JONAS_yammer_report_full

SQL Practice - Search Engine

On this notebook we will analyze the performance of the search engine of yammer.com to assess if it is reasonable to deploy resources in improving it. We will use Mode's databases following the description on https://community.modeanalytics.com/sql/tutorial/understanding-search-functionality/

First, let's examine the data. The events log looks like this:

1_examine_data

	user_id	occurred_at	event_type	event_name	location	device	user_type
1	10522	2014-05-02 11:02:39	engagement	login	Japan	dell inspiron notebook	3
2	10522	2014-05-02 11:02:53	engagement	home_page	Japan	dell inspiron notebook	3
3	10522	2014-05-02 11:03:28	engagement	like_message	Japan	dell inspiron notebook	3
4	10522	2014-05-02 11:04:09	engagement	view_inbox	Japan	dell inspiron notebook	3
5	10522	2014-05-02 11:03:16	engagement	search_run	Japan	dell inspiron notebook	3
6	10522	2014-05-02 11:03:43	engagement	search_run	Japan	dell inspiron notebook	3
7	10612	2014-05-01 09:59:46	engagement	login	Netherlands	iphone 5	1
8	10612	2014-05-01 10:00:18	engagement	like_message	Netherlands	iphone 5	1
9	10612	2014-05-01 10:00:53	engagement	send_message	Netherlands	iphone 5	1
10	10612	2014-05-01 10:01:24	engagement	home_page	Netherlands	iphone 5	1

We note that there is not a real session log, that will be useful for the future. By rearranging the data in terms of user id and event time we can define a session as the set of events occurring between pauses of 20 minutes or more. In that case, we can identify the following

2_sessions

	user_id	event_name	occurred_at	lag_time	session
1	4	login	2014-05-13 09:31:47		1
2	4	home_page	2014-05-13 09:32:10	0 years 0 mons 0 days 0 hours 0 mins 23.00 secs	1
3	4	search_autocomplete	2014-05-13 09:32:26	0 years 0 mons 0 days 0 hours 0 mins 16.00 secs	1
4	4	search_autocomplete	2014-05-13 09:32:58	0 years 0 mons 0 days 0 hours 0 mins 32.00 secs	1
5	4	login	2014-05-20 09:31:30	0 years 0 mons 6 days 23 hours 58 mins 32.00 secs	2
6	4	search_autocomplete	2014-05-20 09:31:55	0 years 0 mons 0 days 0 hours 0 mins 25.00 secs	2
7	4	home_page	2014-05-20 09:31:56	0 years 0 mons 0 days 0 hours 0 mins 1.00 secs	2
8	4	search_run	2014-05-20 09:31:59	0 years 0 mons 0 days 0 hours 0 mins 3.00 secs	2
9	4	search_autocomplete	2014-05-20 09:32:31	0 years 0 mons 0 days 0 hours 0 mins 32.00 secs	2
10	4	search_autocomplete	2014-05-20 09:33:01	0 years 0 mons 0 days 0 hours 0 mins 30.00 secs	2
11	4	login	2014-05-24 11:39:53	0 years 0 mons 4 days 2 hours 6 mins 52.00 secs	3
12	4	home_page	2014-05-24 11:40:20	0 years 0 mons 0 days 0 hours 0 mins 27.00 secs	3
13	4	login	2014-05-27 15:09:09	0 years 0 mons 3 days 3 hours 28 mins 49.00 secs	4
14	4	like_message	2014-05-27 15:09:36	0 years 0 mons 0 days 0 hours 0 mins 27.00 secs	4
15	4	search_run	2014-05-27 15:09:40	0 years 0 mons 0 days 0 hours 0 mins 4.00 secs	4

