README

Yin Zhao, Yue Mao

4/21/2020

a. Team Members

- Yue Mao (ym2749)
- Yin Zhao (yz2426)

b. Files

- main.py
- job_data.csv
- sample_run.txt
- README

c. How To Run

```
python3 main.py job_data.csv <min_sup> <min_conf>
```

d. Dataset Description

- We used "NYC Jobs" dataset (https://data.cityofnewyork.us/City-Government/NYC-Jobs/kpav-sd4t)
- Preprocessing of the dataset includes removing columns that do not contain interesting information we need:

```
colums_to_drop = ["Division/Work Unit", "Work Location 1", "Recruitment Contact", "Post Until", "Hours/S
```

and removing rows that contain empty fields. This leaves 1688 rows in the final dataset.

- The dataset is interesting because it contains job category, status, level, as well as requirement and job description, which would give information about what requirement a certain field needs, what kind of work is involved in certain levels, etc.

e. Description of Internal Design

The program passes through the rows (market baskets) and finds items that have min support. Then generates new candidates based on subsets of the seed items. For each candidate set, compute its support and adds to the items set. Then it generates rules for the items with confidience higher than min confidence.

f. Sample run

- command:

```
python3 main.py job_data.csv 0.57 0.5
```

- This generates interesting and revealing result because it reveals the patterns such as how the details of the job would be discussed in various locations, what is the residency requirement for variosu jobs, and that annual pay is almost always associated with fulltime positions.