Query Refinement in Dataset Search

Helen Borchart

Supervised by Professor Haiyan Jia and Professor Brian Davison

Introduction

As the availability of data on the web continues to grow, the tools used to access such data becomes increasingly important. While these tools may be designed with the goal of helping people discover data on the web, the results are not always satisfactory (Vincent, 2018). It is important to consider how people are modifying their queries in order to better understand their searching behaviors. This is known as query refinement, which is defined as when a user "modif[ies] a previous search query in hope of retrieving better results" (Huang, 2009).

Through a general query log analysis in dataset search on a website for collaborative open data community and catalog, it became clear that query refinement was an important aspect of dataset search. Despite variances in topics, query length, number of results returned per query, etc., users were repeatedly issuing multiple queries in the same session. This is indicative of users' interest in seeing more search results from the system based on the repeated modification of queries.

It appears that query refinement is a major challenge in dataset search, especially for non-expert users when conducting cross-domain dataset searches. Evaluating results that are given from the query could be difficult for non-expert users in particular and they could be modifying their queries in hopes of receiving better results, hence the multiple queries being issued in the log data that was examined. A series of think-aloud interviews with non-expert have revealed more struggles with query refinement than coming up with initial queries. Using the state-of-art dataset search engines and repositories, the participants experienced difficulty with determining the contents of the dataset and relevance of the dataset to their needs based on the description or preview of the dataset. This leads to the question of what sort of relationship exists between content level preview and dataset search experience. If users are modifying their queries in hopes of receiving better results, what characteristics of their searching experience factor into that and how does that contribute to the dataset search experience as a whole?

RQ: What is the relationship between preview and dataset search experience?

If the current design of dataset search tools cannot suffice user needs for modifying queries toward obtaining satisfactory search results, what alternatives could facilitate this process? This study

proposes to explore the potential of a content-level preview of the datasets in assisting query refinement and achieving optimal search experience.

Literature Review

Previous work on dataset search shows that one of the key processes in dataset search is the refinement of queries (Chapman, 2109; Koesten, Laura M., et al. 2019). Even though some scholars argue that users are less likely to engage in query refinement in dataset search than in general web search (Koesten et al., 2019), studies have shown empirical data indicating that they would refine their queries several times. For instance, a study that examines dataset search query log data from two data portals in the U.K. shows that refinement was recorded in 22.77% of the sessions for one of the portals and 36.08% for the other (Kacprzak et al., 2018). This number may not be higher because users do not have faith in the search engine to give them satisfactory results (Kacprzak et al., 2018). The lack of faith in the search engine might influence these users to modify their search, as query refinement is the process of changing a query to get more satisfactory results. This could be for a variety of reasons; one reason being whether or not the user does not realize that the results being presented to them or relevant to their dataset needs based on the way in which they are being viewed by the user.

This leads us to ask:

RQ1: What is the relationship between previewing content-level data information and query refinement?

Google has built a search engine specifically for dataset search and in a paper describing the process, the authors comment on how “metadata is often limited and minimalistic, it may not provide enough signal to decide whether a dataset is relevant to the user’s query” (Brickley, et al., 2019). If the metadata is not sufficient for users to make a decision about the relevancy of the dataset, it is then important to consider what features of a dataset could be enough of a signal for users to make informed decisions on the relevancy of a dataset. A paper examining query refinement in a database found that

allowing users to examine the content of the bibliographic database led to users i) finding relevant documents faster and (ii) higher levels of satisfaction with the relevance of documents for their needs (Stojanovic, 2005). If it is true that content preview has lead to higher levels of reported relevance, we hypothesize that:

H1: Content preview in dataset searches will reduce user needs for query refinement.

H1a: Preview of the cells of the dataset will lead to higher levels of certainty of the content;

H1b: Higher levels of certainty in the content will lead to increased confidence in relevance judgment of the dataset results;

H1c: Higher levels of confidence in the relevance judgment will lower the need for query refinement.

Another paper exploring dataset search also emphasized the quality of relevance when considering dataset search experience. In the paper examining facets of dataset search that users found useful done by Koesten et. al, they found that "specific relevance, usability and quality aspects were perceived to be different for data than for documents - for example, the methodology used to collect and clean the data, missing values, the granularity of the captured information" (Koesten, 2020). Similarly to the paper exploring bibliography database search, one way of providing users a way of assessing granularity of data could be "achieved by providing visual or textual indicators of these aspects on the interface" (Koesten, 2017). The suggestion of using the preview of data to aid dataset search is seen again when discussing relevance. Koesten argues, "To support relevance assessments, we recommend also displaying information about the granularity of the data. One approach would be to display headers, summarising statistics or previews of the data, all of which could be provided alongside the search results." (Koesten, 2017).