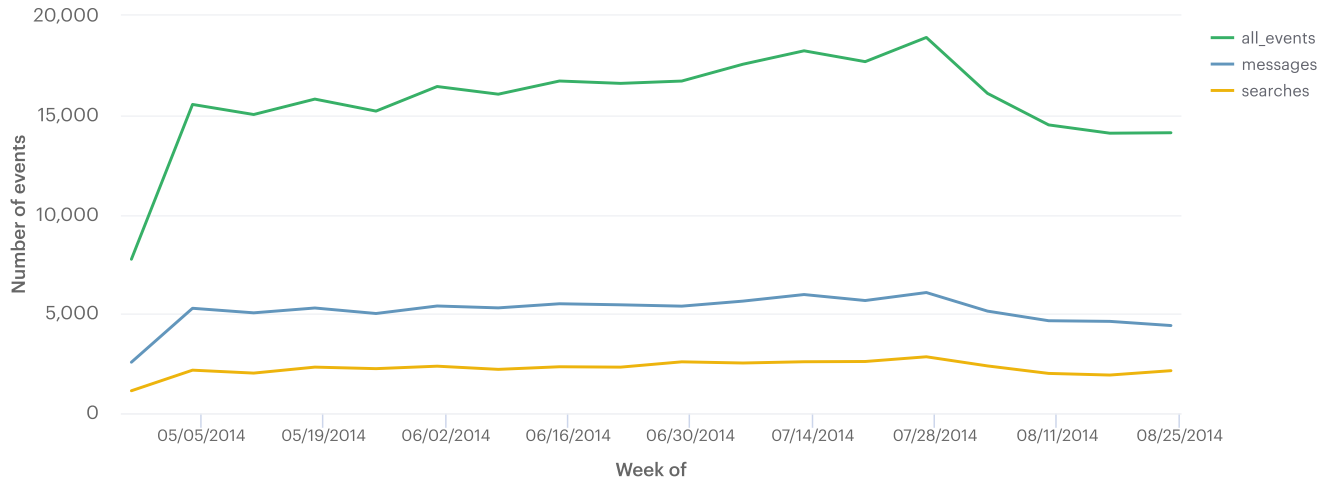
This seems to define the sessions successfully, especially if we look at more data. In particular, there's a large enough number of sessions:

