**Отчет**

Изучение полученого дампа было выполнено в следующем порядке:

1) Общий анализ данных (наиболее/наименее часто встречающие значения, количество уникальных значений)  
2) Анализ IP адреса  
3) Анализ user-agentа  
4) Сопоставление идентификаторов устройств  
5) Сопоставление логинов  
6) Построение графа связей

**Общий анализ данных**

В результате было получено, что:

* больше всего запросов было выполнено с IP 188.234.164.151
* самая популярная платформа -- Android
* самый популярный браузер Google Chrome
* есть некоторое количество неповторяющихся данных

**Анализ IP адреса**

Для анализа IP адресов были выбраны следующие инструменты:

* пакет Python ipwhois для извлечения whois-данных. Наиболее популярный и хорошо документированный.
* библиотека netaddr для работы с IP адресами.
* API ресурса sypexgeo.net для получение географической информации с помощью базы данных Sypex Geo Max. Был выбран именно это метод получения местоположения в связи с его простотой и отсутствием необходимости проверки актуальности баз данных.

В качестве опорных данных были выбраны следующие: страна, город, долгота, широта, подсеть, имя подсети.

**Анализ user-agentа**

Для парсинга user-agent было выбрана библиотека ua\_parser из-за простоты использования полученных данных. Были получены семейства операционной системы и браузера, а также их версии. На основе этих данных было сделано предположение, является ли устройство мобильным.

Собранные данные были объединены с данными об идентификаторе устройства и логине, и записаны в три файла, содержание которых соответсвует названию файлов: mobile\_devices.csv, not\_mobile\_devices.csv, all\_info.csv.

**Сопоставление идентификаторов устройств**

При поиске идентификаторов устройств, принадлежащих одному и тому же реальному устройству, были выдвинуты следующие преположения:

* устройство находится в одной подсети
* устройство не сменяет операционные системы
* устройство не подменяет user-agent
* устройство может сменить/обновить браузер
* IP адреса в подсети могут быть динамическими

Таким образом, использовался следующий алгоритм:  
1) устройства делились на мобильные/не мобильные на основе анализа user-agent  
2) далее поиск ограничивался одной подсетью  
3) записи с данной подсетью сортировались по ОС и ее версии  
4) для каждой группы ОС определялось два словаря: с ключами в виде IP адресов и с ключами в виде семейства браузера.  
5) если IP адресс был только один, то полученная на данном этапе информация относилась к одному устройству.  
6) в противном случае находились пересечения между записями, сгруппированными по IP адресам, и записями, сгруппированными по браузеру. Каждое пересечение определялось как одно устройство.

В результате были получены множества идентификаторов, с определенной долей вероятности принадлежащих одному устройству.

Из-за особенности алгоритма необходимо было еще раз обработать полученные данные, так как некоторые данные могли дублироваться.

**Сопоставление логинов**

Для каждого полученного устройства по его списку идентификаторов были определены связи с логинами. Было установлено, что некоторые логины принадлежат как мобильным, так и не мобильным устройствам.

Данные были записаны в две таблицы: not\_mobile\_devices\_ids.csv, mobile\_devices\_ids.csv.

**Построение графа связей**

По вышеописанным данным был построен граф, в котором:

* узлы являются:
  + идентификаторами устройств (красный цвет - мобильное устройство, зеленый - не мобильное)
  + логинами (серый цвет - мобильное устройство, коричневый - не мобильное)
* связи определяются, исходя из того, что:
  + идентификаторы устройства связаны между собой
  + логины устройства связаны между собой
  + каждый идентификатор связан с каждым логином одного устройства

Здесь можно было бы определить пересечения между множествами логинов, что говорило об одном владельце этих устройств.

Укладка Force Atlas, видны кластеры



