Overview

In cloud environment, each machine has a unique id, unique IP and is associated with multivalued fields like securitty groups, subnets and so on. More specifically

1. a machine is assiciated with set of subnets, and security groups
2. a machine is identified by the machine id and IP address

The data is available in the following format

```
machineid|ip-addr|security-group|subnet
mch_c0c46596|172.31.0.1|sg_10,sg_3,sg_2,sg_9,sg_4,sg_7|net_6,net_5,net_7,net_3,net_9,net_4
mch_c0c97964|172.31.0.2|sg_6,sg_10,sg_8,sg_4,sg_1,sg_7,sg_5,sg_3,sg_9|net_4,net_7,net_9
mch_c0cde904|172.31.0.3|sg_9,sg_6,sg_5,sg_4,sg_7,sg_1|net_3
mch_c0d1fba2|172.31.0.4|sg_2,sg_7,sg_5,sg_3,sg_6,sg_8|net_5,net_8,net_10
mch_c0d603dc|172.31.0.5|sg_3,sg_9|net_5,net_2,net_3,net_10,net_6,net_1,net_4
mch_c0da15a8|172.31.0.6|sg_6,sg_8,sg_1,sg_5,sg_7|net_7
mch_c0dd84ea|172.31.0.7|sg_1,sg_7,sg_3,sg_2,sg_4,sg_9,sg_5,sg_6|net_4,net_1,net_10,net_5,net
mch_c0e171d6|172.31.0.8|sg_5,sg_3|net_3,net_2,net_1,net_8,net_5
mch_c0e4e32a|172.31.0.9|sg_7,sg_2,sg_8,sg_5,sg_9,sg_6,sg_4|net_1,net_5,net_8,net_9,net_2,net
mch_c0e8a26c|172.31.0.10|sg_3,sg_5,sg_10,sg_8|net_5,net_3
```

As you can see, the fields are pipe separated and the fields security-group and subnet are multivalued. This data doesn't lend itself for easy analysis.

Goal

The goal is to explode the data so that these multivalued columns have single values. For e.g

```
the following example row
mch_c0c46596|172.31.0.1|sg_10,sg_3,sg_2,sg_9,sg_4,sg_7|net_6,net_5,net_7,net_3,net_9,net_4

should be transformed to

machineid|ip-addr|security-group|subnet
mch_c0c46596|172.31.0.1|sg_10|net_6
mch_c0c46596|172.31.0.1|sg_10|net_5
mch_c0c46596|172.31.0.1|sg_10|net_7
mch_c0c46596|172.31.0.1|sg_10|net_3
mch_c0c46596|172.31.0.1|sg_10|net_9
mch_c0c46596|172.31.0.1|sg_10|net_4

mch_c0c46596|172.31.0.1|sg_3|net_6
mch_c0c46596|172.31.0.1|sg_3|net_5
mch_c0c46596|172.31.0.1|sg_3|net_7
mch_c0c46596|172.31.0.1|sg_3|net_3
```

```
mch_c0c46596|172.31.0.1|sg_3|net_9
mch_c0c46596|172.31.0.1|sg_3|net_4

...
so on
```

In summary, for each multiple valued fields, we are creating multiple rows for the machine. In the above example, we can assume, that machineid and ip-addr are single valued

Base Assignment

Write a python program to convert the given file to the required format. The program structure would be approximately as follows

1. command line parsing for accepting the input file
2. parse the given file
3. determine which columns are multivalued
4. determine which are single valued
5. for each row explode the row and create the rows based on each of multi-valued fields.

Analysis

Based on the above generated rows, use pandas to calculate some frequency counts

- count of usage of each subnet
- count of usage of each security group
- count of usage of subnet + security group combination (optional)

Advanced (Optional)

We have given a simple bash script to generate data sets for the above example. document the bash script to the best of your abilities

```
# how to run the program
# the program needs jq utility to be be installed on the machine.
bash datagen.sh
```

Evaluation Criteria

- Working program for Base Assignment
- Variable names and program structure
- Error handling as appropriate
- format the code using black utility
- a pylint score above 7.0

Tips

- use any python library as suitable
- please refrain from using chatgpt for documenting, while it does document well, the expectation is that the author is able to explain the design choices and language usage.
- use python 3.8 and above
- Using pylint and black shall show the errors upfront, to help easier development.
- for easier development, write the functions in single file.
- please do document the functions as required for easier understanding.
- understanding the provided bash program shall help you to test the program better. This is entirely optional.