3_total_num_of_sessions

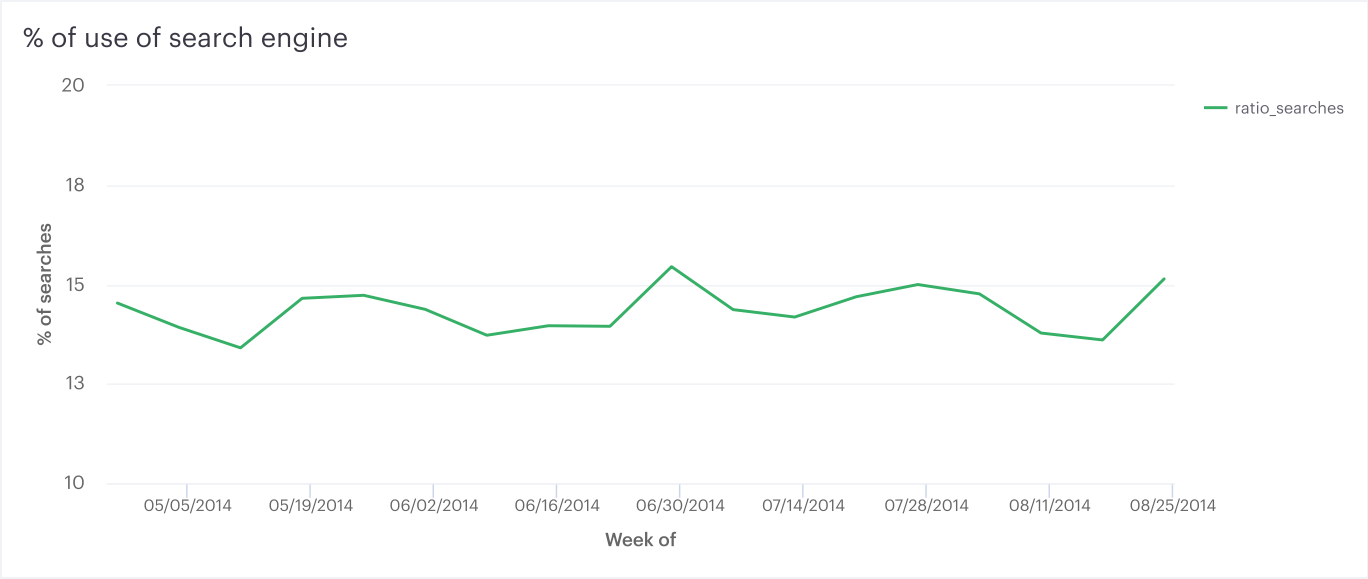
	max
1	38401

Now that we are familiar with the data, let's look at the importance of the search engine. First, we will look at how much is used proportionally, and we will break it into autocompleted searches, full run searches, and clicks on the results provided

Engagement events per week



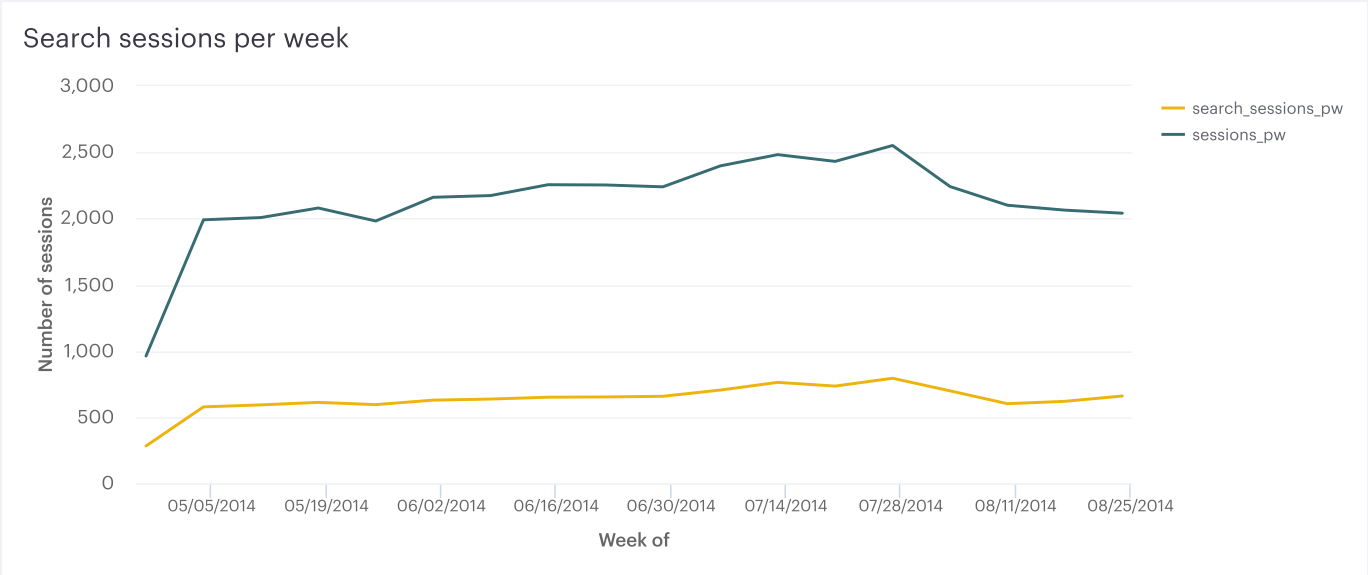
We see that the number of searches is about half of the interaction with messages, but in a percentage they share about 15% of the usage of the platform. It is therefore not a residual tool.

