Language detection via Accept-Language

Mikhail Popov (Analysis and Reporting), Erik Berhardson (Engineering and Data Collection), Dan Garry (Management), Oliver Keyes (Review)

January 4, 2016

Summary

This was an A/B test to determine the impact of switching search languages in the case that a query produces zero results or fewer than 3 results. While we previously used the elasticsearch plug-in "langdetect" to detect the query's language, we postulated that we should first try finding a valid language using the Accept-Language header, if it is detected.

We found evidence that the Accept-Language header detection makes a slight positive difference to the zero results rate, with 3.18-3.34% more requests getting some results in the test group than the control group, and the test group being 1.026-1.029 more likely to get some results than the control group when we found a valid language via Accept-Language detection.

As with the previous test, we recommend collecting data about users' interaction with the results to see whether the language detection (via AL and/or es-plugin) produces useful results, not just some results.

Introduction

Using the first non-English Accept-Language HTTP header will provide a good proxy for the language the query is in when the query returns no results against the wiki it was already run against. We postulated that this is a better proxy than the existing elasticsearch "langdetect" plugin we tested previously (Keyes et al., 2015).

There were three total groups of requests: the control group with status quo settings, the requests for which we performed a language detection based on Accept-Language (AL) header and then elasticsearch plug-in (es-plugin) when there were 0 results returned, and the requests for which we performed a language detection when there were less than 3 results returned. If the AL detection could not find a valid language, we fell back on the es-plugin detection.

Methods

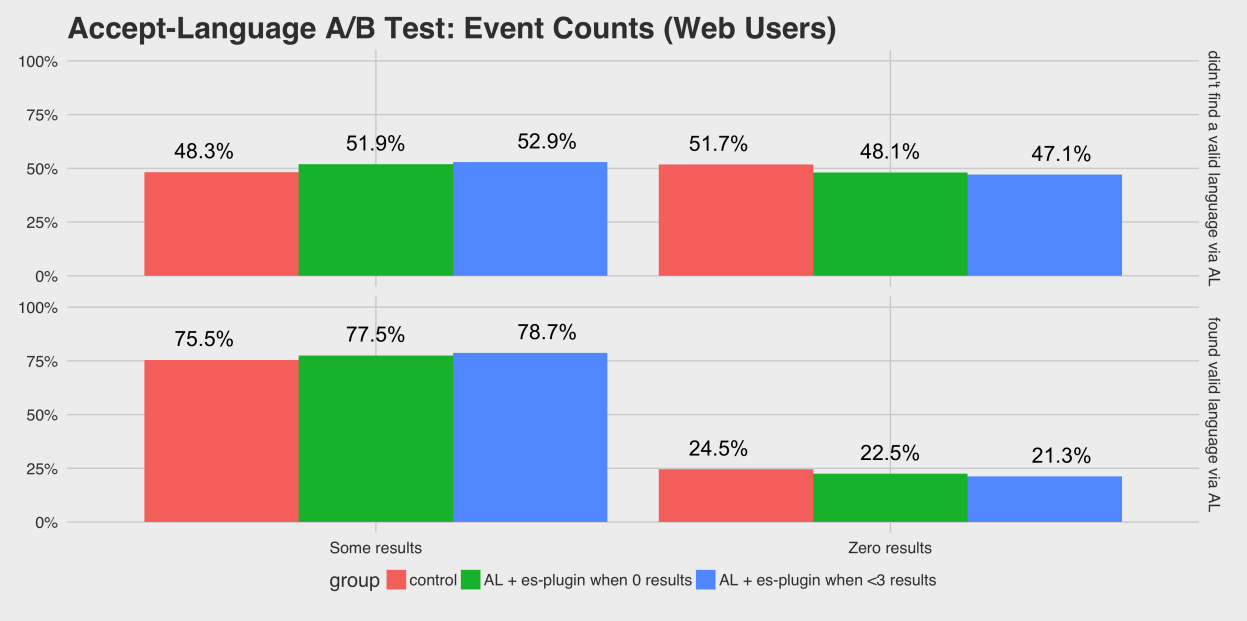
1 out fo 7 requests were sampled. Since the sampling process is pseudo-random yet deterministic, if a user was selected for this test on their first request, they would be selected for the test on subsequent requests. Thus, each group had a sampling rate of 1 in 21 sample requests.

First, we counted the number of events per user (uniquely identified by their IP, User Agent, and X-Forwarded-For). After looking at the counts of events, we restricted our **Web** dataset to users who made less than 7 search requests, which accounts for 99% of the data. The top 1% of **Web** users was excluded from analysis because they would skew the results.

While API requests are included in this test, the API user must pass an explicit flag opting into the general query rewriting feature which basically no-one enables. Any measurable effect will be entirely within users of the **Web**-based search.

To perform the analysis, we employed Agresti and Min's (2005) methods for computing Bayesian confidence intervals.

Results



We compared controls with the group where AL & es-plugin were used when 0 results were returned, and computed the following 95%-probability intervals:

Among those for who we *were* ***not*** *able to detect a valid language via AL*, 2.8%-4.4% more users received results in the test group and were 1.058-1.093 times more likely to receive results than the control group.

Among those for who we ***were*** *able to detect a valid language via AL*, 2%-2.2% more users received results in the test group and were 1.026-1.029 times more likely to receive results than the control group.

Less than 3 results We also peformed this test with a test group who had received less than 3 results.

Among those for who we *were* ***not*** *able to detect a valid language via AL*, 3.87%-5.46% more users received results in the test group and were 1.08-1.11 times more likely to receive results than the control group.

Among those for who we ***were*** *able to detect a valid language via AL*, 3.18-3.34% more users received results in the test group and were 0.95-1.044 times more likely to receive results than the control group, which is to say we cannot conclude that their search results were impacted in a significantly positive way.

References

- Keyes, O., Bernhardson, E., Causse, D., Garry, D., Popov, M. (2015). Results of Second Language Switching A/B Test. *Wikimedia Commons*. (PDF)
- User:EBernhardson (WMF)/Notes/Accept-Language
- Discovery/Testing: Language detection via Accept-Language
- Popov, M. (2015). Assessment of bucketing used for backend tests which report to CirrusSearchUserTesting log. *Wikimedia Commons*. (PDF)
- Agresti, A. and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2x2 contingency tables, *Biometrics*, **61**, 515-523. doi:10.1111/j.1541-0420.2005.031228.x