Query sampler: Generating query sets for analyzing search engines using keyword research tools

Sebastian Schultheiß [Corresp., 1] , **Dirk Lewandowski** [1, 2] , **Sonja von Mach** [1] , **Nurce Yagci** [1]

[1] Department of Information, Hamburg University of Applied Sciences, Hamburg, Germany

[2] Department of Computer Science and Applied Cognitive Science, University Duisburg-Essen, Duisburg, Germany

Corresponding Author: Sebastian Schultheiß
Email address: sebastian.schultheiss@haw-hamburg.de

Search engine queries are the starting point for studies in different fields, such as health or political science. These studies usually aim to make statements about social phenomena. However, the queries used in the studies are often created rather unsystematically and do not correspond to actual user behavior. Therefore, the evidential value of the studies must be questioned. We address this problem by developing an approach (query sampler) to sample queries from commercial search engines, using keyword research tools designed to support search engine marketing. This allows us to generate large numbers of queries related to a given topic and derive information on how often each keyword is searched for, that is, the query volume. We empirically test our approach with queries from two published studies, and the results show that the number of queries and total search volume could be considerably expanded. Our approach has a wide range of applications for studies that seek to draw conclusions about social phenomena using search engine queries. The approach can be applied flexibly to different topics and is relatively straightforward to implement, as we provide the code for querying Google Ads API. Limitations are that the approach needs to be tested with a broader range of topics and thoroughly checked for problems with topic drift and the role of close variants provided by keyword research tools.

1 # Query sampler: Generating query sets for analyzing search
2 # engines using keyword research tools

5 Sebastian Schultheiß[1], Dirk Lewandowski[1,2], Sonja von Mach[1], Nurce Yagci[1]

7 [1] Department of Information, Hamburg University of Applied Sciences, Hamburg, Germany
8 [2] Department of Computer Science and Applied Cognitive Science, University Duisburg-Essen,
9 Duisburg, Germany

11 Corresponding Author:
12 Sebastian Schultheiß[1]
13 Finkenau 35, 22081 Hamburg, Germany
14 Email address: sebastian.schultheiss@haw-hamburg.de

38

39

Abstract

Search engine queries are the starting point for studies in different fields, such as health or political science. These studies usually aim to make statements about social phenomena. However, the queries used in the studies are often created rather unsystematically and do not correspond to actual user behavior. Therefore, the evidential value of the studies must be questioned. We address this problem by developing an approach (query sampler) to sample queries from commercial search engines, using keyword research tools designed to support search engine marketing. This allows us to generate large numbers of queries related to a given topic and derive information on how often each keyword is searched for, that is, the query volume. We empirically test our approach with queries from two published studies, and the results show that the number of queries and total search volume could be considerably expanded. Our approach has a wide range of applications for studies that seek to draw conclusions about social phenomena using search engine queries. The approach can be applied flexibly to different topics and is relatively straightforward to implement, as we provide the code for querying Google Ads API. Limitations are that the approach needs to be tested with a broader range of topics and thoroughly checked for problems with topic drift and the role of close variants provided by keyword research tools.

Introduction

When investigating social phenomena in search engines, researchers build lists of queries, which they enter in the search engine and analyze the returned results. For instance, to determine whether Google prefers results of a particular political leaning, a list of queries can be designed, and the returned results analyzed. However, a central issue is whether the study results would be valid if another set of queries on the same topic was used. In the literature, we find that query sets are developed rather unsystematically, and therefore, the evidential value of such studies may be questioned. We address this research gap by developing an approach to sample queries from commercial search engines, including information on how often each query is entered, using commercial keyword research tools provided by search engine companies, such as Google. A myriad of studies uses query lists to generate search results to be further analyzed. In the information retrieval and information science communities, this is a standard procedure (e.g., Verma & Yilmaz, 2016; Fu, 2017). However, such query lists are also used in other fields. For instance, in medicine, studies examine the information quality of search results, for example, related to cancer diet (Herth et al., 2016) or breast cancer (Janssen et al., 2018). Moreover, in media and communications, researchers may aim to identify fake news for COVID-19-related queries (Mazzeo, Rapisarda & Giuffrida, 2021). An example from computational social sciences is a study investigating vaccine-related webpages using predefined terms such as "vaccine + danger" (Xu, 2019). Other exemplary studies assess Wikipedia's coverage when searching for psychological concepts (Schweitzer, 2008), examine results of political queries during election campaigns (Unkel & Haim, 2021), and in the field of sociology, analyze public information-seeking related to gun control and gun rights (Semenza & Bernau, 2022). The common aspects

80 of these studies are that authors use lists of search terms, enter the terms into Google or other
81 search engines, and analyze the retrieved results according to their research questions. The
82 question is where these queries come from. How do researchers generate query sets that reflect
83 what users search for? To what degree do researchers succeed in assembling such query sets? In
84 some studies, the authors identify this problem and discuss whether the queries used in their
85 research, such as medical or political terms, may not fully match the phrases people actually use
86 when searching for information on the respective topic (Herth et al., 2016; Unkel & Haim, 2021;
87 Semenza & Bernau, 2022). Inevitably, the question arises: Would the respective study have
88 yielded different results if the authors had used other queries?
89 From the thematic diversity of studies using query sets and the problem of considering real
90 search queries, two requirements of an ideal query set are derived, (1) consideration of query
91 popularity and (2) topic coverage. First, the query set should reflect the users' search interests by
92 considering the search volume (popularity) of the queries. This prerequisite is essential for
93 studies that aim to draw conclusions about social phenomena, for example, from health (Janssen
94 et al., 2018) or political science (Unkel & Haim, 2021). Second, the ideal query set should cover
95 the queries on any topic as comprehensively as possible.
96 Queries submitted by real search engine users of all search engines worldwide can be considered
97 the "ideal" of a query set. These queries would be most likely yield valid information about
98 social phenomena since they would map the search interests of the entire online society and
99 would be independent of specific search engine providers. The closest to this holy grail would be
100 the use of panel data from permanently observed users. Such data, however, would be only a
101 sample and no longer a complete survey. While such a sample may be appropriate for many
102 studies, it may not be sufficient when the aim is to cover topics in their entirety, especially with
103 long-tail phenomena. Another promising approach is the Archive Query Log (AQL). Once the
104 data collection is completed, AQL will contain approximately 356 million queries from several
105 search engines extracted from the Internet Archive (Reimer et al., 2023). Instead of data from
106 multiple search engines, some researchers have access to queries from individual search engine
107 providers. However, such data are not freely available to researchers, and studies based on this
108 data usually focus on optimizing the respective search systems (e.g., Dang, Kumaran & Troy,
109 2012; Baytin et al., 2013). Thus, a central methodological challenge is to find alternative
110 approaches for developing query sets.
111
112 **Research questions.**
113 In this paper, we answer the following research questions (RQs):
114 **RQ1**: What approaches are used by scholars to generate query sets, and what are the strengths
115 and weaknesses of these approaches?
116 **RQ2**: What is a suitable approach for generating a query set that considers query popularity and
117 allows the coverage of a specific topic?
118
119 **Research gap.**

120   Scholars reuse or generate query sets by applying numerous methods. In information retrieval
121   research, test collections consisting of queries, documents, and relevance judgments are essential
122   (for overview of Text Retrieval Conference (TREC) test collections see Harman & Voorhees,
123   2007). However, queries in test collections are static, limited to specific languages, and partially
124   prefiltered, even if a test collection includes queries provided by a search engine provider (e.g.,
125   MS MARCO dataset with Bing data; Craswell et al., 2021).
126   We found seven types of query sets researchers generate:

127   1.   Queries delivered by a search engine provider, for example, Microsoft (Azzopardi et al.,
128   2020).
129   2.   Popularity data; this includes the use of all tools that provide data on query popularity,
130   while the scope of functions differs greatly in these tools, for example, Alby (2020).
131   3.   Autocomplete suggestions users receive as a drop-down list when entering queries, for
132   example, Haak & Schaer (2022).
133   4.   Content extracted from online communities, such as the AskDocs section of Reddit
134   (Zuccon et al., 2016).
135   5.   Queries provided by subjects asked to generate queries, for example, utilizing an online
136   survey (Bilal & Ellis, 2011).
137   6.   Queries developed by the study authors based on specific criteria, for example, query
138   type (Schultheiß, Sünkler & Lewandowski, 2018).
139   7.   Predetermined lists of terms, for example, names of political candidates (Hinz, Sünkler &
140   Lewandowski, 2023).