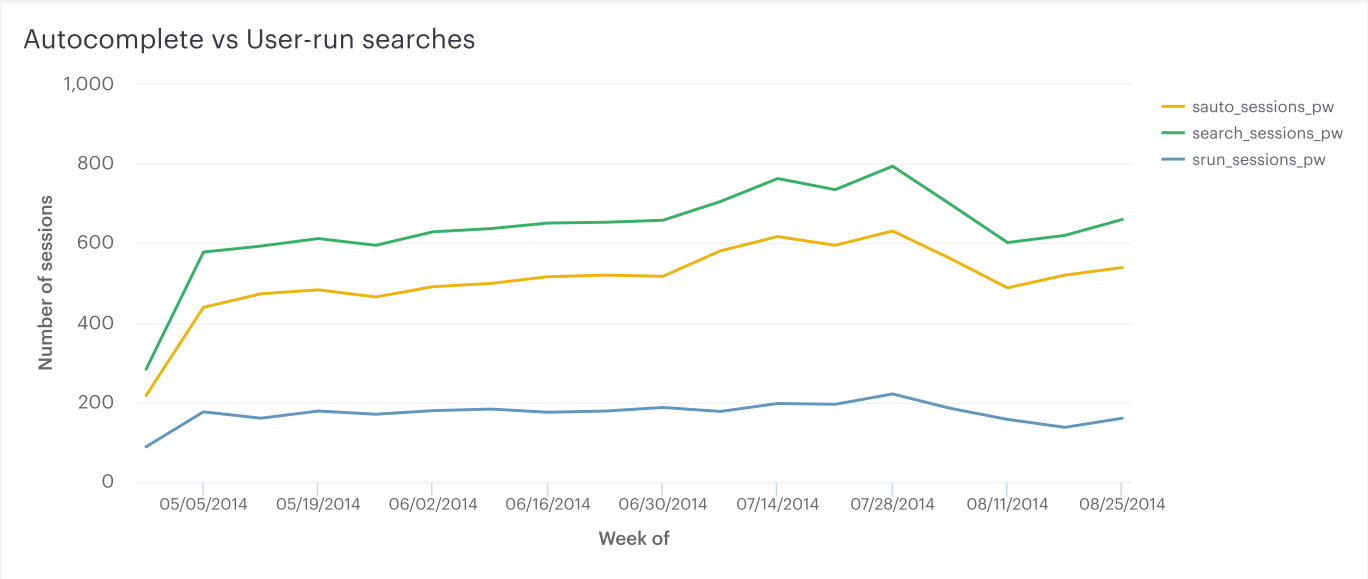


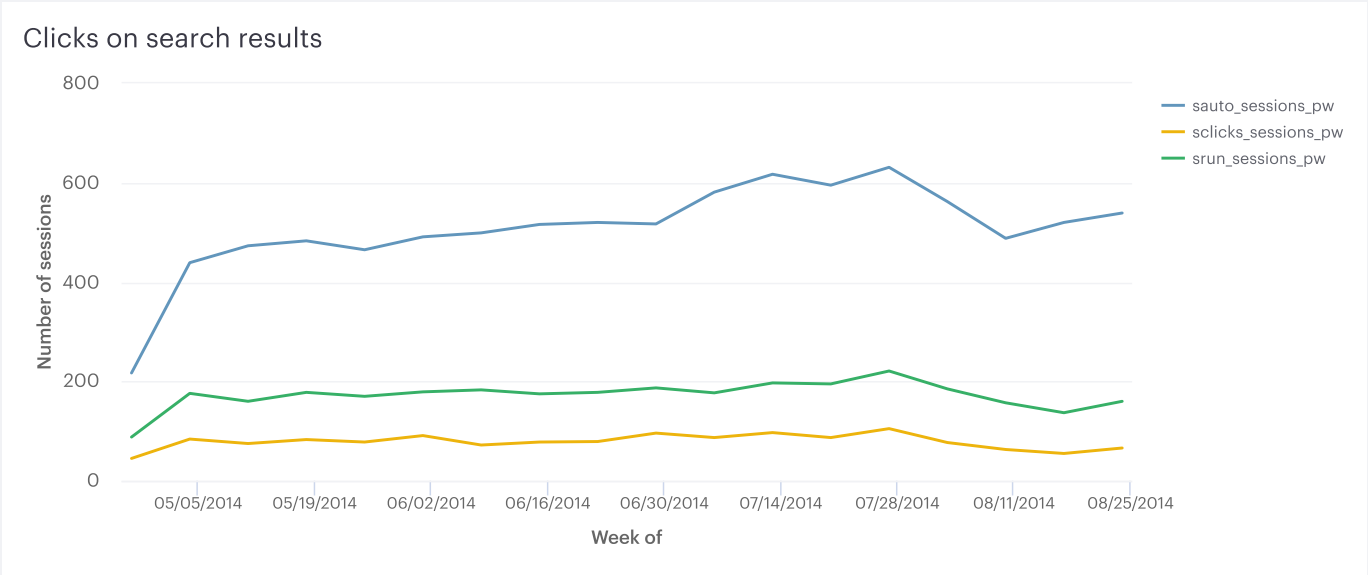
Let's look at the data per session now.

