Аналитика пользовательского поведения на основе сырых логов

Шитова Анастасия, DA-11

Цель исследования

Информация из логов запросов к серверу, как правило, доступна сравнительно легко. Задача исследования: извлечь максимально полное представление о пользователях на основе их запросов к серверу:

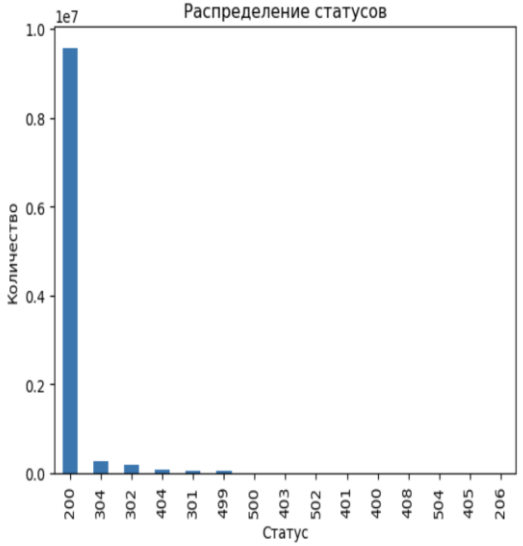
- Геолокация
- Периоды активности
- Характеристики активности
- Характеристики пользователей
- Нагрузка по запросам

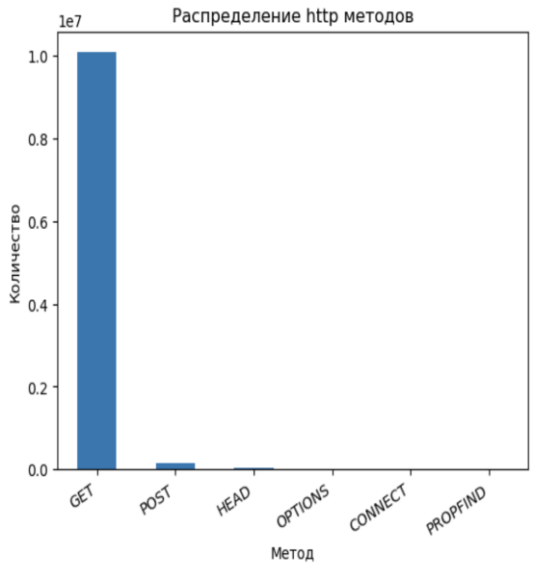
Данные

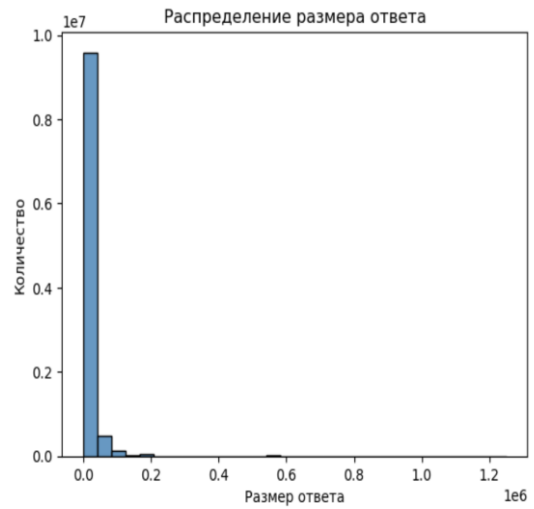
- Данные серверных логов иранского маркетплейса zanbil.ir за январь 2019 год (файл формата *.log)