141

142   The commonality in the mentioned approaches is that they are only partly suitable for drawing
143   conclusions about social phenomena since they were either created for other objectives (e.g., test
144   collections for information retrieval research, as described above), contain terms that users may
145   not use (e.g., lists of technical jargon), or are limited in scalability (e.g., content from online
146   communities). The overall problem with these approaches is that the query sets represent actual
147   user querying behavior in some way, but researchers cannot guarantee that they represent the
148   entire user population (e.g., all users or a particular user group). Therefore, the evidential value
149   of such studies can be questioned.
150   In this study, we address this issue and present an approach that allows building extensive sets of
151   queries on any topic. We do this by using lists of terms and keyword research tools. Such tools
152   are offered by search engine marketing companies (e.g., Semrush) and commercial search engine
153   companies, such as Google, to allow marketers to plan their campaigns (see Section *Selecting a*
154   *keyword research tool*). Keyword research tools find additional keywords based on already
155   known relevant keywords; that is, the tools make suggestions for more keywords that can be
156   used to address customers. The tools also predict the query volume for each search query, that is,
157   the predicted number of searches per month. By offering keyword ideas, the tools aim to support
158   marketers. In this context, keywords are terms or phrases a website should show up on a search

159   engine, while queries refer to the actions of the users1. In this paper, we will use both terms
160   synonymously.
161
162   **Structure of the paper.**
163   The rest of the paper is structured as follows. First, we show which approaches for creating a
164   search query set have been used in the literature and their advantages and disadvantages.
165   Subsequently, we describe our approach to generating a query set. The approach consists of (1)
166   selecting an initial list of terms, (2) including synonyms and alternative spellings, (3) selecting a
167   keyword research tool, and (4) generating keyword ideas. The description of the approach is
168   followed by its empirical verification using query data from two published studies where queries
169   formed the basis of analysis (Herth et al., 2016; Lewandowski, Sünkler & Yagci, 2021). Finally,
170   we discuss the results and present suggestions for future research.
171

172   ## Literature review
173   Through an extensive literature search, we identified seven approaches to generating query sets.
174   Using Scopus and Google Scholar, we searched for articles containing words related to query
175   sets (e.g., set of queries, list of queries) together with words related to search engines (e.g.,
176   search engine, Google). We focused on finding articles describing how the query sets were built,
177   regardless of the study's objective. Table 1 details these approaches, summarizing the extent to
178   which the prerequisites of popularity consideration and topic coverage are considered. The
179   following sections describe the approaches we identified in the literature in more detail.
180
181   (please insert Table 1 here)
182
183   **Queries delivered by a search engine provider.**
184   Using queries delivered by search engine providers is the most promising of the feasible
185   approaches listed in Table 1 since both criteria (popularity consideration and topic coverage) are
186   met. Examples are studies where researchers have access to transaction logs from search engines
187   such as Google (Kinney, Huffman & Zhai, 2008), Bing (Dang, Kumaran & Troy, 2012; Das et
188   al., 2017), Yahoo (Goel et al., 2010), AOL (Lucchese et al., 2013), Yandex (Baytin et al., 2013),
189   Excite (Gravano, Hatzivassiloglou & Lichtenstein, 2003), Sogou (Whiting, Jose & Alonso,
190   2016), or T-Online (Lewandowski, 2015). The authors of these studies focus on improving the
191   performance of their company's search systems, for example, regarding autocorrection (Baytin et
192   al., 2013) or query reformulations (Dang, Kumaran & Troy, 2012). A significant disadvantage is
193   that analyses depend on search engine providers' willingness to provide researchers with data. It
194   is highly unlikely that search engine providers grant access to their data when researchers wish to
195   investigate topics that are outside the providers' self-interests (Lewandowski, Sünkler &
196   Schultheiß, 2020). Thus, independent decisions by researchers regarding the thematic and
197   quantitative scope of the data are barely possible. Furthermore, the provided queries refer only to

---

1 https://www.searchenginejournal.com/understanding-difference-queries-keywords/126421/#close

198   a single search engine without allowing comparisons between different search engines. Even
199   when researchers have access to transaction logs of multiple search engines, comparing the data
200   is quite challenging. Jansen & Spink (2006) enumerated differences between nine transaction
201   logs. These differences concern, for example, different time spans when the logs were created
202   and missing numbers of sessions and terms in two logs.
203
204   **Popularity data.**
205   Since most researchers do not have access to queries from search engine providers, one possible
206   solution is popularity data. Popularity data contain queries including information on their
207   popularity (i.e., search volume), with the accuracy of search volume data varying considerably.
208   Popularity data are made available through tools, mainly Google Trends. Other studies use
209   keyword research tools, such as Google Keyword Planner. Predetermined terms or lists of terms,
210   for example, on specific topics, serve as the basis (i.e., seed terms, see Section *Selecting a*
211   *keyword research tool*). This differentiates the popularity data approach from the predetermined
212   lists approach (see Section *Predetermined lists of terms*), in which the list entries are
213   synonymous with the queries used. The studies by Ballatore (2015) and Fumagalli, Bailoni &
214   Giunchiglia (2020) are examples of studies using Google Trends. Ballatore (2015) selected the
215   most popular queries from Google Trends for several conspiracy theories, while Fumagalli et al.
216   (2020) used Google Trends to generate queries relating to Schema.org types, for example, book
217   series or creative work. Tana (2018) used Google Trends to retrieve the top queries for seed
218   terms such as "depression." Google Trends allows access to actual searched terms during a
219   specific time episode. However, Google Trends provides not absolute but normalized data that
220   express the search volume of the respective term in relation to the search volume of all other
221   searches at a given time[2]. Absolute numbers on search volume are provided by keyword research
222   tools such as Google Keyword Planner[3]. For instance, in a study commissioned by a German
223   health insurer (Central, 2015), the authors used a predetermined list of $N = 50$ common diseases
224   as seed terms for forming term clusters consisting of the disease terms (e.g., hyperkinetic
225   disorder) as well as frequently used synonyms (e.g., ADHD) and additional terms (e.g., doctor)
226   by using a keyword research tool[4]. Similar approaches were taken by Alby (2020) and
227   D'Ambrosio et al. (2015). Regarding the topic of skin diseases, Alby (2020) used disease terms
228   and synonyms to build search queries ($N = 2{,}397$) via Google Keyword Planner, while
229   D'Ambrosio et al. (2015) used preconception-related keywords to obtain queries ($N = 29{,}132$)
230   that were actually searched for by Italian Internet users. For Google Keyword Planner, however,
231   the limitation must be added that Google greatly reduced the accuracy of the data for accounts
232   with low AdWords sales in 2016. Since then, search volumes are provided only in broad ranges
233   (e.g., $10 - 100$ or $100 - 1K$ average monthly searches)[5]. This means that studies similar to the
234   pre-2016 studies using Google Keyword Planner could not be conducted anymore, at least not

---

[2] https://support.google.com/trends/answer/4365533?hl=en
[3] https://support.google.com/google-ads/answer/7337243?hl=en
[4] The Authors did not mention Google Keyword Planner explicitly, but its usage can be assumed.
[5] https://www.seroundtable.com/google-keyword-planner-throttled-22535.html

235   with accurate search volume data. Waller (2011) used data from web analytics company Hitwise
236   (now a division of Connexity)[6]. The sample covered queries typed into Google Australia over a
237   4-week period in 2009.
238   Exact values for the search volume of individual keywords are also provided by Keyword Magic
239   Tool[7] from Semrush. Some tools use more than one source to provide their keywords, but the
240   exact methods from which sources keywords and search volume are generated are kept secret.
241   However, not having insights into the origin of the data is a serious issue in academic research.
242
243   **Autocomplete suggestions.**
244   Search engines deliver ideas to help the user to formulate their information need. Users receive
245   common queries as a drop-down via autocomplete suggestions. The predictions match what a
246   user started to enter and incorporate other factors, such as trending interest in the query. Search
247   engines do not provide autocomplete suggestions for all content. For example, Google prevents
248   predictions that are in violation of Google policies, such as sexually explicit, dangerous, or
249   harassing content[8]. The same holds for Bing, as it filters spam and adult and offensive content
250   from the suggestions[9]. Nevertheless, in some cases, autocomplete suggestions can contain
251   misinformation that may hurt organizations or individuals (Hiemstra, 2020).
252   Autocomplete predictions are used by researchers to create query sets. Haak & Schaer (2022)
253   crawled $N = 21,407$ autocomplete suggestions from Google to analyze person-related
254   suggestions for biases. Wu et al. (2016) developed a system that discovers query patterns (e.g.,
255   "jobs in [location]") by using query autocomplete features. The authors aimed to discover a
256   focused set of queries that center around an entity. In contrast, Bar-Yossef & Gurevich (2008)
257   developed algorithms for sampling random autocomplete suggestions. Fumagalli, Bailoni &
258   Giunchiglia (2020) used "Answer the public"[10], a tool that uses autocomplete suggestions from
259   Google and Bing and organizes the queries according to different criteria, such as question type.
260   Haider (2016) used autocomplete to define the queries for her study on informational structures
261   on waste sorting.
262   Since autocomplete suggestions rely on actions taken by users, real user behavior is reflected.
263   However, since the suggestions come from the search engine provider, the underlying algorithm
264   and ranking factors for the autocomplete predictions can only be understood rudimentarily from
265   the outside. In addition, it is crucial to remember that creating suggestions is complex, based on
266   many influencing factors, such as a user's past searches. Furthermore, suggestions excluded due
267   to inappropriate content, as described above, result in incomplete sets of user searches.
268
269   **Content extracted from online communities.**

---

[6] https://connexity.com/ (Hitwise was acquired by Connexity in 2015.)
[7] https://de.semrush.com/analytics/keywordmagic/?q=adhs&db=de
[8] https://support.google.com/websearch/answer/7368877
[9] https://blogs.bing.com/search/2013/03/25/a-deeper-look-at-autosuggest
[10] https://answerthepublic.com/

270 Using content from online communities differs significantly from the previous approaches, as no
271 queries from web search engines are considered. Instead, based on discussions in online
272 communities, researchers map popular search queries for the topics discussed.
273 Yilmaz et al. (2019) used questions posted on an educational Q&A website to build a query set
274 in the Turkish language. Related to medical topics, Zuccon et al. (2016) created a query set
275 modeled after distinct topics from forum posts from the AskDocs section of Reddit, designed to
276 resemble laypeople's health queries. The same approach was followed by Soldaini & Goharian
277 (2017). Similarly, Liu, Fang & Cai (2015) selected question-like queries from topics of medical
278 forums such as drugs.com, while Zhang (2012) selected tasks from Yahoo! Answers (to search
279 for in MedlinePlus). Finally, Azizan, Bakar & Rahman (2019) used content from online forums,
280 blogs, social media, and Google Instant to create a query set related to agriculture.
281 A major disadvantage of query generation via online communities is that the queries and their
282 popularity can only be modeled in the context of the respective online community but not
283 beyond. Whether or how often the queries generated in this manner are searched for via search
284 engines remains unclear.
285

