Tanner Prynn

ISTA 116

Professor Surdeanu

TA: Nathan Dykhuis

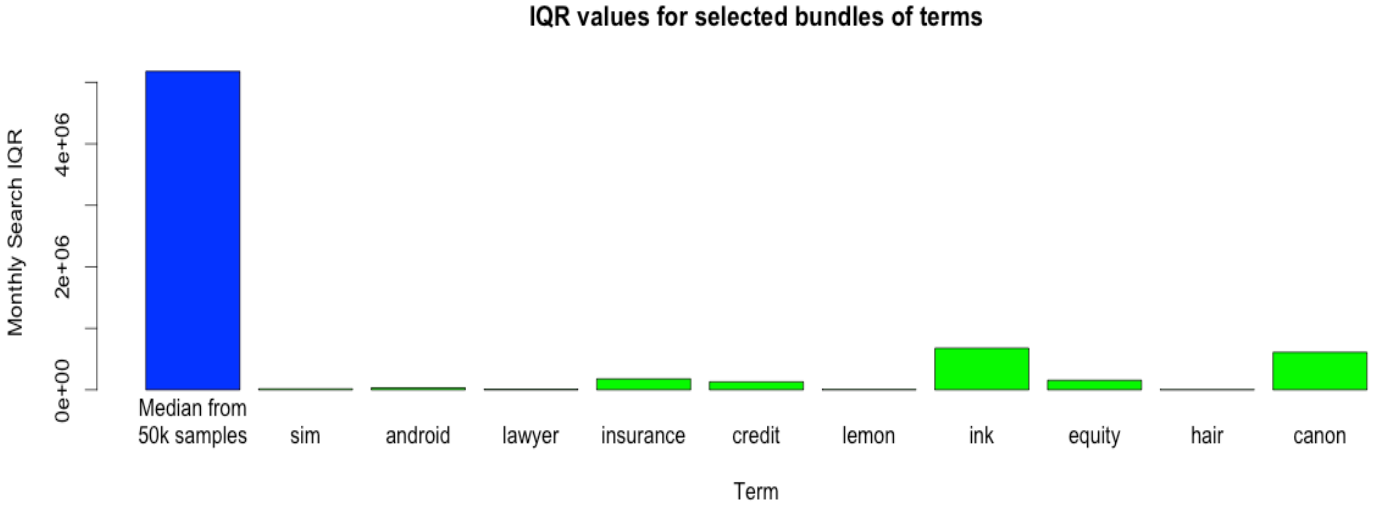
1 May 2013

Introduction

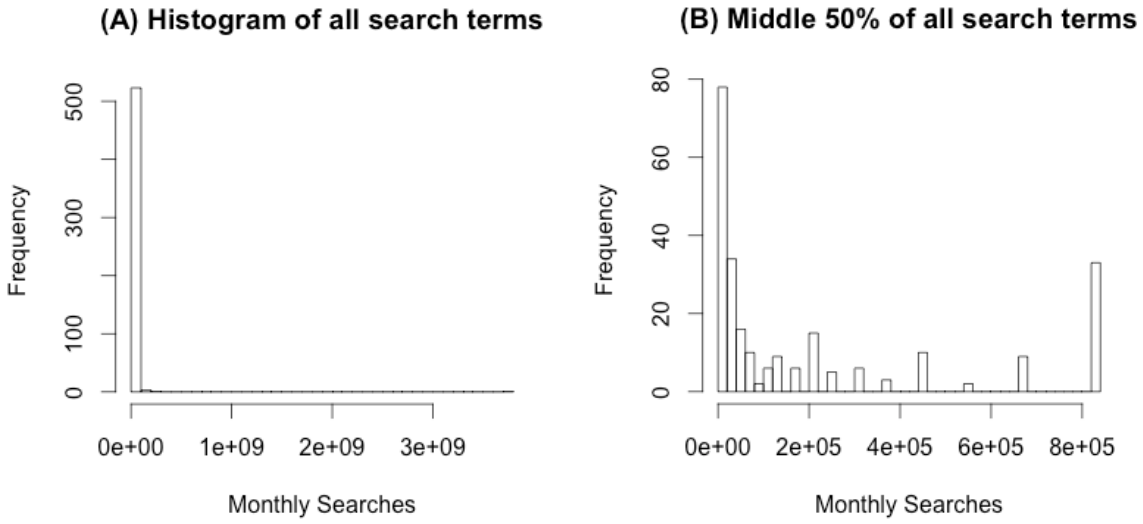
The goal of this paper will be to observe various Google Search terms in order to draw conclusions about the grouping of related search terms into "bundles."