Most of the searches are from the autocomplete engine, which is a good sign in the sense that it works very well, or, at least, it has a lot of users. We would not recommend working on the autocomplete unless there's specific complains about its performance.

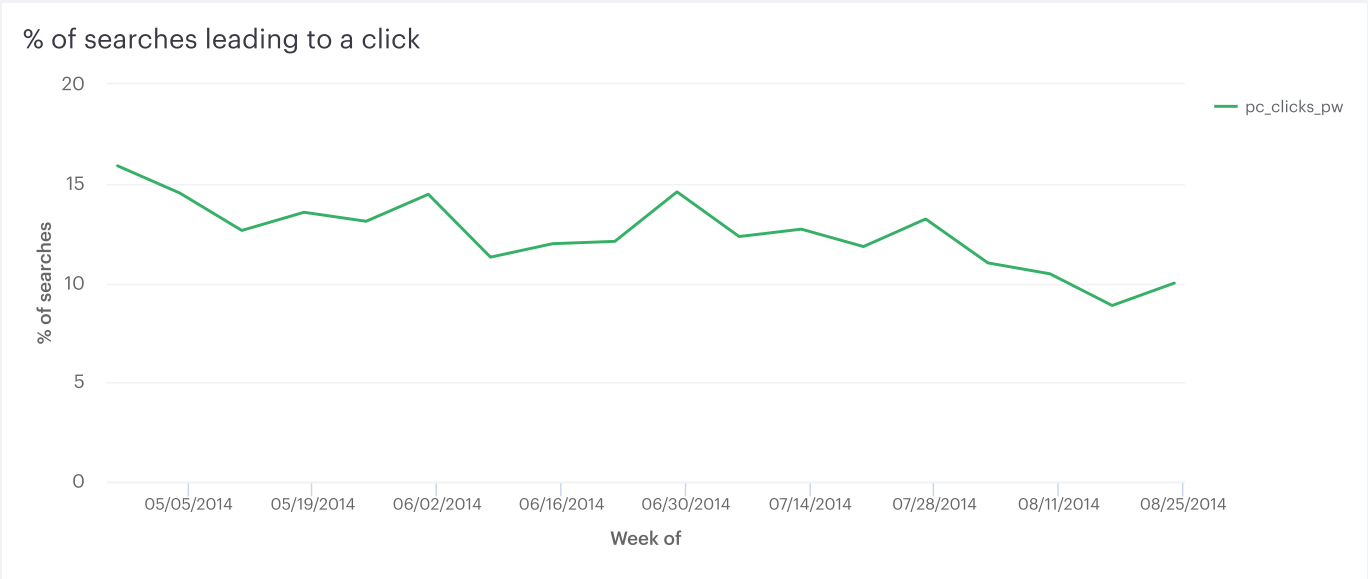
We also observe very few clicks on the results of searches. That is typically a bad sign, unless the user is looking for some kind of information provided already in the search query, like a name or a date. More information is required to address that question.