286 **Queries provided by subjects.**
287 The approach of queries provided by subjects encompasses all studies in which the authors use a
288 group of subjects who are asked to generate queries based on certain specifications.
289 To obtain data generated by real people, Bilal & Ellis (2011) identified $N = 130$ tasks in the
290 literature from 1989 to 2011 that were assigned to children and/or self-selected by them. Then
291 the way children queried was examined, and the words used built the foundation of the query set.
292 During the 2018 U.S. midterm elections, Trielli & Diakopoulos (2022) analyzed whether search
293 results differ for members of different ideological groups. As a basis for the search results to be
294 analyzed later, queries were needed. For this purpose, the authors conducted online surveys in
295 several states, asking the subjects what terms they would use when searching for information
296 about a candidate. To build a test collection (UQV100), Bailey et al. (2016) used crowdsourcing
297 and collected $N = 5,764$ unique queries from $N = 263$ workers.
298 Using this approach has limitations. First, researchers receive queries from real people but not
299 from a natural situation (using a search engine). In addition, self-reported behavior ("What
300 search terms would you use?") does not necessarily reflect natural user behavior. Thus, it
301 remains unclear whether the subjects would have searched in the way that they stated in the
302 survey.
303

304 **Queries developed by the study authors.**
305 This approach includes all studies in which researchers develop the queries themselves. The
306 authors do not consider whether or how frequently real search engine users use the queries, and
307 they do not use predetermined lists of terms, as described in the Section "*Predetermined lists of*
308 *terms*."

309   The only basis for self-creating the queries are criteria such as query complexity (Singer,
310   Norbisrath & Lewandowski, 2012), query type (e.g., Schultheiß, Sünkler & Lewandowski, 2018;
311   Schultheiß & Lewandowski, 2021), or other criteria, such as the number of content farm articles
312   per query (McCreadie et al., 2012). Queries developed by researchers is the least appropriate
313   among the approaches described in this paper since queries are arbitrarily arranged, and
314   popularity is not considered. However, queries developed by authors can serve as a starting point
315   (i.e., the initial list of terms, see Section "*Selecting an initial list of terms*") for creating further
316   queries.
317
318   **Predetermined lists of terms.**
319   Another approach to generating search queries is using predetermined lists with different
320   thematic focuses. The lists differ in terms of their coverage range from relatively small samples
321   to complete lists, for example, of all political parties or the names of all candidates running for
322   an election.
323   An exemplary study using a complete list is the analysis by Hinz, Sünkler & Lewandowski
324   (2023). For the 2021 German federal election, the authors analyzed whether candidates use
325   search engine optimization (SEO) on their personal websites. The analysis was based on the
326   complete list of all candidates in the election ($N = 6,211$). Other studies also used predetermined
327   but sampled lists. For instance, Torres & Rogers (2020) combined the names of political parties
328   with specific issues associated with the political agendas found on official party websites or in
329   Facebook comments. Hussain et al. (2019) used keyword captions of images to form queries for
330   a retrieval effectiveness study regarding image search engines. Leontiadis, Moore & Christin
331   (2011) generated a query set focusing on search-redirection attacks. The authors issued a seed
332   query ("no prescription Vicodin") and then collected search phrases found on the retrieved pages
333   linking to websites the attackers wished to promote, for example, online pharmacies. Another
334   approach using predetermined lists is the project "data donation" ("Datenspende"). A plugin
335   installed in the browser of the participants ("donors") conducted searches for predefined terms at
336   regular intervals and sent the results of the first search engine result page (SERP) back to the
337   researchers (Krafft, Gamer & Zweig, 2019). However, the queries selected by the researchers
338   were the precise names of political parties and selected politicians. Whether or how frequently
339   search engine users actually used these queries remains unknown (e.g., one can easily see from
340   tools like Google Trends that the query "Bündnis90/Die Grünen" for the German Green party is
341   searched only seldom, as the party is usually referred to as "Grüne" or "Die Grünen"). Another
342   data donation study, with the same query selection limitations, focused on health-related queries
343   (disease + clinical term; Reber et al., 2020).
344   To summarize, for queries on predetermined lists, it remains unknown whether and how
345   frequently they were used by search engine users. While using lists of predetermined terms alone
346   has limitations, lists can serve as a basis for the query set to be created by using the approach
347   described in the following section.
348

349 **Query sampler: an approach for generating query sets**
350 We propose an approach for generating a query set under the precondition that researchers do not
351 have direct access to queries from a commercial search engine provider like Google but still aim
352 for query sets that are representative in terms of query popularity and topic coverage. The
353 approach aims to cover the search interest related to an initial list of terms by building extensive
354 sets of queries on any topics. In doing so, we combine initial lists of terms and popularity data by
355 utilizing keyword research tools.
356 The approach is illustrated in Fig. 1 and consists of the following steps: (1) selecting an initial
357 list of terms, (2) including synonyms and alternative spellings (optional), (3) selecting a keyword
358 research tool, and (4) generating keyword ideas.
359
360 (please insert Fig. 1 here)
361
362 **Selecting an initial list of terms.**
363 First, a list of initial terms is selected. The initial list includes seed terms that form the basis of
364 the query set to be created (i.e., initial list = seed list for round 1). The entries of the initial list
365 can come from various sources; for example, they can be compiled through brainstorming by the
366 researchers or be predetermined lists with or without a thematic focus (see *Literature review*).
367 Existing sources suitable for reuse are, for example, randomly selected Wikipedia articles[11] when
368 a cross-topic list is to be created or Google Trends[12] or Twitter Trends[13] if trending topics are to
369 form the basis of the list. Topic-specific lists, in contrast, can also come from multiple sources
370 such as online communities, for example, the AskDocs section of Reddit (Zuccon et al., 2016),
371 or public authorities, for example, a list of important terms regarding topics of domestic policy
372 published by the German Federal Ministry of the Interior and Community[14] or a term index
373 provided by the German Federal Department for Media Harmful to Young Persons[15].
374
375 **Including synonyms and alternative spellings (optional).**
376 When the aim is to achieve high thematic coverage in the resulting query set, researchers should
377 consider including synonyms and alternative spellings in the initial list of terms. It can be
378 assumed that the approach described in this paper will suggest several synonyms and alternative
379 spellings, making this step obsolete, at least in theory. However, especially in the case of
380 specialized vocabularies, such as medical terms, it cannot be assumed that this applies to all
381 synonyms and alternative spellings. We illustrate the inclusion of synonyms and alternative
382 spellings with a brief example in Section "*Synonyms and alternative spellings*".
383

---

[11] https://en.wikipedia.org/wiki/Special:Random
[12] https://osf.io/q7wt3
[13] https://twitter.com/i/trends
[14] https://www.bmi.bund.de/DE/service/lexikon/lexikon-node.html
[15] https://www.bzkj.de/resource/blob/197826/5e88ec66e545bcb196b7bf81fc6dd9e3/2-auflage-gefaehrdungsatlas-data.pdf

Selecting a keyword research tool.

384 The next step is selecting a suitable keyword research tool (Fig. 1, step three). We explain the

385 reasons for using Google services as the first choice from our point of view and the limitations of

386 Google Keyword Planner and alternative tools.

387

388 We used Google Keyword Planning services for our approach for two main reasons. First,

389 Google is the most popular search engine on the web. In the U.S., about 87% of all queries are

390 submitted to Google (StatCounter, 2023a), and in Europe, 92% (StatCounter, 2023b). Therefore,

391 Google achieves the highest coverage of the Internet user community, allowing more reliable

392 statements on socially relevant topics. Second, even if it remains unclear how the keyword ideas

393 are generated, their origin can be limited to Google, which does not apply to alternative tools as

394 described at the end of this section.

395 However, using Google Keyword Planner[16] through the standard user interface has limitations.

396 For instance, the tool provides keyword ideas for up to 10 seed terms only[17], which restricts its

397 usefulness, especially when it comes to extensive lists of seed terms. Additionally, one must run

398 an ad campaign of considerable size to obtain precise data on the average search volume of the

399 generated keyword ideas. Otherwise, the Google Keyword Planner delivers only approximate

400 search volume estimates (e.g., 100–1,000; 10,000–100,000)[18], which are not very useful for

401 research studies.

402 The limitations of the regular Google Keyword Planner do not apply to the Google Ads API. The

403 Google Ads API enables users to generate large sets of keyword ideas, including precise data on

404 search volume. To send requests to the Google Ads API, users need to authenticate the usage of

405 their Google account via the Google Cloud Console. A client id (username) and client secret

406 (password) are generated by creating a new project. Additionally, a refresh token must be

407 generated by using the previous parameters. This token needs to be updated weekly to ensure the

408 account security. The Python library GoogleAds requires these parameters together with the

409 developer token and the ID of the Google Ads account to make calls to the API. Keyword ideas

410 are generated by calling the KeywordPlanIdeaService[19] and using the

411 GenerateKeywordIdeasRequest. The input parameters are a keyword, location id, and language

412 id. No active ad campaign is necessary to use the API, as with the regular Google Keyword

413 Planner, but a basic access token must be applied for[20]. The application must include several

414 details, such as the reasons for applying to use the API. Our statement that we will use the API

415 for research led to the approval of our application.

