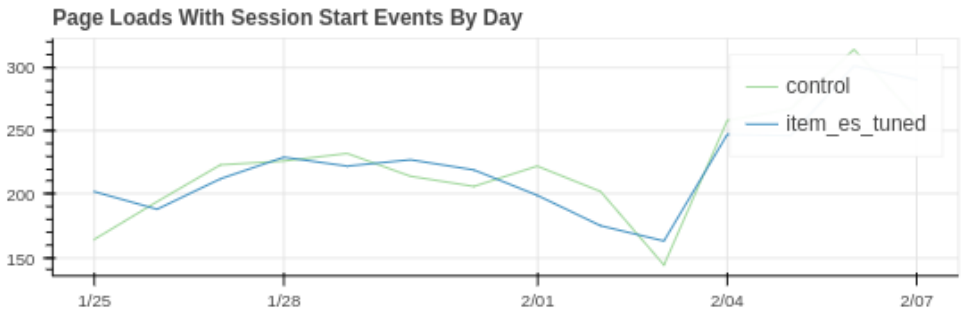
spanish wbsearchentities a/b test analysis

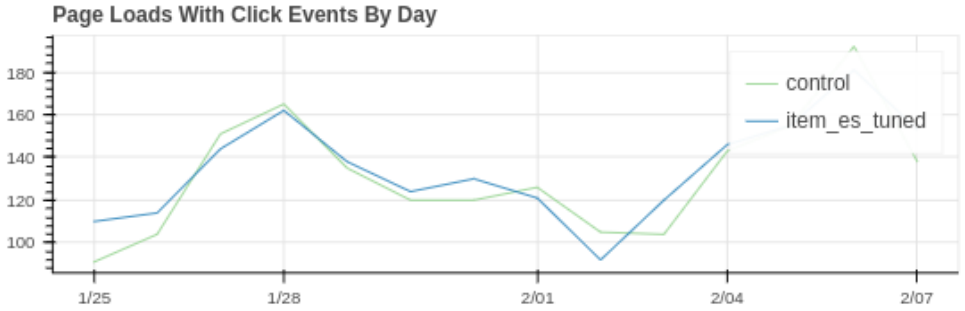
As part of T208917 (https://phabricator.wikimedia.org/T208917) the weights of the wbsearchentities prefix search on www.wikidata.org were tuned using historical click logs as a guide. To determine the effectiveness of this tuning an AB test was run from Jan 25 at 00:00 UTC until Feb 8 at 00:00 UTC. The test was limited to users performing searches for wikidata items in the german language. Users were divided equally on a per-page load basis into either the control or test bucket.

The analytics for this test collect usage from the various entity selectors used throughout the interface for editing Wikidata. The data also includes usage of the autocomplete on the top-right of all Wikidata pages, but due to a bug in the data collection the usage of the top-right autocomplete was only logged from entity pages.

The graphs below are all probability densities and annotated with 95% confidence intervals. The confidence intervals are constructed by running five thousand rounds of the bootstrap method.

