Обучаемые сегменты Крипты

Обучающая выборка

5.10.2021

Яндекс

Обучаемые сегменты Крипты. Обучающая выборка. Версия 1

Дата подготовки документа: 5.10.2021

Этот документ является составной частью технической документации Яндекса.

Предупреждение об исключительных правах и конфиденциальной информации

Исключительные права на все результаты интеллектуальной деятельности и приравненные к ним средства индивидуализации юридических лиц, товаров, работ, услуг и предприятий, которым предоставляется правовая охрана (интеллектуальную собственность), используемые при разработке, поддержке и эксплуатации службы Обучаемые сегменты Крипты, включая, но не ограничиваясь, программы для ЭВМ, базы данных, изображения, тексты, другие произведения, а также изобретения, полезные модели, товарные знаки, знаки обслуживания, коммерческие обозначения и фирменные наименования, принадлежат ООО «ЯНДЕКС» либо его лицензиарам.

Использование результатов интеллектуальной деятельности и приравненных к ним средств индивидуализации в целях, не связанных с разработкой, поддержкой и эксплуатацией службы Обучаемые сегменты Крипты, не допускается без получения предварительного согласия правообладателя. Настоящий документ содержит конфиденциальную информацию ООО «ЯНДЕКС». Использование конфиденциальной информации в целях, не связанных с разработкой, поддержкой и эксплуатацией службы Обучаемые сегменты Крипты, а равно как и разглашение таковой, не допускается. При этом под разглашением понимается любое действие или бездействие, в результате которых конфиденциальная информация в любой возможной форме (устной, письменной, иной форме, в том числе с использованием технических средств) становится известной третьим лицам без согласия обладателя такой информации либо вопреки трудовому или гражданско-правовому договору.

Отношения ООО «ЯНДЕКС» с лицами, привлекаемыми для разработки, поддержки и эксплуатации службы Обучаемые сегменты Крипты, регулируются законодательством Российской Федерации и заключаемыми в соответствии с ним трудовыми и/или гражданско-правовыми договорами (соглашениями). Нарушение требований об охране результатов интеллектуальной деятельности и приравненных к ним средств индивидуализации, а равно как и конфиденциальной информации, влечет за собой дисциплинарную, гражданско-правовую, административную или уголовную ответственность в соответствии с законодательством Российской Федерации.

Контактная информация

ООО «ЯНДЕКС»

https://www.yandex.ru

Тел.: +7 495 739 7000

Email: pr@yandex-team.ru

Главный офис: 119021, Россия, г. Москва, ул. Льва Толстого, д. 16

Содержание

Возможности Крипты

Крипта предоставляет типовые сегменты Аудиторий в различных сферах: банкинг, страхование, медицина, индустрия недвижимости и другие. Типовые модели обычно не хуже, а зачастую и лучше специально настроенных (по метрике AUC-ROC).

Термины

Партнер

Организация — заказчик услуг Яндекса, предоставляемых на базе Крипты.

Клиент

Клиент партнера.

Передача данных Яндексу

Партнер предоставляет Яндексу обучающую выборку по своим клиентам, а также следит за обновлением выборки, чтобы модель оставалась качественной.

Передать данные можно одним из способов:

- отправить файл менеджеру;
- загрузить файл в Яндекс.Аудитории и сообщить ID сегмента менеджеру.

Формат файла

Обучающую выборку по клиентам следует предоставлять в виде файла в формате CSV, в кодировке UTF-8.

Первая строка должна содержать наименования полей. Каждая последующая строка представляет собой запись об одном клиенте. Поля записи отделяются друг от друга запятой.

В файле должен присутствовать хотя бы один атрибут клиента, а также данные для обучения: `retro_date` и `target`. При загрузке файла в Яндекс.Аудитории обязательно наличие либо атрибута `phone`, либо `email`.

Описание атрибутов клиента и данных для обучения представлено в таблице ниже.