416 Besides Google Keyword Planning services, a number of alternative tools are available, with

417 Keyword Magic Tool from Semrush[21] being among the most popular. Alternative keyword

---

[16] https://support.google.com/google-ads/answer/7337243?hl=en

[17] https://support.google.com/google-ads/answer/9327909?hl=en

[18] https://www.seroundtable.com/google-keyword-planner-throttled-22535.html

[19] https://developers.google.com/google-ads/api/docs/keyword-planning/generate-keyword-ideas

[20] https://developers.google.com/adwords/api/docs/access-levels?hl=en

[21] https://de.semrush.com/analytics/keywordmagic/?q=adhs&db=de; Please note that Semrush is only meant to allow a comparison to Google services and is representative of many similar tools.

418   research tools are used when the query set to be created addresses a topic that is regarded as
419   inappropriate by Google. Google does not serve ads for inappropriate content, which means that
420   Google Keyword Planner does not provide any keyword ideas for such content either. Google
421   defines inappropriate content, among other things, as dangerous or derogatory content (e.g.,
422   content promoting hate groups or hate group paraphernalia), sensitive events (e.g., ads appearing
423   to profit from a tragic event with no discernible benefit to users), or sexually explicit content[22].
424   Tools such as Semrush do not have such restrictions, so keyword ideas are generated even for
425   content that Google considers inappropriate. One severe disadvantage of Semrush and similar
426   tools is that the origin of the keyword ideas is not transparent. According to Semrush, the
427   keyword ideas are based on data from third-party suppliers[23]. However, it remains unclear who
428   the third-party suppliers are and which keyword idea comes from which source.
429

430   **Generating keyword ideas.**
431   We intend to cover the search interest related to the initial list of terms using the Google Ads API
432   service "KeywordPlanIdeaService." Our approach is to resend the keyword ideas generated by
433   the initial list of terms to the Google Ads API to gradually receive not only more but also more
434   specific keyword ideas. As illustrated in Fig. 1, step four, this procedure is repeated in several
435   rounds until no new keyword ideas emerge and saturation for the initial terms can be assumed
436   (Strauss & Corbin, 1998, p. 143).
437   The process for each round is outlined below.
438   Round 1:
439   1. Collecting keyword ideas for all terms from the initial list
440   2. Cleaning the keyword ideas from ideas without search volume to ensure the criterion of
441   popularity
442   Round 2 and all further rounds:
443   1. Collecting keyword ideas for all remaining keyword ideas from the previous round
444   2. Cleaning the keyword ideas from ideas without search volume to ensure the criterion of
445   popularity
446   3. Removing duplicates within the same round (i.e., keyword ideas generated by more than
447   one initial term of the respective round)
448   4. Removing duplicates with previous rounds (i.e., keyword ideas that have already been
449   generated in a previous round)
450

451   **Proof of concept**
452   To test our approach, we selected two published studies for comparison purposes. Study one is
453   about the quality of information on cancer diet (Herth et al., 2016), and study two is on SEO for
454   COVID-19 and radical right topics (Lewandowski, Sünkler & Yagci, 2021). The studies were

---

[22] https://support.google.com/adspolicy/answer/6015406
[23] https://www.semrush.com/kb/998-where-does-semrush-data-come-from

455 selected for their differences in terms of topic and scope, allowing a first impression of the
456 generalizability and scalability of our approach.
457   1. Both studies would have benefited from our approach, as a greater variety of queries and,
458   thus, web pages would have strengthened the analyses.
459   2. The studies come from different subject areas. The radical right topics (Lewandowski,
460   Sünkler & Yagci, 2021) allow a test of the described problem regarding Google's position
461   on inappropriate content (see Section "*Selecting a keyword research tool*"), that is, whether
462   keyword ideas are generated at all for such terms.
463   3. The initial lists of terms used in the studies vary in size.
464 We tested our approach by generating keyword ideas for the queries used in the studies and
465 comparing the resulting keyword ideas in terms of number and search volume with the original
466 studies. For better comparability with the example studies, we omitted considering synonyms
467 and alternative spellings (see Section "*Including synonyms and alternative spellings (optional)*").
468

469 **Study on cancer diet.**
470 The first study is from the medical field. The authors evaluate the quality of online patient
471 information about cancer diet (Herth et al., 2016). For the term "Krebsdiät"[24] (English: "cancer
472 diet"), the authors manually collected the first $N = 100$ organic results using the German version
473 of Google and analyzed the quality of the results according to formal and content criteria, for
474 example, transparency concerning provider and completeness. In their discussion, the authors
475 present a short keyword analysis they conducted a few months after the study using Google
476 Keyword Planner. The analysis showed that most users do not search for "cancer diet" but for
477 more specific information on cancer diet or cancer diets by name, such as "ketogenic diet" or
478 "nutrition in cancer." Hence, the authors conclude that a more detailed evaluation of patient
479 information with more specified keywords is needed in future studies.
480 Table 2 shows keyword ideas we generated for the initial term "cancer diet" in five rounds. In
481 columns two and three, the number of seed terms and the number of generated keyword ideas for
482 these terms in each round are presented. In addition, the excluded keyword ideas are shown, that
483 is, the number of keyword ideas with a search volume of 0, the number of duplicates within the
484 current round, and the number of duplicates with already existing keyword ideas. The difference
485 between generated and excluded keyword ideas is shown in the column of remaining keyword
486 ideas. These form the basis ("seed terms") for the next round. The two rightmost columns show
487 the search volume of the remaining keyword ideas per month, on average and sum[25].
488

---

[24] After consulting one of the authors, the use of the term "Krebsdiät" was confirmed since the original
German term is not explicitly mentioned in the paper.
[25] Google Keyword Planning services provide aggregated search volume data for keywords, e.g., "MBA" and
their close variants, e.g., "masters of business administration." Keywords and their close variants are reported
with identical search volumes. Thus, the sum of the search volume may be higher than the *actual* search
volume. See https://www.searchenginewatch.com/2016/09/26/reliable-search-volume-data-a-glimmer-of-
hope/

489   (please insert Table 2 here)

490

491   In five rounds of collecting keyword ideas, we generated $N = 98$ unique keyword ideas (the sum

492   of the remaining keyword ideas of each round) for the initial term "cancer diet." The keyword

493   ideas of rounds three and four contain the highest number of duplicates. Due to excluding

494   duplicates, the number of remaining keywords decreased considerably with each iteration. While

495   from round two, 64% of the generated keyword ideas serve as seed terms for the next round, in

496   round three, it is only 3%, and in round four, only one term. In round five, no new keyword ideas

497   were generated. Together with the decreasing added monthly search volume in each round, this

498   finding indicates a saturation regarding the general topic of the study (cancer diet).

499   The original study examined one term ("cancer diet") with a search volume of $N = 260$ average

500   monthly searches at the time the study was conducted, according to the analysis by the authors[26].

501   By applying our approach, we expanded the term to a list of $N = 98$ terms, with an added search

502   volume of $N = 2,144$ monthly searches on average. Thus, the query used in the original study

503   only covers 1% of queries and 12% of the projected search volume generated through our

504   approach. By using our approach, researchers would have achieved a better evidential value even

505   if they had cut off the list due to limited resources to analyze data for all queries. For instance, a

506   cut-off after round two still would have covered $N = 77$ queries and 83% of the total search

507   volume. This shows that researchers to do necessarily need to use all queries generated using our

508   approach and still can increase the evidential value of their studies.

509

510   **Study on search engine optimization for radical right and COVID-19 topics.**

511   The second study we tested our approach on is about SEO (Lewandowski, Sünkler & Yagci,

512   2021). Using SEO indicators such as the usage of a site title, page speed, or usage of HTTPS, the

513   authors built a rule-based classifier to determine the probability of SEO on a web page. To test

514   the classifier, three query sets from Google Trends, including one on radical right content and

515   one on the topic of COVID-19 were used. Through screen scraping, Google search results for the

516   queries were collected and then classified according to their SEO probability. The results show

517   that a large fraction of web pages found on Google are optimized (Lewandowski, Sünkler &

518   Yagci, 2021). To test our approach, we used a sample of $N = 15$ queries of the COVID-19

519   ($N = 271$) dataset[27] and the full dataset of the radical right ($N = 82$) queries.

520   As Table 3 shows, we generated $N = 385$ keyword ideas for the COVID-19-related queries in

521   three rounds since no new keyword ideas were generated in round three. Most keyword ideas

522   were delivered in round one. As in the previous study on cancer diet, many keyword ideas were

523   excluded because they were duplicates. The queries used in the sample ($N = 15$) from the

524   original study cover 4% of the queries we cover with our approach. For the initial terms of the

---

[26] In our replication, "cancer diet" had an average search volume of $N = 149$ monthly searches.

[27] As "dataset," we refer to the initial lists of terms included in the research data. In three columns, the research data lists (1) the keyword ideas, (2) the round in which the idea was generated, and (3) the search volume of the respective keyword idea. Keyword ideas that are indicated with round "0" are the terms of the initial lists that were used to generate the keyword ideas in all further rounds.

525   published study, we identified an aggregated search volume of $N = 1{,}473{,}536$ monthly searches,

526   which is 35% of the search volume generated by our approach ($N = 4{,}170{,}514$ monthly

527   searches).

528

529   (please insert Table 3 here)

530

531   For the radical right queries, we generated $N = 278$ keyword ideas in five rounds, as shown in

532   Table 4. The queries used in the original study cover 29% of queries we cover with our

533   approach. For the list of initial terms of the published study, we identified an aggregated search

534   volume of $N = 27{,}117$ monthly searches, which is 16% of the search volume generated by our

535   approach ($N = 166{,}227$ monthly searches). No ideas were generated for 46% of the initial terms

536   ($N = 38$), including right-wing extremist numeric codes such as "1488." It can be assumed that

537   Google classifies such unambiguous right-wing terms as inappropriate, so no keyword ideas are

538   generated (see the explanation in Section "*Selecting a keyword research tool*").