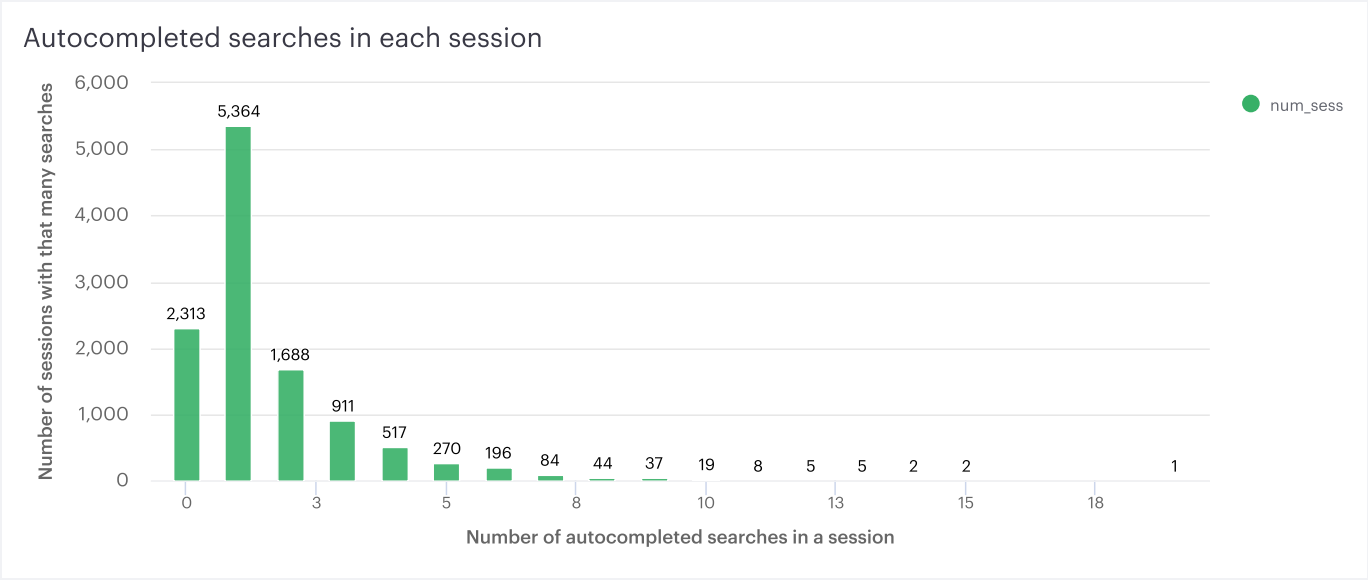


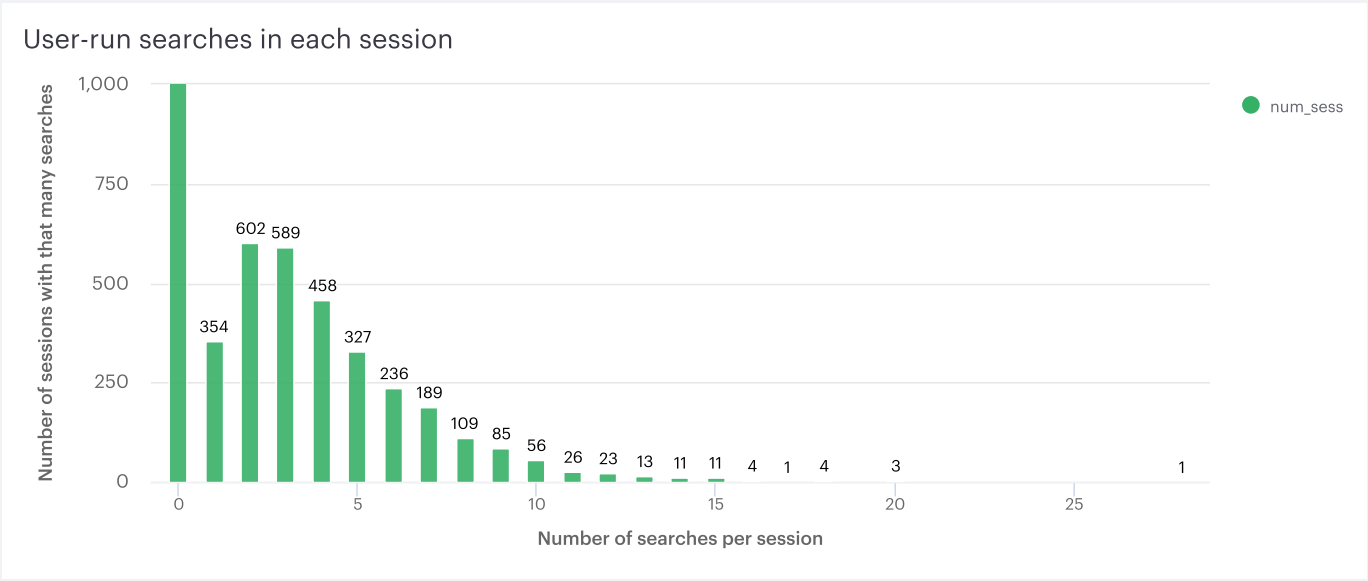


Indeed, if we look at the searches leading to a click, it's about 10%, and it keeps going down per week! Definitely worth spending some time here.

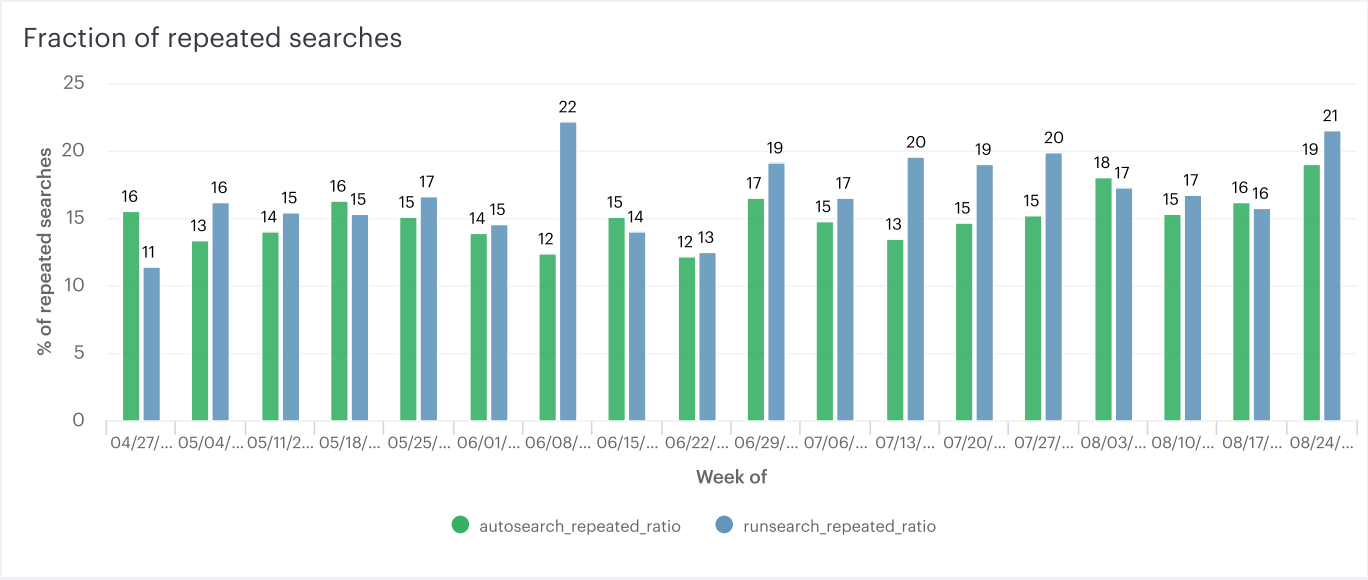


Let's look at how much use is given to the tool per session. Intuitively, running many searches in each session should indicate poor performance, except for users who search very different items very frequently. From the bar charts we observe that many sessions running autocomplete do it only once, while those who run hardcore entry searches do it many times, typically 2 or 3 times per session. That indicates us that there may be room for improvement.

