HiveTable ReadMe

Introduction: (The code is in HiveRelatedCode.txt)

Line 2 - 5 Create table 'yelp_business' using business dataset
Line 8 - 12 Create table 'yelp_review' using review dataset
Line 15 – 18 Create the schema for the resulting table of joining the above two tables together
Line 21 – 26 Join the two tables by 'business_id' column and insert the resulted table 'yelp_businessReview' into the schema created above.
Line 29 – 30 Find distinct locations from yelp_business table. The result shows the number of locations that contain businesses in them.
Line 33 – 34 Find distinct locations from table 'yelp_businessReview' The result shows the number of locations that contain businesses with reviews.
Line 37- 38 Export table 'yelp_businessReview' for Ngram analysis (Details can be found in Ngram folder)

Line 40 – 54 Create five tables with two columns. One is location number, the other is the number of businesses with five star, four star, three star, two star and one star in the corresponding locations.

Line 56 – 58 Create a table 'business_num' with two columns. One is location number, the other is the number of businesses in the corresponding locations.

Line 60 – 62 Create table 'review_num' with two columns. One is location number, the other is the number of reviews in the corresponding locations.

Line 65 – 67 Create table ‘ngram’ with two columns. One is location number, the other is the mostly used phrase in review in the corresponding location. (Details about the generation of these phrases can be found in Ngram folder)

Line 70 - 72 Create table ‘bizstar’ with two columns. One is location number, the other is the average star of the corresponding location. The formula used to calculate the average star is: (Details can be found in AvgBizStar folder)

$$\text{AvgStar} = \frac{\sum_{businesses\ b\ within\ the\ location} b.reviewcount * b.star}{\sum_{businesses\ b\ within\ the\ location} reviewcount}$$

Line 75 – 78 Create table ‘bizreview_sentiment’ with results of sentiment analysis. (Details can be found in SentimentAnalysis folder)

Line 80 – 92 Create four tables about the sentiment analysis. The following describes the tables. (Details about calculating these values can be found in SentimentAnalysis folder)

1. senti_common: There’re four columns. One is location number. The other three are the average negative, neutral and positive values in the area.
2. senti_positive: Two column. One is location number. The other is the number of businesses with positive sentiment value in the corresponding location.
3. senti_negative: Two column. One is location number. The other is the number of businesses with negative sentiment value in the corresponding location.
4. senti_neutral: Two column. One is location number. The other is the number of businesses with neutral sentiment value in the corresponding location.

Line 95 – 109 Create the feature table ‘yelp_feature’ from the above table by joining on location number.
There’re 17 features in all.
The final schema is as the following:

col_name	data_type	comment
avgstar	double	
location	smallint	
business_count	bigint	
reviewcount	bigint	
five_star_count	bigint	
four_star_count	bigint	
three_star_count	bigint	
two_star_count	bigint	
one_star_count	bigint	
senti_neg_avg	double	
senti_pos_avg	double	
senti_neu_avg	double	
senti_comp_avg	double	
senti_compound_pos_count	bigint	
senti_compound_neg_count	bigint	
senti_compound_neu_count	bigint	
phrase	string	