539

540   (please insert Table 4 here)

541

542   **Cumulative keyword ideas and search volume of all studies.**

543   The following figures show the cumulative keyword ideas (Fig. 2) and search volume (Fig. 3) of

544   all studies and rounds, that is, the sum of keyword ideas and search volume that have been added

545   up to a certain round.

546   From Fig. 2, we see that the most keyword ideas were collected after two rounds. It is noticeable

547   that for "cancer diet" the highest gain in new keyword ideas was achieved from round one to

548   round two. This can be explained by the fact that the initial list of terms consisted of only one

549   query (namely "cancer diet"), while the lists of terms for the other two topics were more

550   extensive ($N = 15$ queries for COVID-19 and $N = 82$ queries for radical right). No new ideas

551   were generated after four rounds (at the latest).

552

553   (please insert Fig. 2 here)

554

555   From the cumulative search volume shown in Fig. 3, differences between the dataset on radical

556   right topics and the other two datasets become clear. The search volume of the keyword ideas on

557   cancer diet and COVID-19 increased considerably in round two (from 19.2% to 83% and from

558   33.9% to 100%, respectively). However, the situation is different with the radical right queries.

559   A total of 96.7% of their search volume was already covered in round one, so the further rounds

560   increased the search volume only slightly.

561

562   (please insert Fig. 3 here)

563

564   **Synonyms and alternative spellings.**

565   Here, we illustrate with a short example that it is worthwhile to include synonyms and alternative
566   spellings in the initial list of terms to achieve high thematic coverage of the resulting query set
567   (see Section "*Including synonyms and alternative spellings (optional)*"). In her study on skin
568   diseases, Alby (2020) built an initial list of queries related to spinalioma[28] together with $N = 15$
569   synonyms and alternative spellings. She then entered all initial terms into Google Keyword
570   Planner and collected keyword ideas. For all keyword ideas, Alby collected Google results and
571   analyzed them (e.g., regarding their information quality). We repeated Alby's approach by using
572   Google Ads API and generated $N = 1,616$ keyword ideas with only $N = 108$ (7%) duplicates for
573   "spinalioma" together with the synonyms and alternative spellings. Thus, due to the low
574   duplicate rate, it is worthwhile to include synonyms and alternative spellings in the initial list of
575   terms since many new keyword ideas can be generated.
576

577   **Discussion**
578   This paper describes an approach to sample queries from commercial search engines using
579   keyword research tools. First, an initial list of terms is selected and keyword ideas are generated
580   (round one of collecting keyword ideas) from this list. These keyword ideas then serve as seed
581   terms, that is, the starting point, for the second round to collect more keyword ideas. This
582   procedure is repeated in further rounds until no new keyword ideas are found and, therefore, the
583   initial list is saturated. The seed terms used to generate keyword ideas in round one are the
584   researcher's predefined terms, whereas from round two onwards, the seed terms are the keyword
585   ideas received in the preceding round.
586   We empirically tested our approach by using the queries of two published studies as initial terms.
587   Study one is about information quality regarding cancer diet (Herth et al., 2016), and study two
588   is on SEO for COVID-19 and radical right topics (Lewandowski, Sünkler & Yagci, 2021). The
589   number of queries and the total search volume covered could be significantly expanded when
590   comparing the original studies with our query collection approach. After three rounds of
591   collecting keyword ideas, no more new ideas were generated. Two rounds were sufficient to
592   cover most of the total generated keyword ideas and search volume. Hence, both studies would
593   have benefited from our approach, as the foundation of the studies, that is, the search results
594   analyzed, would have been more consistent with what users really search for and see on the web.
595   Our approach considers the popularity of the queries and allows to cover a self-selected topic
596   comprehensively. Both are advantages over other approaches for generating query sets identified
597   in the literature. Previous approaches either consider the popularity of the queries only to a
598   limited extent, for example, by extracting content from online communities, or do not or only
599   partially allow for full thematic coverage, for example, when using queries developed by
600   researchers. Both prerequisites, considering query popularity and allowing topic coverage, are
601   also met by queries delivered by search engine providers, which, however, are only made
602   available to few researchers for specific purposes.

---

[28] The German term used in the study is "Spinaliom."

603 The described approach and its testing come with several limitations. Firstly, when choosing the
604 keyword research tool, it should be noted that a dependency on a provider, such as Google,
605 arises. This dependency is also reflected in the required application for API access. If access is
606 not granted or withdrawn, the implementation of our approach in its current form is no longer
607 possible. Second, the studies we used to test the approach illustrate only a fraction of the possible
608 use cases. Third, a content analysis of the generated keyword ideas has yet to be performed.
609 These limitations point to the need for future studies. First, the approach should be conducted
610 with other keyword research tools, and the results should be compared; this would make the
611 results more reliable. Moreover, this could counteract the dependency on one provider. Second,
612 the approach should be tested on other topics and with more extensive initial lists of terms to
613 check the applicability and scalability beyond the replicated studies. Third, an analysis should be
614 conducted to identify possible topic drifts for the generated keyword ideas, for example, through
615 human evaluators. Topic drifts could occur if the topic of the initial terms is no longer reflected
616 at a certain point, for example, after a particular round of generating keyword ideas (Hobbs,
617 1990). Topic drifts must be identified to exclude affected keyword ideas. Fourth, the effects of
618 expanded query sets on study outcomes should be examined. This is particularly important for
619 studies that aim to make statements about the quality of information a user is confronted with
620 when searching, for example, for health-related topics. The study results may also change when
621 the number of examined queries grows. Finally, it needs to be discussed how to deal with the so-
622 called close variants, which are output by Google Keyword Planning services (see Section
623 "*Study on cancer diet*"). As close variants lead to an unrealistically high total search volume of
624 the query set, they will likely have to be excluded.
625 The approach described in this paper for generating a query set has many possible applications
626 since it can be applied flexibly to different topics and is relatively straightforward to implement.
627 Studies investigating search results related to social phenomena would particularly benefit from
628 the approach. The search interest of real users is covered by systematically obtaining keyword
629 ideas for an initial list of terms. When researchers retrieve and analyze search results on this
630 basis, the search results are more likely to correspond to those seen by real search engine users
631 than if queries without user reference, such as queries developed by researchers or using only
632 technical terms, had been examined. This is especially relevant when statements about social
633 phenomena are to be made, for example, when examining the quality of patient information on
634 the Internet. Otherwise, authors risk analyzing search results that users do not see with their
635 queries, limiting the reliability of their study results.
636

Conclusion

638 In this paper, we described an approach to sample queries from commercial search engines using
639 Google Keyword Planning services. We empirically tested our approach with two published
640 studies on the quality of patient information and SEO. The results show that the number of
641 queries and total search volume could be significantly expanded. When comparing the *number of*
642 *queries* of the original studies with the queries generated by our approach, the original studies

643 only cover 1% (cancer diet), 29% (radical right), and 4% (COVID-19) of the queries generated
644 by our approach. The same holds for the *total search volume*. The search volume of the queries
645 of the original studies only covers 12% (cancer diet), 16% (radical right), and 35% (COVID-19)
646 of the search volume generated by our approach. This leads us to the conclusion that the studies
647 would have benefited from our approach since the queries generated by our approach better
648 reflect actual user behavior. In general, we found that researchers can improve the evidential
649 value of studies that use search queries by extending their initial query set by using our approach.
650 Thus, the approach offers a wide range of applications for studies that seek to draw conclusions
651 about social phenomena using search engine queries. The approach is relatively easy to apply to
652 different topics and use cases. Future research should test the approach with other keyword
653 research tools and topics and conduct content analyses of the generated keyword ideas.
654

655 **References**

656 Alby A. 2020. Muster und Limitationen der Internet-basierten Selbstdiagnose bei häufigen

657   Dermatosen. Christian-Albrechts-Universität zu Kiel.

658 Azizan A, Bakar ZA, Rahman NA. 2019. Construction of Durian Dataset from Web Collection for

659   Query Reformulation Research. *International Journal of Recent Technology and*

660   *Engineering* 8:630–634. DOI: 10.35940/ijrte.B1098.0982S1119.

661 Azzopardi L, White RW, Thomas P, Craswell N. 2020. Data-Driven Evaluation Metrics for

662   Heterogeneous Search Engine Result Pages. In: *Proceedings of the 2020 Conference*

663   *on Human Information Interaction and Retrieval*. Vancouver BC Canada: ACM, 213–

664   222. DOI: 10.1145/3343413.3377959.

665 Bailey P, Moffat A, Scholer F, Thomas P. 2016. UQV100: A Test Collection with Query

666   Variability. In: *Proceedings of the 39th International ACM SIGIR conference on*

667   *Research and Development in Information Retrieval*. Pisa Italy: ACM, 725–728. DOI:

668   10.1145/2911451.2914671.

669 Ballatore A. 2015. Google chemtrails: A methodology to analyze topic representation in search

670   engine results. *First Monday* 20:1–17. DOI: 10.5210/fm.v20i7.5597.

671 Bar-Yossef Z, Gurevich M. 2008. Mining search engine query logs via suggestion sampling.

672   *Proceedings of the VLDB Endowment* 1:54–65. DOI: 10.14778/1453856.1453868.

673   Baytin A, Galinskaya I, Panina M, Serdyukov P. 2013. Speller performance prediction for query

674   autocorrection. In: *Proceedings of the 22nd ACM international conference on*

675   *Conference on information & knowledge management - CIKM '13*. New York, New York,

676   USA: ACM Press, 1821–1824. DOI: 10.1145/2505515.2507871.

677   Bilal D, Ellis R. 2011. Evaluating Leading Web Search Engines on Children's Queries. In: Jacko

678   JA ed. *Human-Computer Interaction. Users and Applications*. Lecture Notes in

679   Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 549–558. DOI:

680   10.1007/978-3-642-21619-0_67.

