**Задача:** В связи с высоким уровнем проникновения услуг сотовой связи многие (не все) люди покупают для себя более, чем одну сим-карту, либо пользуются служебными сим-картами. Перед ОАО «МегаФон» стоит задача определить абонентов (msisdn), которые являются одной и той же Персоной (человеком).

Для решения данной задачи предлагается использовать данные о перемещениях абонентов, собранные с базовых станций. Перемещения абонентов мобильной связи регистрируются на базовых станциях. Регистрация абонента на базовой станции означает нахождение в зоне действия данной базовой станции. Зоны действия различных базовых станций пересекаются. В одну и ту же точку на местности обычно ставят несколько базовых станций, которые либо «смотрят» в разные стороны, либо поддерживают различные стандарты (2G/3G).

**Требуется:**

**а) сформировать алгоритм определения Персоны на основании входных данных и представить его:**

**а1. на языке Python (присылать код в формате \*.ipynb. Комментарии по тексту кода желательны)**

**а2. Презентацию в формате \*.ppt или \*.pdf, в которой кратко изложить суть алгоритма**

**б) сформировать сущность Персона по предоставленному набору данных. Результат должен быть представлен в файле формата \*.txt с двумя полями: 1. Идентификатор Персоны (сгенерировать самостоятельно, например 1,2..N);**

**2. Номера телефонов из входных данных, которые данной Персоне с большой долей вероятности принадлежат**

**Входные данные:**

А. Информация о перемещении случайного набора абонентов (**файл 02\_Data.csv**) за период 23-25 мая 2013 года, в котором точно присутствуют номера телефонов, принадлежащие одному и тому же человеку. Описание полей данных:

1. Lac – идентификатор группы базовых станций
2. cid – идентификатор базовой станции (уникален в рамках Lac)
3. msisdn – номер телефона абонента (в нашем случае зашифрованный)
4. imei - http://ru.wikipedia.org/wiki/IMEI

Первые 8 цифр imei называются TAC и определяют устройство, которым пользуется абонент. Сокращенный справочник TAC (**файл 03\_устройства**) состоит из полей:

- tac – непосредственно сам TAC

- vendor – производитель телефона

- platform – операционная система на телефоне

- type – тип устройства (Phone – телефон, SmartPhone – смартфон, Laptop – планшет, ноутбук, Modem – модем, etc.)

5. event\_type – каждая регистрации на базовой станции имеет свой тип. Справочник типов регистраций с расшифровкой в **файле 04\_event\_type.** Справочник получен от технической службы, полностью понять всё, что в нем написано не требуется.

6. tstamp – timestamp = количество 1/1000 секунды, прошедших с 01.01.1970 до времени регистрации на БС.

7. long – гео долгота размещения базовой станции

8. lat – гео широта размещения базовой станции

9. max\_dist – максимальная дистанция приема в метрах

10. cell\_type - типа размещения (может быть METRO - станция под землей, INDOOR - станция размещена внутри какого либо торгового центра или чего-то подобного, OUTDOOR - уличное размещение).

11. start\_angle - начало угла покрытия станции (0-360 по часовой стрелке, начиная с направления на север)

12. end\_angle - конец угла покрытия станции (0-360 по часовой стрелке, начиная с направления на север)

В. Файл **01\_Факты –** обучающая выборка: набор пар номеров телефонов (msisdn), каждая из которых с большой долей вероятности принадлежит одной Персоне. Данные по передвижениям данных msisdn присутствуют в файле **02\_Data.csv**