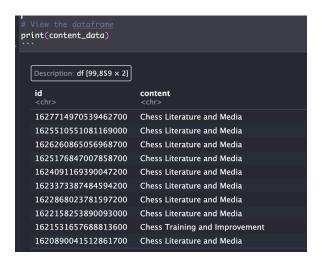
Problem: Using the content classification text result from Paul, I tried to match it with the tweet in the rachel\_new\_chess\_community. It shows that there should be over 90,000 tweets being classified, however, when I extract the tweet and the corresponding ID from the original JSON file to merge with the classification, only about 4,000 tweets are being identified successfully.





#### Steps:

- 1. For those tweets, identify the influencer in 2 ways:
- Users with the rank of less than 10
- Manually chosen the 10 influencers in the previous project
- 2. Generate tables that contain the tweets created by influencers and consumers with the content type (e.g. like the table below)

Α	В	С	D
created_at	content	user_type	count
2020-06-29	Chess Events and Tournaments	consumer	1
2020-06-29	Political and Social Issues in Chess	consumer	1
2020-06-30	Chess Events and Tournaments	consumer	2
2020-07-01	Chess Literature and Media	consumer	1
2020-07-01	Social Media Impact on Chess	consumer	1
2020-07-02	Chess Training and Improvement	influencer	1
2020-07-02	Social Media Impact on Chess	consumer	2

- 3. Transform the table into 3 columns: the date, number of tweets by consumers, number of tweets by influencers. I have selected the top 2 classified content types: Chess Events and Tournaments & Chess Training and Improvement. Here is the sum of tweets in those 2 contents.
  - Problem: there are some dates that have 0 tweets generated by consumer nor influencer. Should we just not include such dates in the time series, or should we fill out with 0, or should we smooth the line to have a reasonable guess on such dates (like impute missing values with linear interpolation)? Here I just ignore such dates, to only use those dates with non-sparse date to form the time series.

A	В	С
created_at =	consumer =	influencer =
2020-06-29	1	0
2020-06-30	2	0
2020-07-02	0	1
2020-07-03	1	0
2020-07-10	1	0
2020-07-11	1	0
2020-07-12	0	1
2020-07-13	0	1

4. Create time series on the dataset and perform the Granger test. The result is not so good since I got p-values larger than 0.05 for every time lag I tested (lag from 1 to 10)

Improvement: Here the first experiment I did is for all the influencers with top 10 rank aggregated. I haven't tested the result for a single influencer yet. Also, I can do the same experiment for the other influencer identification (manually selected the influencer id in the previous project). The missing values in the time series might also affect the results.

## Experienment 2:

Select the top 10 content types with the supply from the top 2 core agents. (by rank). The Result is not good. Every p value (1-10 lag) is larger than 0.05.

## Experiment 3:

Select single core agent (the top core agent manually selected by user id) with the top 10 content types. The relationship is not significant: every p value is larger than 0.05.

# Experienment 4:

Select top 10 rank as an influencer, get the aggregated tweet counts for consumer ang influencers. Get the panel data with index: content type, the best result was lag = 7 (p=0.14).

#### Experiment 5:

Select 2 influencers (rank at 2 and 4), and get the top 10 content types tweet data. The result shows at lag 2 the causality is significant but in other lags the p values are larger than 0.05.