**Data Collection**

First I downloaded the tweets containing the hashtag "#healthy" ranging from February 21, 2018 to February 28, 2018 using the GOT. GOT is an open source platform for downloading old twitter data. I got around 23515 tweets containing “#healthy” hashtag. Then using pandas’ value\_counts built-in method I calculated the frequency of the all hashtags presented in the tweet. Then I took top 25 frequent hashtags (get\_hashtag\_frequency.py) and downloaded again twitter data for the same date range using GOT. After removing the duplicates we got in total 256860 tweets.

**Analysis**

**Most Common Hashtags and Frequent Tweeters**

After collecting the data, I extracted the the common 50 tweets. During tweet collection, I also collected tweeter id (id of the person who tweets). I calculated their frequency using pandas value\_count() function and extracted the top 50 tweeters.

**Extract Hashtags which are used together**

To get the hashtags which are used together more often than others I applied association rule mining. I wrote my own methods for association rules mining which follows apriori principle. I used pandas dataframe for storing the tweets where the hashtags present in a tweet is stores as a column. I used only this column for association mining. First I generated candidate dataset containing 1 hashtag only. Using this I calculated the frequent itemset based on the provided support. The choose of support and confidence is mainly trial and error basis which will be explained later. Then I calculated the frequent itemset containing more than 1 hashtag. Based on the support and confidence I applied apriori principle to prune out infrequent subset. Then I generated the rules considering the subset for each frequent item and based on support and confidence I kept only the valid rules. For the valid association rule, I also calculated the conviction and interesting value.

**Choice of Support And Confidence:** The choice of support was made based on trial and error. But prior that I made a small calculation. One of the top frequent hashtag “#healthy” has the support around 0.04. As the number of tweets containing top 25 hashtags is much larger than the frequency of the individual hashtag, I chose a small number for support. I tried with support = 0.02 and confidence = .6 but only got 6 association rules and the frequent itemset counted up to 2 elements only. Then I tried with support = 0.019 and confidence = .6,. I got a frequent itemset containing up to 6 elements itemset and 571 association rules. I chose this support and confidence value for this project as it’s giving better result.

**Skyline Hashtag Rules**

From the association rules, generated in the previous step, I calculated the skyline for the 2,3,4,5 element frequent itemset. I loaded the association rules in pandas dataframe. Then for each row I check whether any row is dominated by any other rows in terms of confidence, conviction and interest. When I found any row that is not dominated by any other row, I declared this as skyline rule. Then stored them in a csv file.

**Observation and Findings**

When I chose the support value 0.02 I got only 6 association rules with frequent containing up to 3 elements. But when I chose the value as 0.019 I got 571 rules with the frequent itemset containing up to 6 elements. So, there is a lot of hashtags of support between 0.019 and 0.02 which got pruned due to choice of support. There are lot of rules with confidence 1. For example: ['#bakeryfreshfriends']==>['#granola'], ['#granola','#sunbeltbakery']==>['#bananabread'], ['#sunbeltbakery']==>['#bananabread', '#granola', '#recipe']. There are a lot of tweets containing hashtags '#granola', '#recipe', '#bananabread' which suggests that many tweets are tweeted for providing interesting recipes made of granola and banana-bread. The tag “#contest” is also very much frequent with the tag “#recipe”, and it’s explanation can be many tweets coming from organizations which arrange recipe contest.

For skyline rules, there are lot of rules fall in the skyline with the same value because the number of total tweets is much much bigger than the frequency distribution of each tweet and as we count up to 3 decimal points for support and confidence, lots of hashtags got the same value.

**Challenges**

The first challenge I faced was to collect twitter data. Twitter API allowed to download tweets for past 7 days only. I tried some existing python package for downloading tweets but none of them seem working perfectly. And I got less number of data because twitter API data limit. Then I used GOT twitter data scraper which is open source and easy to use. It uses twitter advanced search for scrapping the data, so it doesn’t have any data limit problem. Moreover, I found the data more accurate than other packages.

The second challenge I faced during association rule mining. It was taking a long time to calculate the candidate\_1 itemset. I then reduced the size considering the tweets which has more than 500 frequency. Because tweets less than that frequency will be pruned automatically due to less support. It made the program faster. I also struggled to find a suitable support value. I had to run the program several times for finding a suitable value.