Wrangle Report

Jinsoo Kim

9 June 2018

Quality issues.

1. **twittter_archive** file

    1) Delete retweet data

        Some tweets are retweeted which have the identical photo and ratings. Since duplicated data are not needed, I removed.

    2) Many of values in 'expanded_urls' are empty, None. As noticed in 'Project Motivation'

tweet_id is the last part of the tweet URL after "status/" → https://twitter.com/dog_rates/status/889531135344209921

        For empty 'expanded_urls' I put the tweet_id at the last part after *https://twitter.com/dog_rates/status/*.

    3) There was one identical expanded_url except retweets. This is rating same dog twice. So I deleted later one.

    4) In 'timestamp' +0000 is in every data and useless. So I removed.

    5) 'source' columns' variables including urls are too long. I replaced with short ones.

    6) Some tweets have decimal values for 'rating_numerator'. I checked in 'text' one by one and modified in needed cases.

    7) Like 9/11 and 12/10, some rows have two ??/?? typed texts in 'text'. This may cause mismatching for 'rating_numerator' and 'rating_denominator'. I checked one by one and modified in needed cases.

    8) Some 'rating_denominator' have awkward values. I checked one by one and modified in needed cases.

    9) Dog names in 'name' like 'a', 'an' may not be the real names. I checked one by one and modified in needed cases.

    10) Changed datatype of 'tweet_id' as string. Because this is not for calculating.

11) Changed datatype of 'source' as category. Because it only consists of 4 types.

12) Changed datatype of 'rating_numerator' as float. Because this is for calculating.

13) Changed datatype of 'rating_denominator' as float. Because this is for calculating.

2. **predictions** file

   1) Changed datatype of 'tweet_id' as string. Because this is not for calculating.

3. **tweet_json** file

   1) Changed datatype of 'tweet_id' as string. Because this is not for calculating.

Tidiness issues.

1. In **twitter_archive** file, merge into 'dog_breeds' and remove 4('doggo', 'floofer', 'pupper', 'puppo') columns.
2. Consolidate 3 files('twitter_archive', 'predictions', 'tweet_ids') into '**merged_tweets'**