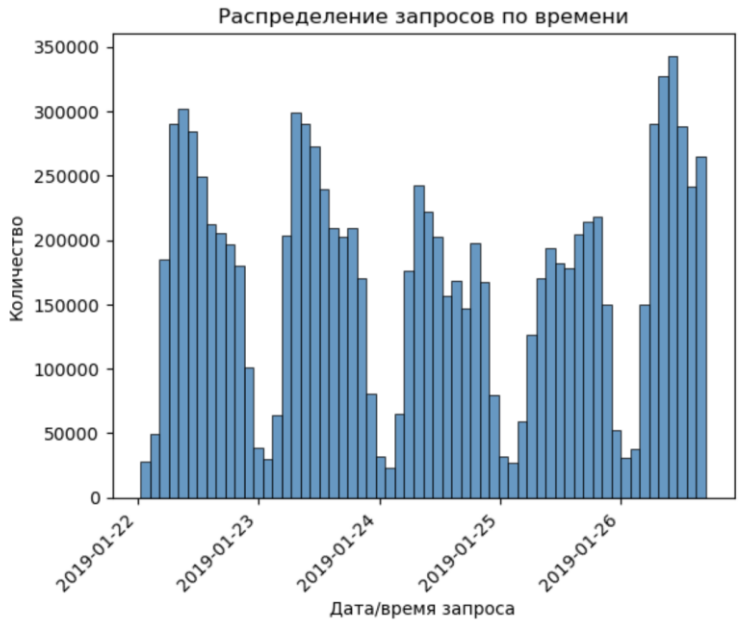
	client	datetime	method	request	status	size	referer	user_agent	traffic_label	ref
0	37.152.163.59	2019-01-22 12:38:27+03:30	GET	/image/29314?name=%D8%AF%DB%8C%D8%A8%D8%A7-7.jpg&wh=50x50	200	1105	https://www.zanbil.ir/product/29314/%da%a9%d8%a7%d9%84%d8%b3%da%a9%d9%87-%d8%af%d9%88%d9%82%d9%84%d9%88-%d8%af%d9%84%db%8c%d8%ac%d8%a7%d9%86-%d9%85%d8%af%d9%84-%d8%af%db%8c%d8%a8%d8%a7-delijan-twin-strollers-	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko	other	www.zanbil.ir
1	37.152.163.59	2019-01-22 12:38:27+03:30	GET	/static/images/zanbil-kharid.png	200	358	https://www.zanbil.ir/product/29314/%da%a9%d8%a7%d9%84%d8%b3%da%a9%d9%87-%d8%af%d9%88%d9%82%d9%84%d9%88-%d8%af%d9%84%db%8c%d8%ac%d8%a7%d9%86-%d9%85%d8%af%d9%84-%d8%af%db%8c%d8%a8%d8%a7-delijan-twin-strollers-	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko	other	www.zanbil.ir
2	85.9.73.119	2019-01-22 12:38:27+03:30	GET	/static/images/next.png	200	3045	https://znbl.ir/static/bundle-bundle_site_head.css	Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36	other	znbl.ir
3	37.152.163.59	2019-01-22 12:38:27+03:30	GET	/image/29314?name=%D8%AF%DB%8C%D8%A8%D8%A7-4.jpg&wh=50x50	200	1457	https://www.zanbil.ir/product/29314/%da%a9%d8%a7%d9%84%d8%b3%da%a9%d9%87-%d8%af%d9%88%d9%82%d9%84%d9%88-%d8%af%d9%84%db%8c%d8%ac%d8%a7%d9%86-%d9%85%d8%af%d9%84-%d8%af%db%8c%d8%a8%d8%a7-delijan-twin-strollers-	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko	other	www.zanbil.ir
4	85.9.73.119	2019-01-22 12:38:27+03:30	GET	/static/images/checked.png	200	1083	https://znbl.ir/static/bundle-bundle_site_head.css	Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36	other	znbl.ir

