Task:

Due to the active expansion of communication services, many (not all) people buy more than one SIM card or use corporate SIM cards. The company is faced with the task of identifying subscribers (msisdn), which are, actually, one Person (person).

To solve this problem, it is proposed to use data about the movements of subscribers, collected from base stations. Moves of mobile subscribers are registered at base stations. Registering a subscriber at a base station means staying within the coverage area of a given base station. The operation zones of different base stations intersect. At the same point on the ground, there are usually several base stations that either "look" in different directions or support different standards (2G / 3G).

Required:

A) To generate the algorithm for determining Person based on the input data and present it:

A1. In the language R (send the code in * .txt format. Comments on the text of the code are desirable)

A2. Presentation in the format * .ppt or * .pdf, which briefly state the essence of the algorithm

B) To form the essence of the Person according to the given data set.

The result should be presented in a * .txt format file with two fields:

1. Person`s Identifier (generate yourself, eg 1,2..N);

2. Phone numbers from the input data that the Person concerned are likely to own

Input data:

A. Information about the movement of a random set of subscribers **(file 02_Data.csv)** for the period May 23-25, 2013, in which telephone numbers belonging to the same person.

Description of data fields:

1. Lac - identifier to a group of base stations
2. cid – identifier to the base station (unique within the Lac framework)
3. msisdn - the phone number of the subscriber (in our case, encrypted)

4. imei - http://en.wikipedia.org/wiki/IMEI

The first 8 digits imei are called TAC and determine the device that the subscriber uses. The shortened TAC directory **(device file 03_)** consists of fields:

- tac - directly TAC itself
- vendor - phone manufacturer
- platform - the operating system on the phone
- type - device type (Phone - phone, SmartPhone - smartphone, Laptop - tablet, laptop, Modem - modem, etc.)

5. event_type - each registration on the base station has its own type. Directory types of registrations with a decryption in the **file 04_event_type.** The directory is received from the technical service, completely understand everything that is written in it is not required.

6. tstamp - timestamp = the number of 1/1000 seconds that passed from 01/01/1970 to the time of registration on the BS.

7. long - geo longitude of base station placement

8. lat-geolocation latitude of the base station

9. max_dist - the maximum distance of reception in meters

10. cell_type - type of placement (can be METRO - station underground, INDOOR - the station is located inside a shopping center or something like that, OUTDOOR - street location).

11. start_angle - the beginning of the coverage angle of the station (0-360 clockwise, starting from the direction to the north)

12. end_angle - the end of the coverage angle of the station (0-360 clockwise, starting from the direction to the north)

B. **File 01_Facts** - training sample: a set of phone number pairs (msisdn), each of which is likely to belong to one Person. Data on the movement of msisdn data is present in the **file 02_Data.csv**