To that end, the data used will be provided by the Open Advertising Dataset project (https://code.google.com/p/open-advertising-dataset/), which tracked a few more than 500 search terms daily over a period of months. I expect that related search terms will be more closely grouped than the overall data: that is, terms which are similar to each other will have similar amounts of searches during the same time period. If this is true, then it would be possible to predict the popularity of a search term T, given the popularities of related terms $\{t_1, t_2, \dots, t_n\}$.

Descriptive Statistics



The IQR for each observed bundle of terms, compared to the median of 50,000 size-15 samples randomly selected from all terms.



(A): The data initially looks narrowly distributed, because it has several massive outliers

(B): Inspecting a specific slice shows that the data is not normally distributed

Inference Statistics

In order to answer the question, "Are bundles of related search terms more closely grouped than unrelated search terms?" some ambiguities must be resolved. First, I should define how different terms are related. Ideally, search terms could be grouped by the category they fall under (e.g. celebrities might include the terms "Justin Bieber", "Neil Patrick Harris", and "Emma Watson" among others). However, sorting and categorizing these terms is a bit above my level of expertise, so I will instead define a bundle of search terms as a collection of different terms which share a unique word in common. Second, I must choose a valid test statistic to allow us to define how "closely related" a bundle of terms is. This statistic should have two properties:

1. It should give an accurate account of the range of the data

2. It should be relatively unaffected by outliers, since there are many large outliers in the dataset (e.g. searches for "facebook" are seven orders of magnitude larger than searches for "computing for idiots")

In order to fulfill these properties, I will be using the interquartile range (IQR) of the total monthly searches for each group of terms as a test statistic. Now, I can define our null and alternative hypotheses:

H_0: Bundles of related search terms will have the same or larger IQR than a randomly selected bundle from the overall population

H_1: Bundles of related search terms will have a smaller IQR than a randomly selected bundle from the overall population

The test will be one-tailed, using a permutation test of 10,000 random samples from the dataset, and comparing the bundle of related search terms according to the likelihood of its IQR appearing as a random sample of the same number of terms. Results follow.

Term	**sim**	**android**	**lawyer**	**insurance**	**credit**
.05 Quantile* (monthly searches)	197750	157075	128400	153722.5	130600
Observed IQR (monthly searches)	14380	31700	7700	177525	129600
Observed *p*-value	.0001	.0089	.0004	.0639	.0487

Term	**lemon**	**ink**	**equity**	**hair**	**canon**
.05 Quantile* (monthly searches)	103943.8	117600	95210	60475	37093.75
Observed IQR (monthly searches)	4575	677000	154800	2800	610000
Observed *p*-value	.0001	.2875	.0802	.0004	.2844

* Computed by taking 10,000 random samples from the entire collection of search terms. Each sample contained the same number of terms as the number of terms in the bundle. For example, there are 18 terms containing the term 'android,' so each of the 10,000 samples for that comparison contained 18 random terms.

Methodology

I began by searching through the full dataset for terms that contained words in common. For example, "android programming" and "android phone" both contain the word "android." I picked out bundles that had between 10 and 20 terms, which should allow for a decent size for each sample from the overall dataset. For each of these bundles of terms, I computed 10,000 samples of the same size as the number of terms in the bundle, and calculated the interquartile range of each sample. Next, I calculated the interquartile range of each bundle of terms, and compared it to the relevant sample of IQRs. I calculated both the value at the .05 quantile for the 10,000 samples and the *p*-value corresponding to the IQR for each bundle of terms.

Discussion

Of the selected terms, six out of ten have a *p*-value below the .05 quantile. The median *p*-value is .0288. Based on the median *p*-value, the null hypothesis is rejected at a significance level of 0.05, and the hypothesis that bundled search terms are more likely to be closely grouped is supported. This conclusion means that for a group of search terms that share some unique word, their popularity in terms of monthly searches will be similar. Although I use the median *p*-value here, the null hypothesis may have been able to be rejected more significantly if it was possible to calculate the overall probability, which is compounded for each term sampled.

The original question asked whether similar search terms are similarly popular. Unfortunately, I had to limit the scope of my research to bundles of terms that share a word. While my hypothesis about such terms was validated, the more general hypothesis about groups of similar terms is left open. With more data, and the ability to categorize

search terms better, the general hypothesis could be accepted. The common-sense explanation for both the limited and general hypotheses, however, is the same: similar search terms are similarly popular.