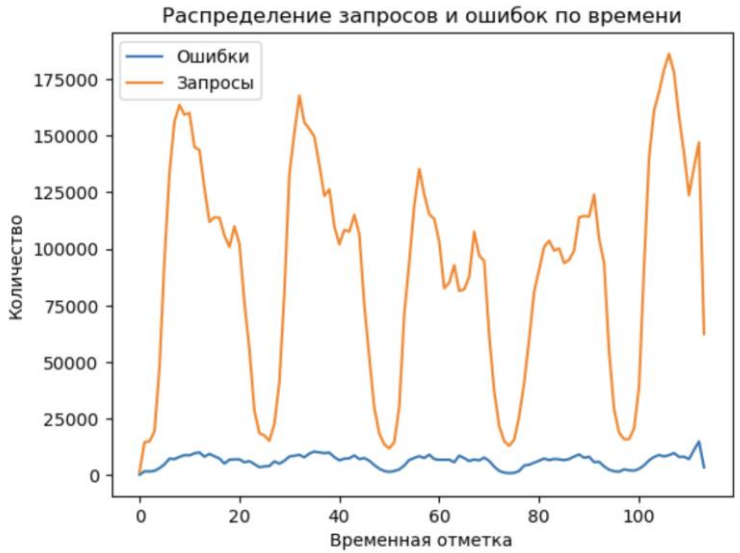
Распределения данных











Фильтрация данных

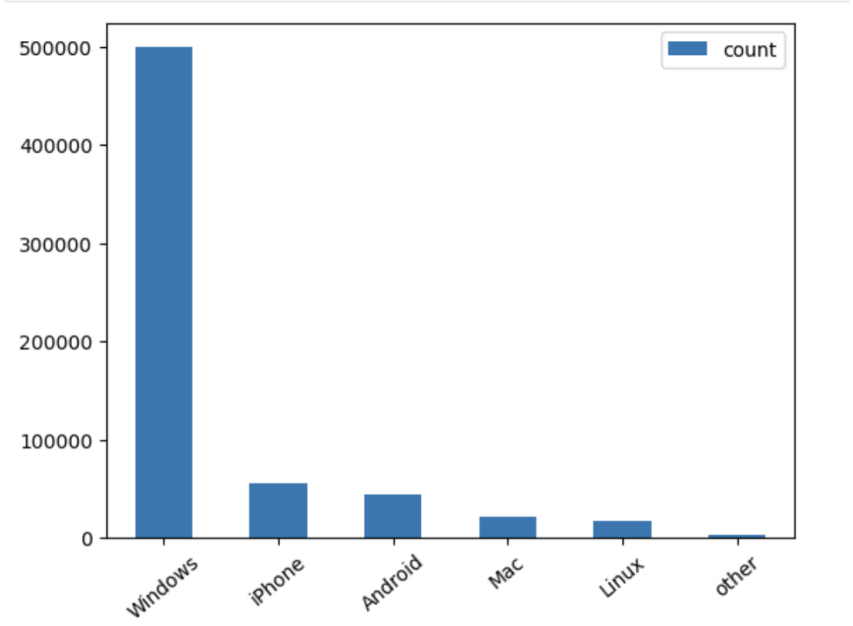
- Дедупликация данных
  данные сократились на 1%
  с 10 364 865 до 10 256 742 записей
- Удаление запросов от ботов
  данные сократились на 40%
  с 10 256 742 до 6 435 015 записей
- Удаление запросов медиафайлов
  данные сократились на 90%
  с 6 435 015 до 747 038 записей
- Удаление ответов с ошибками
  данные сократились на 15%
  с 747 038 до **640 155** записей

Топ посетителей

client	user_agent	count
91.99.72.15	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.7(KHTML, like Gecko) Chrome/16.0.912.36 Safari/535.7	9701
91.99.72.15	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.92 Safari/537.36	9627
91.99.72.15	Mozilla/5.0 (Windows NT 6.2; Win64; x64; rv:16.0)Gecko/16.0 Firefox/16.0	9604
91.99.72.15	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_3) AppleWebKit/534.55.3 (KHTML, like Gecko) Version/5.1.3 Safari/534.53.10	9556
130.185.76.185	Mozilla/5.0 (Windows NT 6.1; rv:65.0) Gecko/20100101 Firefox/65.0	8257
34.247.132.53	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/63.0.3239.84 Safari/537.36	6686
91.99.30.32	Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:64.0) Gecko/20100101 Firefox/64.0	6462
91.99.47.57	Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:64.0) Gecko/20100101 Firefox/64.0	5204
130.185.74.243	Mozilla/5.0 (Windows NT 6.1; rv:42.0) Gecko/20100101 Firefox/42.0	5194
5.78.190.233	Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:64.0) Gecko/20100101 Firefox/64.0	4869

Топ устройств

os	count
Windows	499730
iPhone	54972
Android	43955
Mac	21427
Linux	16599
other	3472



Топ сайтов, с которых пришли пользователи

traffic_label	ref	
direct		100394
other	www.zanbil.ir	461428
	www-zanbil-ir.cdn.ampproject.org	9577
	ptcnovin.com	2751
	emalls.ir	2647
	www.ptcnovin.com	490
search	www.google.com	42167
	www.zanbil.ir	14884
	api.torob.com	1414
	www.bing.com	318
	search.mysearch.com	249
social	org.telegram.messenger	16
	www.facebook.com	4
	l.instagram.com	3
	instagram.com	1
	l.facebook.com	1

