Data Science test

Welcome future sensego talent, here are two problems we would like you to solve to acknowledge you level both logical and computational. There is no expected form asked for the result but it should at least contains the code used to answer the questions so some explanations along.
If you can't answer some of the questions, don't panic this is not eliminatory and if you have any question feel free to send me a mail at michael.weiss@sensego.fr .
The test was designed to last between 2 and 4 hours, but of course it is hard to evaluate it when you already have the answers so don't hesitate to tell me if you had to stop because it was taking you too much time or simple give me an estimation of the time needed to try and answer all the questions.

Problem 1 : Warm up

There are zombies in Seattle. Liv and Ravi are trying to track them down to find out who is creating new zombies in an effort to prevent an apocalypse. Other than the patient-zero zombies, new people only become zombies after being scratched by an existing zombie. Zombiism is transitive. This means that if zombie *0* knows zombie *1* and zombie *1* knows zombie *2*, then zombie *0* is connected to zombie *2* by way of knowing zombie *1*. A zombie *cluster* is a group of zombies who are directly or indirectly linked through the other zombies they know, such as the one who scratched them or supplies who them with brains.

The diagram showing connectedness will be made up of a matrix of values 0 or 1. The value matrix[i, j] indicates if the zombie i knows the zombie j. For example, in a world of 3 zombies the complete matrix of zombie connectedness could be:

```
110
110
001
```

Zombies 0 and 1 know each other. Zombie 2 does not know anyone.

Question:

Your task is to determine the number of connected groups of zombies, or *clusters*, in a given matrix square matrix of size n.

1) Write a function *zombieCluster* that compute this number without using any other library than the standard ones. Typically you method should take as input a list of lists representing the matrix and should output an integer.

We provide some examples in the file *problem1_test.txt* where you will find matrices and the corresponding results your function should return. In addition if you are coding in python we provide the script *problem1_test.py* that allows you to test your function by writing in it and then running the script.

Constraints:

You may consider these constraints to prevent the function testing to use too much memory and take too much time.

- $1 \leq n \leq 300$

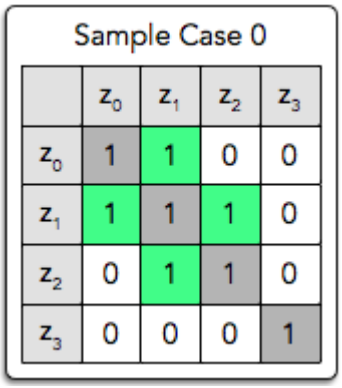
Examples :

Sample Input 0

```
1100
1110
0110
0001
```

Sample Output 0

```
2
```



In the diagram above, the squares highlighting a known connection between two different zombies are highlighted in green. Because each zombie is already aware that they are personally a zombie, those are highlighted in grey.

Explanation 0

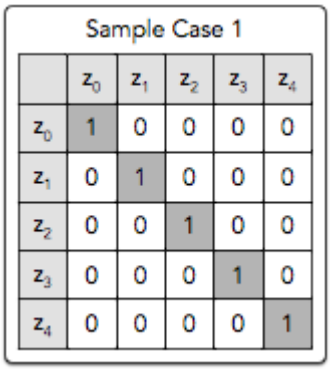
We have *4* zombies numbered . There are *2* pairs of zombies who directly know each another: *(0, 1)* and *(1, 2)*. Because of zombiism's transitive property, the set of zombies *{0, 1, 2}* is considered to be a single zombie cluster. The remaining zombie - *3* -, doesn't know any other zombies and is considered to be his own, separate zombie cluster *{3}*. This gives us a total of *2* zombie clusters.

Sample Input 1

```
5
10000
01000
00100
00010
00001
```

Sample Output 1

```
5
```



Explanation 1

No zombie knows who any other zombie is, so they each form their own zombie cluster: *{0}, {1}, {2}, {3}, and {4}*. This means we have *5* zombie clusters, so we print *5* on a new line.

Question (bonus):

2) In mathematical terms what does *zombieCluster* do ? Do you already know one or more libraries that can do that in the language you used ? Find a library that does that, and use it to write a test function to show that you get the same results using your function and the one from the library.

Problem 2 : Sensego's everyday work

With the given file we can observe during a month the WiFis **seen** by some mobile devices. Here we limited it to 30 devices.
A clear understanding of the given data is important :

- WiFis seen by the phone are not necessarily WiFis to which the device connected to, device often scan the surrounding WiFis to pentially connect to one of these.
- At a given time t, the phone will not necessarily send all the wifis seen by it but potentially only some of it. Most of the time when a phone is connected to a wifi, if will simply send the wifi alone and not all the wifis seen
- The phone will send wifis info at a random time. For example we can observe several hours without any wifi lines. This doesn't necessarily mean no wifi is around but can also say that the application in charge of getting the wifi informations got stopped by the OS or that the wifi is disabled on the phone.

Here are the values you have for every line :

Suid : unique identifier of a mobile device
ssid : name of the wifi
bssid : MAC address of the wifi, this should be a unique identifier of the wifi
sensing_ts : time at which the phone saw the wifi
rssi : Strength for the receive signal of the wifi

Hence a line in this file can be read like this :

at the time **sensing_ts** the device **suid** was seing the wifi **ssid** with the mac address **bssid** with the signal strength **rssi**.

Questions:
Using these data, we want to answer to these problems:

3) Detect when someone is at home or at work.
4) Detect when someone is moving between to separated places
5) What is the average percentage of time when someone is detected at home/at work overall
6) Draw the distribution of these percentages per user
7) What is the average percentage of time when someone is detected at home/ at work over the time where we get wifi infos
8) Does one of these statistics seem accurate to you ? Why ?
9) Imagine a solution where we would have a vision of a device with intervals telling us when a device is being moved or stays in a given place.
10) How would you try to label the intervals where a device staying in a given place

After having answered to theses problems we would like to see some basic statistics :

Instructions:

All the answers should be based on what knowledge we have from the wifis and some functions could be WiFi-wise. For example in 3) we could have home corresponds to wifi=x for suid=y. But we should also have in the end time intervals telling us when someone is at home and so on...
Any of the two first questions can be answered separately or together. The definition of the **used notions** such as **work**, **home** or the **fact of moving** will obviously change the output of your answer so there should be a clear definition of it in your code or in the answers to the questions. What we mean by **separated places** is two places that can be geographically distinguished, for example you would have trouble to distinguish your bedroom from your kitchen so two room of the same flat wouldn't be considered as separated places.