Атрибуты клиента	
`phone`	Номер телефона клиента в виде 11 цифр, включая код страны. Пример: 79995551111 Атрибут можно указать в формате MD5-хеша. Перед хешированием номер телефона нужно привести к стандартному формату: 11 цифр, включая код страны. Пример номера телефона после хеширования: `f09f2c3d48f31e2a802944ade2e5aec5`
`email`	Адрес электронной почты клиента без пробелов и заглавных букв. Пример: example@yandex.ru Атрибут можно указать в формате MD5-хеша. Перед хешированием адрес нужно привести к стандартному формату: без пробелов и заглавных букв. Пример адреса после хеширования: `7385287bf0079ffaa7ffe95ac293c63d`

Атрибуты клиента	
`ClientID`	Анонимный идентификатор ClientID, присвоенный Яндекс.Метрикой (см. раздел getClientID Справки Яндекс.Метрики). Пример: 12345678901234567890
`IDFA`	Рекламный идентификатор iOS-устройства.
`GAID`	Рекламный идентификатор Android-устройства.
`UUID`	Уникальный идентификатор экземпляра приложения, установленного на устройстве (состоит из 32 шестнадцатеричных цифр). Пример: `A4BBD833A29A9F4C74DBD833A29A9FEC`

Данные для обучения	
`retro_date`	Дата совершения конверсии в формате `YYYY-MM-DD`. Необходима для обучения модели.
`target`	Значение целевой переменной. Допустимые значения: • `0` - пользователь относится к плохой конверсии. • `1` - пользователь относится к хорошей конверсии.

Примеры файлов

```
phone,email,retro_date,target
79209386842,melanesian@example.com,2021-01-23,1
79216921288,tumbrels@example.com,2020-11-21,0
```

```
email,retro_date,target
stalagmites@example.com,2021-08-14,1
unadulterated@example.com,2021-06-10,0
```

```
phone,retro_date,target
2dca71211faa68cd92dbdb54ae53a646,2020-12-22,0
a96039fabc05c3f208c67b296dd3d847,2021-01-11,1
```

Обновление данных

Чтобы модель оставалась качественной, обучающую выборку следует регулярно обновлять новыми данными.

В файле обновления необходимы:

- Идентификатор (`phone`, `email`, `ClientID`, `IDFA`, `GAID` или `UUID`).

  Всего в файле обновления должно быть не меньше 50 000 записей с идентификаторами.
- Поле `target` для разбиения клиентов на «хороших» и «плохих». Например, хороший — вернул или возвращает кредит, плохой — не вернул кредит. Представительство минимального класса должно быть не менее 1000 идентификаторов.
- Поле `retro_date` с датой обновления информации по идентификатору.

Например, у банка всего было 100 000 нужных конверсий за 1 год. Среди них 40 000 плохих конверсий и 60 000 хороших конверсий. В данном случае минимальным классом являются плохие конверсии, поэтому в файле обновления таких записей должно быть не меньше 1000. Но так как в файле должно быть как минимум 50 000 записей, чтобы сохранить распределение в файл обновления нужно включить как минимум 20 000 записей о плохих конверсиях и 30 000 записей о хороших.

Вопросы и ответы

Какие атрибуты клиента передавать?

Рекомендуется передавать все имеющиеся атрибуты клиента:

- MD5-хеш номера телефона клиента;
- MD5-хеш адреса электронной почты клиента;
- анонимный идентификатор ClientID, присвоенный Яндекс.Метрикой;
- рекламные идентификаторы IDFA или GAID;
- идентификатор экземпляра приложения UUID.

Эти атрибуты помогают найти соответствие между клиентом партнера и пользователем Яндекса.

Как получить MD5-хеш?

Во многих языках программирования существуют инструменты для получения MD5-хеша, например: пакет hashlib в Python, класс MessageDigest в Java и др.

Перед хешированием значения нужно привести к стандартному формату:

- номер телефона — 11 цифр с кодом страны;
- адрес электронной почты — без пробелов и заглавных букв.

Если значений несколько, каждое нужно хешировать отдельно.

Проверьте правильность работы вашего алгоритма получения MD5-хеша с помощью контрольных пар:

phone	**phone_md5**
79160444381	b356af14c869ed7e31e185cbf075b89b
79137172777	50710784ab8398162e5772f8996eea5a

email	**email_md5**
example@yandex.ru	7385287bf0079ffaa7ffe95ac293c63d
mail@example.org	5b9c2b225b5c4ff91ffe849209153ecc