Name: count, dtype: int64

Сессии пользователей

client	user_agent	start_time	end_time	pages	session_time
185.11.88.198	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36	2019-01-22 13:42:31+03:30	2019-01-26 13:59:26+03:30	[/filter/b72,p8095, /site/alexaGooleAnalitic, /, /, /site/enamad, /site/alexaGooleAnalitic, /filter/p65%2Cb1%2Cstexists, /site/enamad, /site/alexaGooleAnalitic, /ajaxFilter/p65,b1,stexists,6313%7C49%20%D8%A7%DB%8C%D9%86%DA%86?o=6313, /ajaxFilter/p65,b1,stexists,6313%7C49%20%D8%A7%DB%8C%D9%86%DA%86,6313%7C50%20%D8%A7%DB%8C%D9%86%DA%86?o=6313, /product/33598/64267/%D8%AA%D9%84%D9%88%DB%8C%D8%B2%DB%8C%D9%88%D9%86-%D8%A7%D9%84-%D8%A7%DB%8C-%D8%AF%DB%8C-%D8%B3%D8%A7%D9%85%D8%B3%D9%88%D9%86%DA%AF-%D9%85%D8%AF%D9%84-50NU7900, /site/alexaGooleAnalitic, /basket/alert/64267, /basket/add/64267?addedValues=, /browse/fan/%D9%BE%D9%86%DA%A9%D9%87, /site/alexaGooleAnalitic, /ajaxFilter/p42,b41, /product/2352/1389/%D9%BE%D9%86%DA%A9%D9%87-%D9%BE%D8%A7%DB%8C%D9%87-%D8%A8%D9%84%D9%86%D8%AF-%D9%BE%D8%A7%D8%B1%D8%B3-%D8%AE%D8%B2%D8%B1-%D9%85%D8%AF%D9%84-ES4010RWKM, /site/alexaGooleAnalitic, /basket/alert/1389, /basket/add/1389?addedValues=, /basket/checkout, /province/getProvinceCities?id=1, /site/alexaGooleAnalitic, /basket/checkout?currentStep=2, /province/getProvinceCities?id=1, /site/alexaGooleAnalitic, /province/getProvinceCities?id=8, /basket/checkout?currentStep=3, /province/getProvinceCities?id=1, /site/alexaGooleAnalitic, /order/create, /site/alexaGooleAnalitic]	4 days 00:16:55
185.11.88.198	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36	2019-01-26 14:13:31+03:30	2019-01-26 14:23:47+03:30	[/province/getProvinceCities?id=1, /site/alexaGooleAnalitic, /province/getProvinceCities?id=8, /site/alexaGooleAnalitic, /filter/p42,b41, /site/alexaGooleAnalitic, /site/alexaGooleAnalitic, /order/track, /site/alexaGooleAnalitic, /order/track, /site/alexaGooleAnalitic, /order/track, /site/alexaGooleAnalitic]	0 days 00:10:16

Выводы

В данной работе проведён анализ информации из логов сервера, показано, как на основе информации о запросах можно узнать больше о пользователях соответствующего сайта:

- откуда пришло больше всего посетителей,
- какие устройства пользуются наибольшей популярностью у пользователей,
- в какие часы стоит ждать пик посещений,
- что именно делают пользователи на сайте, какие страницы посещают,
- сколько времени пользователи проводят на сайте,
- как много ошибок ответов и как часто они происходят,
- сколько "реальных пользователей" и сколько ботов отправляют запросы на сайт

Развитие исследования

- Провести исследование данных с заполненными user_id пользователей, тогда сессии можно было бы привязать к отдельным пользователям. Если была бы доступна информация о заказах пользователей с данными user_id, можно было бы отследить пути пользователей/переходы по страницам сайта, которые наиболее вероятно приведут к покупке.
- Изучить детальнее, какие именно страницы входят в одну сессию, есть ли какие-то закономерности в переходах между страницами. Возможно, это могло бы улучшить систему рекомендаций: предлагать те товары, которые чаще всего попадают в одну сессию с текущим товаром.
- Изучить географию пользователей: ориентироваться не на группы ip адресов, а на конкретные ip-адреса. Добавить интеграцию с сервисом, предоставляющим информацию о локации по данным об ip. Локации можно нанести на карту и посмотреть географию пользователей в наглядном виде. Также можно сопоставить, в какой локации какие страницы пользуются наибольшей популярностью и, например, таргетировать рекламу по этим данным.