Event Counts



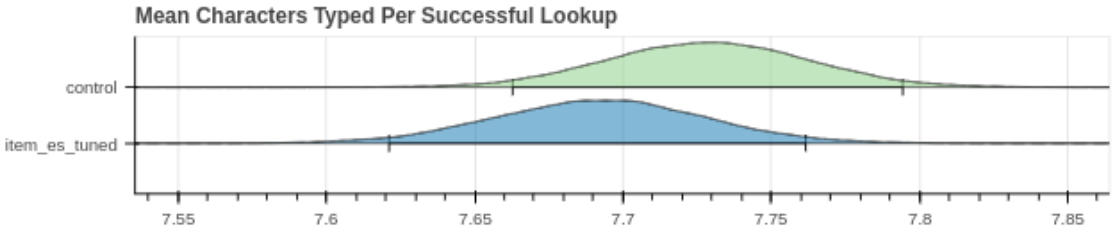


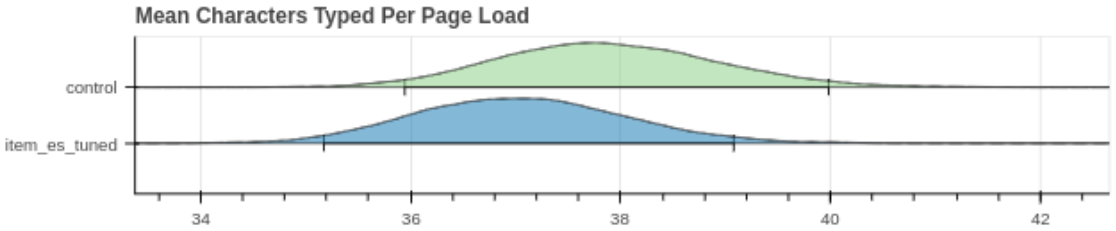
Raw event counts per bucket

bucket	event	count
control	click	1849
control	start	3125
item_es_tuned	click	1890
item_es_tuned	start	3120

Number of characters typed before success

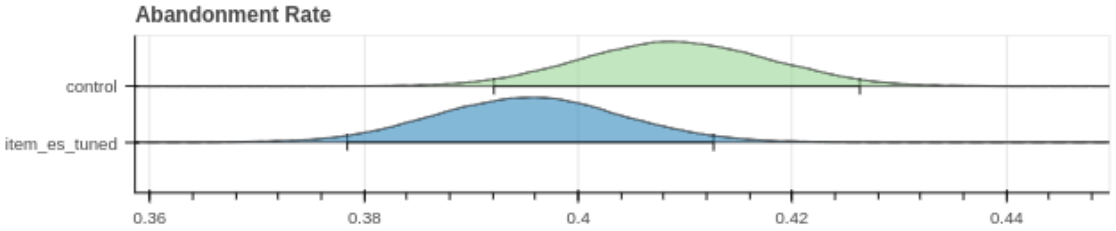
The number of characters typed in each session is a proxy for the amount of effort a user must exert to find the item they are looking for. The 95% CI completely overlaps, suggesting the test treatment had no effect on the number of characters typed





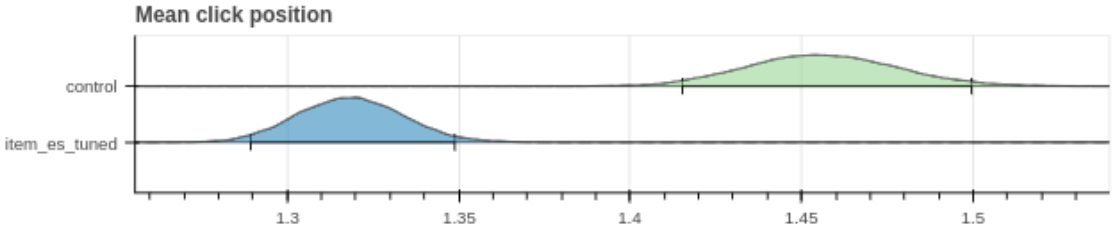
Abandonment Rate

The ratio of page loads with start events against the page loads with click events is interepreted loosly as the abandonment rate of search. The 95% CI completely overlaps, suggesting the test treatment had no effect on abandonment rates. This is the first time we've looked at abandonment rates for wbsearchentities, and further investigation into why it is so high may be called for.



Click Position

The position of clicked result is another proxy for the amount of effort a user must exert to find the item they are looking for. The mean position clicked was decreased from 1.45 to 1.33. Additionally the click position distribution is more concentrated with the new tuning, suggesting users are experiencing more consistent behaviour.



Looking into this result closer while the overall click position doesn't change significiantly, there is an observable change in Clicks@1 from 78% to 85%. Clicks@2 saw a comparable drop from 13% to 9%.