhood rating on Booking.com
• value_b: Hotel’s value rating on Booking.com
• cleanliness_h: Hotel’s cleanliness rating on Hotels.com
• service_h: Hotel’s service rating on Hotels.com
• comfort_h: Hotel’s comfort rating on Hotels.com
• condition_h: Hotel’s condition rating on Hotels.com
• neighborhood_h: Hotel’s neighborhood rating on Hotels.com
• num5_b: Number of 5 star ratings on Booking.com
• num4_b: Number of 4 star ratings on Booking.com
• num3_b: Number of 3 star ratings on Booking.com
• num2_b: Number of 2 star ratings on Booking.com
• num1_b: Number of 1 star ratings on Booking.com
• num5_h: Number of 5 star ratings on Hotels.com
• num4_h: Number of 4 star ratings on Hotels.com
• num3_h: Number of 3 star ratings on Hotels.com
• num2_h: Number of 2 star ratings on Hotels.com
• num1_h: Number of 1 star ratings on Hotels.com
• num5_t: Number of 5 star ratings on TripAdvisor.com
• num4_t: Number of 4 star ratings on TripAdvisor.com
• num3_t: Number of 3 star ratings on TripAdvisor.com
• num2.t: Number of 2 star ratings on TripAdvisor.com
• num1.t: Number of 1 star ratings on TripAdvisor.com
• review_mean_b: Mean of review ratings on Booking.com
• review_std_b: Standard deviation of review ratings on Booking.com
• review_mean_h: Mean of review ratings on Hotels.com
• review_std_h: Standard deviation of review ratings on Hotels.com
• review_mean_t: Mean of review ratings on TripAdvisor.com
• review_std_t: Standard deviation of review ratings on TripAdvisor.com
• title_length_mean_h: Mean length of review titles on Hotels.com
• title_length_std_h: Standard deviation of review title length on Hotels.com
• title_length_mean_t: Mean length of review titles on TripAdvisor.com
• title_length_std_t: Standard deviation of review title length on TripAdvisor.com
• review_length_mean_h: Mean length of review text on Hotels.com
• review_length_std_h: Standard deviation of review text on Hotels.com
• review_length_mean_t: Mean length of review text on TripAdvisor.com
• review_length_std_t: Standard deviation of review text on TripAdvisor.com
• good_length_mean_b: Mean of positive review comment length on Booking.com
• good_length_std_b: Standard deviation of positive review comment length on Booking.com
• bad_length_mean_b: Mean of negative review comment length on Booking.com
• bad_length_std_b: Standard deviation of negative review comment length on Booking.com

• cleanliness_review_mean_h: Mean of review cleanliness ratings on Hotels.com

• cleanliness_review_std_h: Standard deviation of review cleanliness ratings on Hotels.com

• service_review_mean_h: Mean of review service ratings on Hotels.com

• service_review_std_h: Standard deviation of review service ratings on Hotels.com

• comfort_review_mean_h: Mean of review comfort ratings on Hotels.com

• comfort_review_std_h: Standard deviation of review comfort ratings on Hotels.com

• condition_review_mean_h: Mean of review condition ratings on Hotels.com

• condition_review_std_h: Standard deviation of review condition ratings on Hotels.com

• neighborhood_review_mean_h: Mean of review neighborhood ratings on Hotels.com

• neighborhood_review_std_h: Standard deviation of review neighborhood ratings on Hotels.com