681   Central. 2015. *Praxis Dr. Internet: Studie zum Krankheitssuchverhalten in Deutschland sowie*

682   *zur Qualität von Gesundheitsinformationen im Internet*.

683   Craswell N, Mitra B, Yilmaz E, Campos D, Lin J. 2021. Overview of the TREC 2021 Deep

684   Learning Track. :16.

685   D'Ambrosio A, Agricola E, Russo L, Gesualdo F, Pandolfi E, Bortolus R, Castellani C, Lalatta F,

686   Mastroiacovo P, Tozzi AE. 2015. Web-Based Surveillance of Public Information Needs

687   for Informing Preconception Interventions. *PLOS ONE* 10:e0122551. DOI:

688   10.1371/journal.pone.0122551.

689   Dang V, Kumaran G, Troy A. 2012. Domain dependent query reformulation for web search. In:

690   *Proceedings of the 21st ACM international conference on Information and knowledge*

691   *management - CIKM '12*. New York, New York, USA: ACM Press, 1045. DOI:

692   10.1145/2396761.2398401.

693   Das A, Shrivastava S, Agrawal P, Sahoo S, Chinnakotla M. 2017. Discovery and promotion of

694   subtopic level high quality domains for programming queries in web search. In:

695   *Proceedings of the first International Workshop on LEARning Next gEneration Rankers,*

696   *Amsterdam, October 1, 2017 (LEARNER 2017)*. 5.

697   Fu H. 2017. Query Reformulation Patterns of Mixed Language Queries in Different Search

698   Intents. In: *Proceedings of the 2017 Conference on Conference Human Information*

699   *Interaction and Retrieval*. Oslo Norway: ACM, 249–252. DOI:

700   10.1145/3020165.3022126.

701   Fumagalli M, Bailoni T, Giunchiglia F. 2020. Assessing ontologies usage likelihood via search

702   trends. *CEUR Workshop Proceedings* 2708.

703   Goel S, Broder A, Gabrilovich E, Pang B. 2010. Anatomy of the long tail. In: Davison BD, Suel

704   T, Craswell N, Liu B eds. *Proceedings of the third ACM international conference on Web*

705   *search and data mining - WSDM '10*. New York, New York, USA, New York, USA: ACM

706   Press, 201. DOI: 10.1145/1718487.1718513.

707   Gravano L, Hatzivassiloglou V, Lichtenstein R. 2003. Categorizing web queries according to

708   geographical locality. *International Conference on Information and Knowledge*

709   *Management, Proceedings*:325–333. DOI: 10.1145/956863.956925.

710   Haak F, Schaer P. 2022. Auditing Search Query Suggestion Bias Through Recursive Algorithm

711   Interrogation. In: *14th ACM Web Science Conference 2022*. Barcelona Spain: ACM,

712   219–227. DOI: 10.1145/3501247.3531567.

713   Haider J. 2016. The structuring of information through search: sorting waste with Google. *Aslib*

714   *Journal of Information Management* 68:390–406. DOI: 10.1108/AJIM-12-2015-0189.

715   Harman DK, Voorhees EM. 2007. TREC: An overview. *Annual Review of Information Science*

716   *and Technology* 40:113–155. DOI: 10.1002/aris.1440400111.

717   Herth N, Kuenzel U, Liebl P, Keinki C, Zell J, Huebner J. 2016. Internet Information for Patients

718   on Cancer Diets - an Analysis of German Websites. *Oncology Research and Treatment*

719   39:273–281. DOI: 10.1159/000445861.

720   Hiemstra D. 2020. Reducing Misinformation in Query Autocompletions. In: *2nd International*

721   *Symposium on Open Search Technology, 12-14 October 2020, CERN, Geneva,*

722   *Switzerland*. DOI: 10.48550/arXiv.2007.02620.

723   Hinz K, Sünkler S, Lewandowski D. 2023. SEO im Wahlkampf: Welche Kandidierende durch

724   Suchmaschinenoptimierung ihre Sichtbarkeit zu erhöhen versuchen. In: Korte K-R,

725   Schiffers M, von Schuckmann A, Plümer S eds. *Die Bundestagswahl 2021*. Wiesbaden:

726   Springer Fachmedien Wiesbaden, 1–28. DOI: 10.1007/978-3-658-35758-0_19-1.

727 Hobbs JR. 1990. Topic drift. In: Dorval B ed. *Conversational organization and its development*.

728   Norwood, NJ: Ablex, 3–22.

729 Hussain A, Gul S, Shah TA, Shueb S. 2019. Retrieval effectiveness of image search engines.

730   *Electronic Library* 37:173–184. DOI: 10.1108/EL-07-2018-0142.

731 Jansen BJ, Spink A. 2006. How are we searching the World Wide Web? A comparison of nine

732   search engine transaction logs. *Information Processing & Management* 42:248–263.

733   DOI: 10.1016/j.ipm.2004.10.007.

734 Janssen S, Käsmann L, Fahlbusch FB, Rades D, Vordermark D. 2018. Side effects of

735   radiotherapy in breast cancer patients. *Strahlentherapie und Onkologie* 194:136–142.

736   DOI: 10.1007/s00066-017-1197-7.

737 Kinney KA, Huffman SB, Zhai J. 2008. How evaluator domain expertise affects search result

738   relevance judgments. In: *Proceeding of the 17th ACM conference on Information and*

739   *knowledge mining - CIKM '08*. New York, New York, USA: ACM Press, 591. DOI:

740   10.1145/1458082.1458160.

741 Krafft TD, Gamer M, Zweig KA. 2019. What did you see? A study to measure personalization in

742   Google's search engine. *EPJ Data Science* 8:38. DOI: 10.1140/epjds/s13688-019-0217-

743   5.

744 Leontiadis N, Moore T, Christin N. 2011. Measuring and analyzing search-redirection attacks in

745   the illicit online prescription drug trade. In: *Proceedings of the 20th USENIX Security*

746   *Symposium*. 281–297. DOI: 10.5555/2028067.

747 Lewandowski D. 2015. Evaluating the retrieval effectiveness of web search engines using a

748   representative query sample. *Journal of the Association for Information Science and*

749   *Technology* 66:1763–1775. DOI: 10.1002/asi.23304.

750 Lewandowski D, Sünkler S, Schultheiß S. 2020. Studies on Search: Designing Meaningful IIR

751   Studies on Commercial Search Engines. *Datenbank-Spektrum* 20:5–15. DOI:

752   10.1007/s13222-020-00331-1.

753 Lewandowski D, Sünkler S, Yagci N. 2021. The influence of search engine optimization on

754   Google's results. In: *13th ACM Web Science Conference 2021*. New York, NY, USA:

755   ACM, 12–20. DOI: 10.1145/3447535.3462479.

756 Liu X, Fang H, Cai D. 2015. Towards Less Biased Web Search. In: *Proceedings of the 2015*

757   *International Conference on The Theory of Information Retrieval*. New York, NY, USA:

758   ACM, 373–376. DOI: 10.1145/2808194.2809476.

759 Lucchese C, Orlando S, Perego R, Silvestri F, Tolomei G. 2013. Discovering tasks from search

760   engine query logs. *ACM Transactions on Information Systems* 31:1–43. DOI:

761   10.1145/2493175.2493179.

762 Mazzeo V, Rapisarda A, Giuffrida G. 2021. Detection of Fake News on COVID-19 on Web

763   Search Engines. *Frontiers in Physics* 9:1–14. DOI: 10.3389/fphy.2021.685730.

764 McCreadie R, Macdonald C, Ounis I, Giles J, Jabr F. 2012. An examination of content farms in

765   web search using crowdsourcing. In: *Proceedings of the 21st ACM international*

766   *conference on Information and knowledge management - CIKM '12*. New York, New

767   York, USA: ACM Press, 2551. DOI: 10.1145/2396761.2398689.

768 Reber M, Krafft TD, Krafft R, Zweig KA, Couturier A. 2020. Data Donations for Mapping Risk in

769   Google Search of Health Queries: A case study of unproven stem cell treatments in

770   SEM. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*:2985–

771   2992. DOI: 10.1109/SSCI47803.2020.9308420.

772 Reimer JH, Schmidt S, Fröbe M, Gienapp L, Scells H, Stein B, Hagen M, Potthast M. 2023. The

773   Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search

774   Engines from 25 Years of Web Archives.

775   Schultheiß S, Lewandowski D. 2021. How users' knowledge of advertisements influences their

776   viewing and selection behavior in search engines. *Journal of the Association for*

777   *Information Science and Technology* 72:285–301. DOI: 10.1002/asi.24410.

778   Schultheiß S, Sünkler S, Lewandowski D. 2018. We still trust in Google, but less than 10 years

779   ago: an eye-tracking study. *Information Research* 23:paper 799.

780   Schweitzer NJ. 2008. Wikipedia and Psychology: Coverage of Concepts and Its Use by

781   Undergraduate Students. *Teaching of Psychology* 35:81–85. DOI:

782   10.1080/00986280802004594.

783   Semenza DC, Bernau JA. 2022. Information-seeking in the Wake of Tragedy: An Examination

784   of Public Response to Mass Shootings Using Google Search Data. *Sociological*

785   *Perspectives* 65:216–233. DOI: 10.1177/0731121420964785.

