

AUTHORITYHACKER ANCHOR TEXTS STUDY 2019

Introduction

Google is quite open about using anchor texts in various ways (e.g. to determine the missing title, establish relevance, and so on).

Considering this, we were a bit surprised that there aren't many studies on anchor texts. In fact, this might be the only one.

Naturally, we were excited to get started.

Our Dataset

We worked with a bunch of different datasets. First, we gathered together a long list of higher volume keywords using AdWords API.

We then fed this to Ahrefs to generate a more ideas and filter out long-tail keywords. Overall, we ended up with about 1.5 million SERPs for these high-volume, long-tail keywords that we exported into a csv file.

We worked with the following key data points:

Keyword	Organic Traffic	Volume	Country	CPS	URL
Position	Backlinks	External Links	Internal Links	Referring Domains	Internal Backlinks
Anchor URL	Anchors Text	Anchor Count			

■ Adwords Keyword Planner ■ Ahrefs API
■ Ahrefs Keyword Explorer

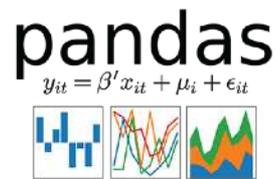
We wanted to work with content pages only. Having worked with some datasets before we wanted to avoid results with root domains, branded queries, wikipedia and youtube results, and so on that would yield slightly different metrics than traditional SEO optimized article or content page.

To achieve that, we simply removed all top 20 pages that either contained some root domains, or certain keywords in URL and using regex we ensured all URLs resembled a content page result.

We then used Ahrefs API to get metrics (internal, external outgoing links) and all anchors and anchor texts, plus some additional data.

We did some scraping with beautiful soup and a bit of more automated research on Ahrefs with Selenium module to confirm some findings or get additional insights.

Our tech stack



We relied on Keyword Planner, Ahrefs Keyword Explorer, Site Explorer and APIs. For automating some manual tasks we used Selenium.

To speed up some processes we used asyncio and aiohttp modules.

All data were stored in SQLite database and processed with Python, Pandas and Numpy.

Our methods and analysis

The process of finding insights was quite a straightforward one. We relied on regex pattern matching or simply string search and correlated the findings with search results position.

As we ensured our dataset was clean from possible outliers these correlations were also straightforward. Typically, we looked at median values.

We ran a number of different tests (i.e. mean values, spearman's correlations) to confirm the results and reviewed the dataset manually for some qualitative insights.

Qualitative insights were the important part as it helped us to avoid many mistakes that would yield inaccurate quantitative results.

For more information about this study, contact AuthorityHacker or Michal Ugor at hello@mchlgr.com