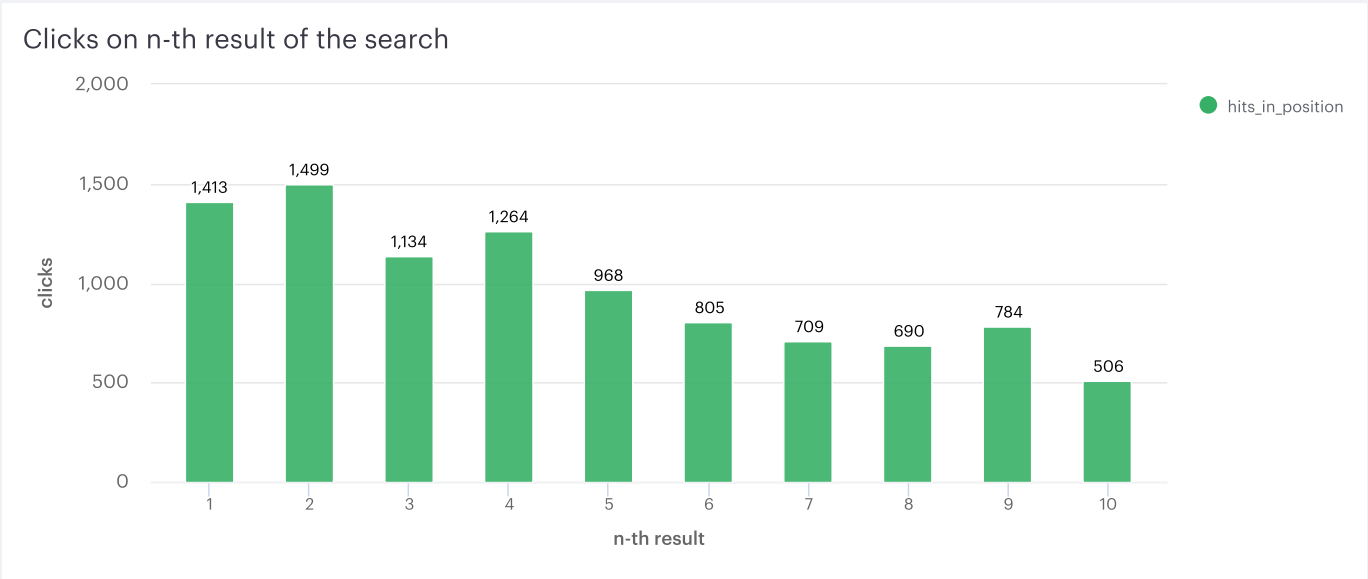




While this seems to point in the direction of a poor performance for the user-run searches, let's look at the repetition ratio. If a user imposes a search, and doesn't like the output, he may try to search something different. We observe many back-to-back searches, slightly more on the user-run events, but substantially larger than what we may expect. It is also noticeable a mild but steady growth for the user-run searches, a possible source of concern for the future.



Finally, the last piece of evidence we want to show is how often the search results display the preferred output as the first(s) results. We observe that is not the case!



Therefore, it is evident that there's room for improvement in the search feature. The recommendation is to leave the autocomplete as is, and work on the results displayed. Minimizing the number of repeated searches and maximizing the clicks after a search should be a priority, and the recommendation is to better assess the relevance of the results: that would pull higher the desired results, reduce duplicate searches, and hopefully lead the user to the desired destination in more cases.