786   Singer G, Norbisrath U, Lewandowski D. 2012. Ordinary search engine users assessing

787   difficulty, effort, and outcome for simple and complex search tasks. In: *Proceedings of*

788   *the 4th Information Interaction in Context Symposium on - IIIX '12*. New York, New York,

789   USA: ACM Press, 110–119. DOI: 10.1145/2362724.2362746.

790   Soldaini L, Goharian N. 2017. Learning to Rank for Consumer Health Search: A Semantic

791   Approach. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in*

792   *Artificial Intelligence and Lecture Notes in Bioinformatics)*. 640–646. DOI: 10.1007/978-

793   3-319-56608-5_60.

794   StatCounter. 2023a.Search Engine Market Share United States Of America | StatCounter

795   Global Stats. *Available at https://gs.statcounter.com/search-engine-market-*

796   *share/all/united-states-of-america* (accessed February 22, 2023).

797   StatCounter. 2023b.Search Engine Market Share Europe | StatCounter Global Stats. *Available*

798   *at https://gs.statcounter.com/search-engine-market-share/all/europe* (accessed February

799   22, 2023).

800   Strauss AL, Corbin JM. 1998. *Basics of qualitative research: techniques and procedures for*

801   *developing grounded theory*. Thousand Oaks: Sage Publications.

802   Tana J. 2018. An infodemiological study using search engine query data to explore the temporal

803   variations of depression in Finland. *Finnish Journal of eHealth and eWelfare* 10:133–

804   142. DOI: 10.23996/fjhw.60778.

805   Torres G, Rogers R. 2020. Political news in search engines. In: *The Politics of Social Media*

806   *Manipulation*. Nieuwe Prinsengracht 89 1018 VR Amsterdam Nederland: Amsterdam

807   University Press,. DOI: 10.5117/9789463724838_ch03.

808   Trielli D, Diakopoulos N. 2022. Partisan search behavior and Google results in the 2018 U.S.

809   midterm elections. *Information, Communication & Society* 25:145–161. DOI:

810   10.1080/1369118X.2020.1764605.

811   Unkel J, Haim M. 2021. Googling Politics: Parties, Sources, and Issue Ownerships on Google in

812   the 2017 German Federal Election Campaign. *Social Science Computer Review*

813   39:844–861. DOI: 10.1177/0894439319881634.

814   Verma M, Yilmaz E. 2016. Category Oriented Task Extraction. In: *Proceedings of the 2016 ACM*

815   *on Conference on Human Information Interaction and Retrieval*. Carrboro North Carolina

816   USA: ACM, 333–336. DOI: 10.1145/2854946.2854997.

817   Waller V. 2011. Not just information: Who searches for what on the search engine Google?

818   *Journal of the American Society for Information Science and Technology* 62:761–775.

819   DOI: 10.1002/asi.21492.

820   Whiting S, Jose JM, Alonso O. 2016. SOGOU-2012-CRAWL. In: *Proceedings of the 39th*

821   *International ACM SIGIR conference on Research and Development in Information*

822   *Retrieval*. New York, NY, USA: ACM, 709–712. DOI: 10.1145/2911451.2914668.

823   Wu W, Meng W, Su W, Zhou G, Chiang Y-Y. 2016. Q2P. *ACM Transactions on the Web* 10:1–

824   29. DOI: 10.1145/2873061.

825   Xu Z. 2019. Personal stories matter: topic evolution and popularity among pro- and anti-vaccine

826   online articles. *Journal of Computational Social Science* 2:207–220. DOI:

827   10.1007/s42001-019-00044-w.

828   Yilmaz T, Ozcan R, Altingovde IS, Ulusoy Ö. 2019. Improving educational web search for

829   question-like queries through subject classification. *Information Processing &*

830   *Management* 56:228–246. DOI: 10.1016/j.ipm.2018.10.013.

831   Zhang Y. 2012. The impact of task complexity on people's mental models of MedlinePlus.

832   *Information Processing & Management* 48:107–119. DOI: 10.1016/j.ipm.2011.02.007.

833   Zuccon G, Palotti J, Goeuriot L, Kelly L, Lupu M, Pecina P, Müller H, Budaher J, Deacon A.

834   2016. The IR Task at the CLEF ehealth evaluation lab 2016: User-centred health

835   information retrieval. *CEUR Workshop Proceedings* 1609:15–27.

836

Figure 1

Query sampler: an approach for generating query sets

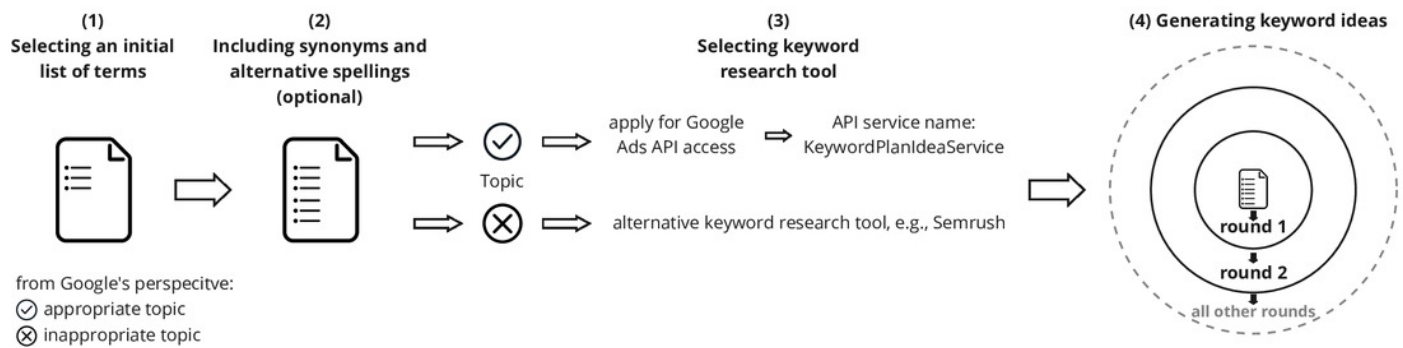



Figure 2

Keyword ideas of all studies and rounds (cumulative)

	Round 1	Round 2	Round 3	Round 4	Round 5
—— Cancer diet	10.2%	78.6%	99.0%	100.0%	100.0%
·········· COVID-19	77.1%	100.0%	100.0%	100.0%	100.0%
− − − Radical right	58.3%	89.9%	98.2%	100.0%	100.0%

