CSCI 581 Project 1: Analyzing Health Related Micro-blog

Description:

This program was created in python 3.6. It uses Twitter API to request twitter data and uses "python-twitter" package to collect the data. It should be installed to collect the data. To see the results run run.py. Uncomment the blocked comments in the files to run the heavy functions.

The thresholds are selected by tuning it with the results. The goal was to get the results between 20-30. I used csv file with 6 attributes because it was easier to visualize the data. Use of tuples while creating the frequency distribution was because of list not able to be a key in dictionary. For the skyline minimum approach was taken.

Observation: Fitness seems to be used with health a lot. It seems like people take physical health as the health than mental or emotional health.

Part 1: Data Collection

1. twitterapp.py has the api and code to run the program. The collected data are stored in data\collection.csv. This file has 6 Attributes and has 113645 rows of data. The total file is of 33.3MB.
2. In the collection.csv file, the attribute 'hashtags' is for all the hashtags used in the tweet. It is saved as a list of strings. The frequency distribution is given by association.
3. The line 100-103 gives when uncommented collects the data from twitter. The collected data can be found in data\top_25.csv.
4. Due to data-collection time being very high. Top 25 hashtags are collected only for a day's worth of data.

Part 2: Analysis

5. S
    a. The frequency distribution of hashtags are generated in hashtags.py
    b. The frequency distribution of users are generated in hastags.py
    c. This is generated in association.py
    d. This is generated in skyline.py using skyline50.csv (generated by code)

Challenges:

1. Collecting data was time consuming. The programs runs quick enough for first hundred iterations but it gradually slow down. It was quicker to get data by terminating the program and starting again. And collecting 7 days of data for health took about 2 days (laptop was moved).
2. This project was fun to do but it was lengthy. Working with big data with no application of reducing methods was time consuming. I used pickle to save the data and work with it again so that I didn't have to wait minutes to read all the files.