Data preparation

In data preparation, we need to import pandas firstly. Then, we can use pd.read\_csv() function to read csv file ’ Sony\_Playstation\_tweets.csv’.

After load data set, we need a descriptive analysis of data. We use following function for analysis:

1. Head(): get the first 5 rows
2. Describe(): calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame
3. Info(): prints information about the DataFrame
4. Isnull().sum(): total number of null
5. Size: find dataframe size
6. Shape: find dataframe shape
7. Ndim: find the number of dimensions

So we observer the first 5 rows of dataframe. And understand some statistical data and data type for each column. Size is 699048. Shape is (19418, 36). Dimension is 2. In the missing value check, we found that some columns have a large number of missing values. For example, ‘place’ has 19417 missing value. We decide to drop these columns.

To delete these columns, we use drop() function. We keep these columns:(id, date, username, tweet, language, replies\_count, retweets\_count, likes\_count, hashtags).

At last, we use isnull().sum() to check the data after cleaning. We can found there is no missing value on the dataset. Data preprocessing is complete, we can proceed to the next step.