10,000 Results

Voice Search Ranking
Factors Study

BACKLINKO VOICE SEARCH SEO STUDY:
Methods & Results

BACKLINKO

As far as we're aware, a comprehensive, data-heavy study of Google Voice Search – specifically on the Google Home device – has not been undertaken to date.

So we set out to investigate the roles of various potential ranking signals that Google may use in their voice search algorithm.

SUMMARY

We asked 10,000 questions to Google Home (location set to USA via both a proxy and updating our location on the Google Home device's settings), and recorded the audio of each answer.

Questions may be asked of the Google Home (or other voice search enabled devices) by firstly 'activating' the device with the phrase *"Hey Google"* (or *"Okay Google"*), and then asking the question.

For example:

🎤 "Hey Google, do fish drink water?"

Answers where a result is found exist in various formats, but mostly a variation of:

"On the website **X**, they say: **Y**"

For example:

"On the website theguardian.com, they say: Only saltwater fish drink. In fresh water, the inside of the fish is saltier than the surrounding environment. Water moves into the fish by osmosis passively through the gills and the skin in the stomach. Fish have to eliminate all this excess water by peeing dilute urine."

We then transcribed these answers, and using a script, pulled the **website** (where present) from the beginning of the answer, and the **text from the website** (from the latter part of the answer) in order to find the URL of the website cited.

We then gathered data on each of these websites, the analysis of which forms the data for the study.

METHODOLOGY

KEYWORDS

We started with an initial seed list of 100 popular keywords from this post on SiegeMedia.

To generate further keywords from the seed keyword list, we decided to use KeywordTool.io. It allows generation of *question keywords or normal keywords* (standard keywords).

We put our seed list into KeywordTool.io, and initially generated further normal keywords for each of the 100 seed keywords.

Experimentation with a subset of these normal keywords, when asked to Google Home, gave a high error rate (responses that were variations of *"Sorry, I don't understand"*) when using these normal keywords.

We further experimented with a small subset of question keywords (generated from the seed list), and found the error rate to be much lower.

Based on this, we chose question keywords from KeywordTool.io as our main keyword list.

We compiled the various question keywords given for each seed keyword into a list (circa 19,000 keywords).

We then randomized this list, and took the first 10,000 keywords from it.

These 10,000 question keywords would be the questions asked to the Google Home device.

❓ QUESTIONS

We created a Google Sheet with the questions, and had a Golang script that would record Google Home's response, upload this audio clip to Google Drive, and then insert it into the Google Sheet next to the question. The audio was then transcribed and then inserted as a further column:

> For a number of reasons, it was not possible to fully automate this process, so our CTO Lloyd had to literally ask Google Home each of the 10,000 questions.

🌐 WEBSITES

Using the text and website parts of the response from Google Home, we were able to (through scraping normal Google SERPs) find the URL of the website, store HTML of the rendered page, and then analyze various aspects of the page, against different criteria.

This was achieved through a Golang script, which itself used Headless Chrome (to make sure that each website was fully rendered, such as those using a front-end JS framework like React) and Colly (to traverse the DOM of the results).

ANALYSIS

We wanted to find out whether the sites that Google Home chose as answers to our questions had attributes in common.

So with our answers in place, we determined the following initial criteria:

- **Percentage of results in Knowledge Graph**

- **Percentage of results in Featured Snippet**

- **Average PageRank in Ahrefs of result**

- **Average DomainRating in Ahrefs of result**

- **Result average/median word count**

- **Result reading level**

- **Schema used in result, and types of Schema often found**

- **Percentage of results that are FAQ pages**

- **Does the keyword appear in a <h2> tag anywhere on the page of the result?**

- **Is answer in Q&A format (ie. the keyword is in a question... And has an answer under it, like in an FAQ page)**

- **Does result contain the query in the <title> tag?**

- **Page load time of result**

- **Time To First Byte of result**

- **Percentage of results running on WordPress**

- **Average SERP position for the voice search result**

With our results in a MySQL database, we then analyzed them against the above criteria, which produced the results outlined above.