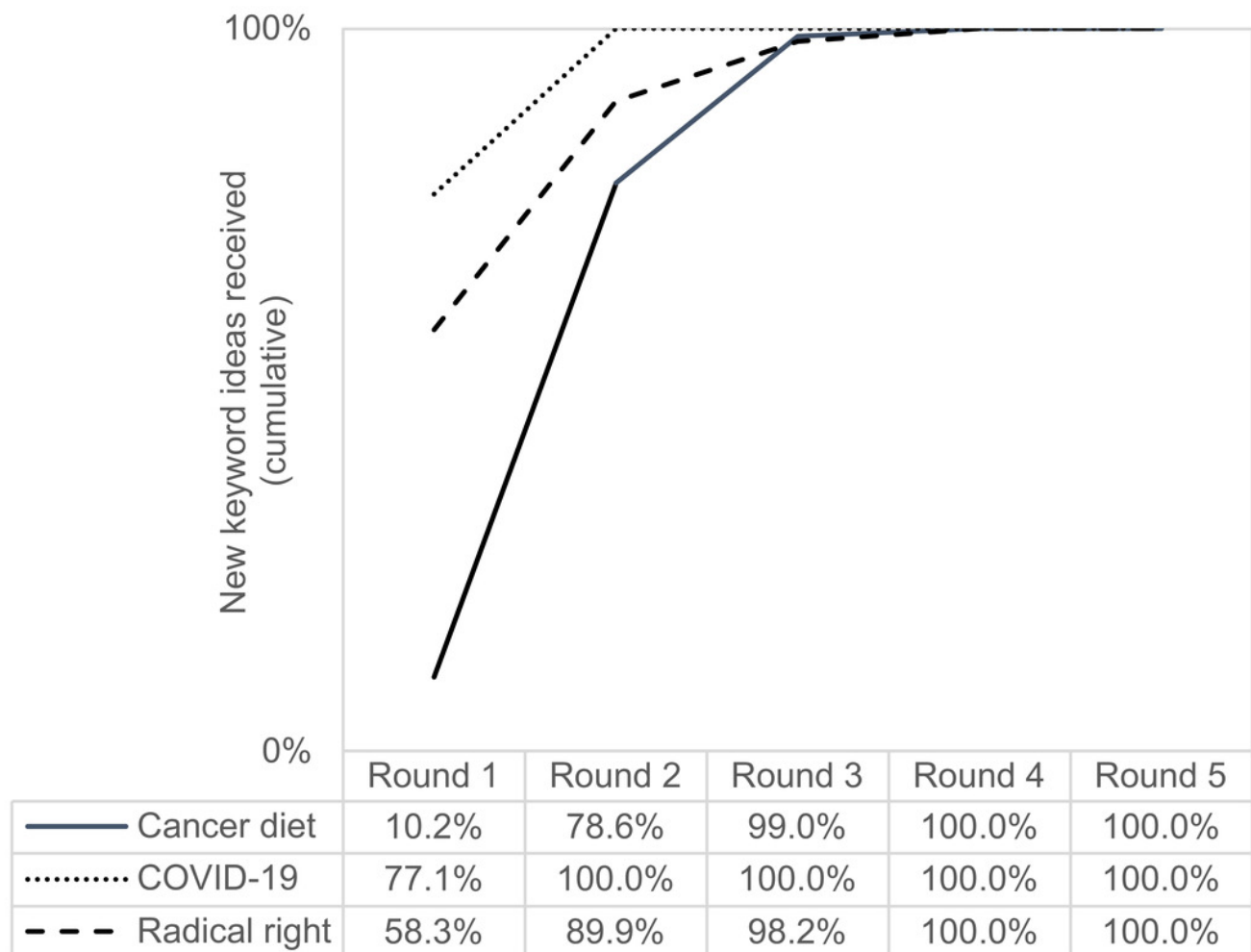
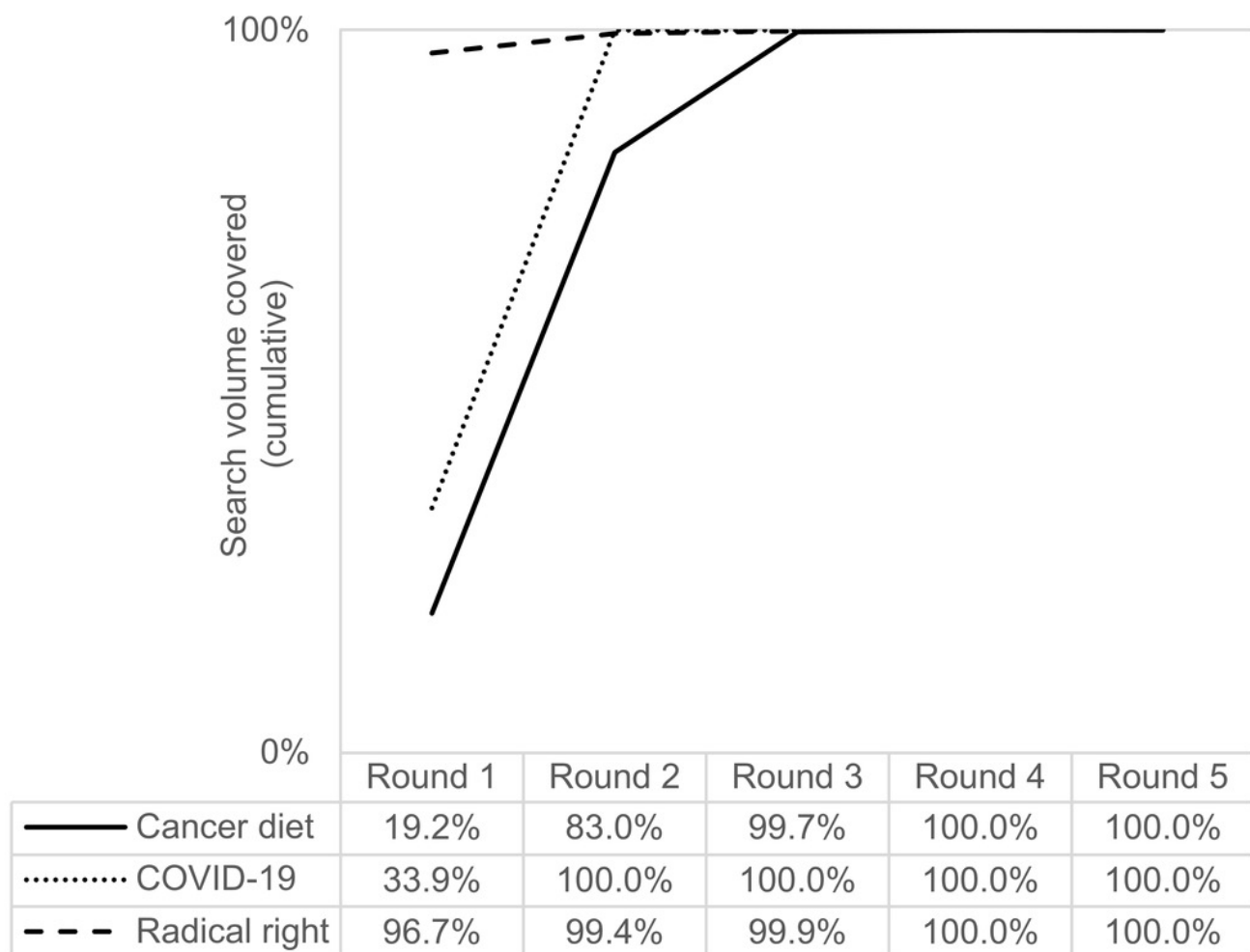



Figure 3

Search volume of all studies and rounds (cumulative)

	Round 1	Round 2	Round 3	Round 4	Round 5
Cancer diet	19.2%	83.0%	99.7%	100.0%	100.0%
COVID-19	33.9%	100.0%	100.0%	100.0%	100.0%
Radical right	96.7%	99.4%	99.9%	100.0%	100.0%

Approaches for generating a query set

[a] The sizes only refer to those in cited studies.

[b] This approach was not identified in the literature, but it represents the best imaginable, albeit unrealistic, fulfilment of both criteria.

[c] This only applies if the data come from a popular search engine (e.g., Google or Bing), assuming that data from this search engine are representative of general search behavior.

Table 1: Approaches for generating a query set

	Criteria		Maximum number of queries[a]
Approach	Popularity consideration	Topic coverage	
Queries of all search engines worldwide[b]	Popularity of topics is considered *across search engines*	All conceivable topics are considered, *independent* of specific search engines	/
Queries delivered by a search engine provider	Popularity of topics is considered within a specific search engine[c]	All conceivable topics searched for in the respective search engine are considered	2,6 B (Goel et al., 2010)
Query sampler, combining the approaches "popularity data" and "predetermined lists of terms" (our approach)	*Popularity is considered through data on search volume using the Google Ads API*	*Most topics are covered (see Section "Selecting a keyword research tool" for information on restrictions regarding search volume data for specific topics)*	*Potentially unlimited*
Popularity data	Popularity is considered through data on search volume	Most topics are covered (see Section "Selecting a keyword research tool" for information on restrictions regarding search volume data for specific topics)	29,132 (D'Ambrosio et al., 2015)
Autocomplete suggestions	Popularity is considered, but no data on search volume are given	Most topics are covered (see Section "Autocomplete suggestions" for information on restrictions regarding autocompletion for specific topics)	21,407 (Haak & Schaer, 2022)
Content extracted from online communities	Popularity is only reflected within the respective online community	Only topics that are discussed in the online community are covered	10,717 (Yilmaz et al., 2019)
Queries provided by subjects	The queries are not created in a natural environment (e.g., within an online forum or by crowdsourcing)	Theoretically, an unlimited coverage can be achieved	5,764 (Bailey et al., 2016)
Queries developed by the study authors	Popularity is not considered	Queries are mostly arbitrarily arranged without specific topics being covered in depth	50 (McCreadie et al., 2012)
Predetermined lists of terms	Depends on the list (see Section "Predetermined lists of terms")	Depends on the list (see Section "Predetermined lists of terms")	6,211 (Hinz, Sünkler & Lewandowski, 2023)

[a] The sizes only refer to those in cited studies.

[b] This approach was not identified in the literature, but it represents the best imaginable, albeit unrealistic, fulfilment of both criteria.

[c] This only applies if the data come from a popular search engine (e.g., Google or Bing), assuming that data from this search engine are representative of general search behavior.

3
4

Generating keyword ideas for study on cancer diet

Table 2: Generating keyword ideas for study on cancer diet

| | Seed terms (N) | Keyword ideas (N) | Exclusion of keyword ideas | | | Keyword ideas: remaining N (%) | Search volume of remaining keyword ideas (mean) | Search volume of remaining keyword ideas (Sum) |
			Search volume of 0 N (%)	Duplicates within round N (%)	Duplicates with already existing keyword ideas N (%)			
Round 1	1	10	0	0	0	10 (100%)	41	412
Round 2	10	105	0	28 (27%)	10 (10%)	67 (64%)	20	1,368
Round 3	67	741	0	645 (87%)	76 (10%)	20 (3%)	18	358
Round 4	20	335	0	275 (82%)	59 (18%)	1 (0.3%)	6	6
Round 5	1	1	0	0	1	0	0	0
Sum						98		2,144

Generating keyword ideas for COVID-19 queries

Table 3: Generating keyword ideas for COVID-19 queries

| | Seed terms (N) | Keyword ideas (N) | Exclusion of keyword ideas | | | Keyword ideas: remaining N (%) | Search volume of remaining keyword ideas (mean) | Search volume of remaining keyword ideas (Sum) |
			Search volume of 0 N (%)	Duplicates within round N (%)	Duplicates with already existing keyword ideas N (%)			
Round 1	15	307	5 (2%)	0	5 (2%)	297 (97%)	4,775	1,413,482
Round 2	297	2,970	79 (3%)	2,205 (84%)	301 (10%)	88 (3%)	31,330	2,757,032
Round 3	88	35	24 (68%)	3 (9%)	8 (23%)	0		
Sum						385		4,170,514

Generating keyword ideas for radical right queries

Table 4: Generating keyword ideas for radical right queries

| | Seed terms (N) | Keyword ideas (N) | Exclusion of keyword ideas | | | Keyword ideas: remaining N (%) | Search volume of remaining keyword ideas (mean) | Search volume of remaining keyword ideas (Sum) |
			Search volume of 0 N (%)	Duplicates within round N (%)	Duplicates with already existing keyword ideas N (%)			
Round 1	82	233	37 (16%)	2 (1%)	32 (14%)	162 (70%)	993	160,819
Round 2	162	854	26 (3%)	575 (67%)	165 (19%)	88 (10%)	51	4,486
Round 3	88	1,194	24 (2%)	1,020 (85%)	127 (11%)	23 (2%)	33	760
Round 4	23	71	0	30 (42%)	36 (51%)	5 (7%)	32	162
Round 5	5	0	0	0	0	0	0	0
Sum						278		166,227