• num_good_reviews_b: Number of 5 star reviews on Booking.com

• num_avg_reviews_b: Number of 2, 3, and 4 star reviews on Booking.com

• num_bad_reviews_b: Number of 1 star reviews on Booking.com

• num_good_reviews_h: Number of 5 star reviews on Hotels.com

• num_avg_reviews_h: Number of 2, 3, and 4 star reviews on Hotels.com

• num_bad_reviews_h: Number of 1 star reviews on Hotels.com
- `num_good_reviews_t`: Number of 5 star reviews on TripAdvisor.com
- `num_avg_reviews_t`: Number of 2, 3, and 4 star reviews on TripAdvisor.com
- `num_bad_reviews_t`: Number of 1 star reviews on TripAdvisor.com
- `num_good_then_bad_b`: Number of 5 star reviews followed by a 1 star review on Booking.com
- `num_bad_then_good_b`: Number of 1 star reviews followed by a 5 star review on Booking.com
- `num_good_then_bad_h`: Number of 5 star reviews followed by a 1 star review on Hotels.com
- `num_bad_then_good_h`: Number of 1 star reviews followed by a 5 star review on Hotels.com
- `num_good_then_bad_t`: Number of 5 star reviews followed by a 1 star review on TripAdvisor.com
- `num_bad_then_good_t`: Number of 1 star reviews followed by a 5 star review on TripAdvisor.com
- `num_empty_b`: Number of empty reviews on Booking.com
- `num_empty_h`: Number of empty reviews on Hotels.com
- `num_empty_t`: Number of empty reviews on TripAdvisor.com
- `num_susp_zip_reviews`: Number of reviews written by users who have written more than 5 reviews in one postal code on TripAdvisor.com
- `num_susp_date_reviews`: Number of reviews written by users who have written more than 3 reviews on the same day on TripAdvisor.com
- `burst_b`: Max(number of reviews in a given day) - Avg(number of reviews per day) on Booking.com
- `burst_h`: Max(number of reviews in a given day) - Avg(number of reviews per day) on Hotels.com
• burst_t: Max(number of reviews in a given day) - Avg(number of reviews per day) on TripAdvisor.com

• text_sim_b: Measure of the number of repeated sentences per reviewer, aggregated for each hotel on Booking.com

• text_sim_h: Measure of the number of repeated sentences per reviewer, aggregated for a given hotel on Hotels.com

• text_sim_t: Measure of the number of repeated sentences per reviewer, aggregated for a given hotel on TripAdvisor.com

• clique_size: Measure of the maximum number of users that all rated the same group of hotels on TripAdvisor.com

1 Multi-site features

All of these features are the combination of ones described above. Div means that the feature consists of the quotient of the normalized values.

• star_int_b_div_h
• star_int_b_div_t
• star_int_h_div_t
• num_reviews_b_div_h
• num_reviews_b_div_t
• num_reviews_h_div_t
• cleanliness_b_div_h
• service_b_div_h
• comfort_b_div_h
• condition_b_div_h
• num_bad_then_good_b_div_t
• num_bad_then_good_h_div_t
• num_empty_b_div_h
• num_empty_b_div_t
• num_empty_h_div_t
• burst_b_div_h
• burst_b_div_t
• burst_h_div_t
• text_sim_b_div_h
• text_sim_b_div_t
• text_sim_h_div_t

• rating_correlation_b_h: Correlation coefficient between a hotel’s rating distribution on Booking.com and Hotels.com
• rating_correlation_b_t: Correlation coefficient between a hotel’s rating distribution on Booking.com and TripAdvisor.com
• rating_correlation_h_t: Correlation coefficient between a hotel’s rating distribution on Hotels.com and TripAdvisor.com

• Mann Whitney U test b_h rating distribution: p-value of the Mann Whitney rank test of a hotel’s rating distribution on Booking.com and Hotels.com
• Mann Whitney U test b_t rating distribution: p-value of the Mann Whitney rank test of a hotel’s rating distribution on Booking.com and TripAdvisor.com
• Mann Whitney U test h_t rating distribution: p-value of the Mann Whitney rank test of a hotel’s rating distribution on Hotels.com and TripAdvisor.com
• Mann Whitney U test h_t review length distribution: p-value of the Mann Whitney rank test of a hotel’s review length distribution on Hotels.com and TripAdvisor.com

• Mann Whitney U test h_t title length distribution: p-value of the Mann Whitney rank test of a hotel’s review title length distribution on Hotels.com and TripAdvisor.com