# Общее описание

Модуль Headers это часть системы антинакрутки MGID. Каждый браузер формирует специфическую структуру заголовков, отличающуюся от других браузеров/версий/платформ. Обучив систему на обучающей выборке, по содержанию и структуре заголовков мы можем предсказать браузер, версию и платформу (ОС). Некоторые версии браузеров а даже отдельные браузеры могут иметь одинаковую структуру заголовков. Поэтому нам нужно сверять предсказанное значение с группой допустимых значений, которые мы также определяем по обучающей выборке

# Требования

1. Система должна быть динамичной. Постоянно выходят новые версии браузеров и в связи с этим систему регулярно переобучаться по новой выборке.
2. Ошибка первого рода (когда мы принимаем нормального пользователя за бота) должна быть не более 5%
3. Ошибка второго рода H1 (когда мы принимаем бота за нормального пользователя) должна быть минимально возможной.

В основном H1 возникает в следующей ситуации: боты используют один из существующих браузерных движков, например Gecko (Firefox). Для снятия подозрений они имитируют естественное распределение браузеров/версий и подставляют соответствующие useragent. Но они имитируют также и firefox, коим они и так являются. Таким образом, если по распределению у нас 10% firefox и мы умеем очень точно определять браузер по заголовкам, то H1 будет у нас более 10%. Ситуация еще хуже с браузером chrome которого 60%. Поэтому нам важно минимизировать группы и предугадывать не только браузер, но и версию и платформу.

# Текущее состояние дел

В качестве фич используются признаки наличия заголовков с определенным содержанием и их последовательность.

Игнорируется содержание следующих заголовков:

**'User-Agent'**,

**'Referer'**,

**'Accept-Language'**,

**'Host'**,

**'X-Host'**,

**'X-Forwarded-For'**,

**'Forwarded-For'** ,

**'X-Network-Type'**,

**'Cookie'**,

**'X-Wap-Profile'**,

**'X-Gateway'**,

**'Ua-Cpu'**,

**'Rvbd-Csh'**,

**'Dris-Orig-Src'**,

**'X-Proxy-Id'**,

**'Rvbd-Ssh'**,

**'X-Requested-With'**

(но их участие в последовательностях учитывается)

Также игнорируются очень редкие и с уникальными именами, которые нужно исключать подсчитывая их частоту.

Наиболее быстрой и точной моделью для прогнозирования заголовков оказалась [DecisionTree](http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier)

Тренировочные выборки находятся в архиве headers.zip в папке с этим документом. Данные в файлах в csv формате со следующими колонками:

1. Время (timestamp)
2. Ip
3. Заголовки в json формате

# Задачи

1. Определить оптимальный объем тренировочной выборки (модель DecisionTree склонна к переобучению)
2. Реализовать скрипт, который будет периодически переобучать рабочую модель на свежих данных.
3. Описать требования к разработке рабочего модуля для антинакрутки на базе заголовков.