This leads to the second hypothesis:

H2: Higher level of perceived relevance of the search results will lead to lower levels of need for query refinement.

The feature of previewing data is not only valuable for determining the relevance and granularity of the data but is important for understanding datasets as a whole. When searching for datasets, it is

important that users understand what they are looking at to understand what they are looking for. Koesten argues this is achieved by “visualizing datasets side by side to facilitate understanding” (Koesten, 2017). In another paper examining data, a proposed method for facilitating understanding was creating interfaces that showed data that was, “displayed graphically, so the user can view the information from different perspectives. Users can choose how the data is represented (e.g., a visualisation, spreadsheet, samples), which means that a wider set of people can potentially use the data)” (Marchionini, 2005). Providing users with a preview of the dataset also allows the user the ability to bypass downloading the dataset. In an interview with a participant in a paper with Koesten exploring the trouble with dataset search, the participant says that due to poor support from the interface, they have to download the data to decide whether or not the dataset is very relevant and that, “they often give you a preview of the first few rows and that’s like a nice starting point” (Koesten, 2017).

Another valuable measure in query refinement and user experience is confidence in the results being presented to the user. Confidence and relevance have been correlated in previous research examining searching behavior. One group of researchers defined the link between the two as when a user “determines [the result] to be ‘highly relevant (i.e., a high confidence that the document is the document that the user would be most interested in viewing)” (“Document search engine including highlighting of confident results”). While confidence is linked to relevance, confidence can also be considered in the context more broadly of finding a result that best matches their needs. In a study looking at query refinement in a knowledge base, the researchers defined confidence in this way. They considered the quality of results, marked by, select[ing] relevant results... each participant had to express his confidence in these results. The confidence describes a participant’s sureness that the selected results are the best possible ones (i.e. that there is no better result for his need” (Suresh, 2014). The desire to find the best possible result can be examined through query refinement then.

Thus far, the relationship between dataset search experience and content preview has been examined. For this project, it is also important to review the literature regarding the relationship between query refinement and dataset search and how that could perhaps be a mediating variable in a participant’s

overall dataset search experience. When exploring the variables mentioned before when searching for data, such as relevance and satisfaction, previous research suggests that there is a relationship that does exist between the two. In a paper exploring search behavior and satisfaction, one group of researchers found that users "issued fewer queries when satisfied and more queries when dissatisfied" (Hassan, 2014). In a paper looking specifically at refinement techniques, researchers found that through their experiment when a user's needs are satisfied, they are less likely to refine and that "users refine queries to direct the information retrieval systems into a new result space because they were not completely satisfied with the results for the original query" (Ooi, 2015).

If query refinement is linked to satisfaction, a variable used to measure dataset search experience, we can ask:

RQ2: What is the relationship between query refinement and dataset search experience?

If it is true that fewer queries are issued when satisfied, and satisfaction is one variable when looking at dataset search experience, we then hypothesize that:

H3: Need for query refinement is negatively associated with dataset search experience.

When considering all of these aspects of dataset search, we take into consideration content preview, certainty in content, relevance, and satisfaction. These factors influence the dataset search experience as a whole, but it is the content level preview that we hypothesize and have reason to believe is contributing positively to those factors. Therefore, we hypothesize:

H4: Content preview will lead to enhanced dataset search experience.

Research Questions & Hypotheses

The most overarching and first research question being asked in this study is

RQ: What is the relationship between preview and dataset search experience?

This question came from two previously stated ideas that we already know from research that has already been conducted: i) query refinement is a part of the dataset search process ii) users will exit out of a search session because metadata is not sufficient enough for them to believe they found a relevant enough match. However, some datasets do allow users to preview cells of the dataset they are looking at. If users were presented with only a textual description of the dataset versus being presented with potentially relevant cells, how or would that change their searching behavior and correlated query refinements?

After exploring previous literature, this research question can be broken down into two different questions to try and answer this.

RQ1: What is the relationship between previewing content-level data information and query refinement?

If a preview of the data truly facilitates a better understanding of the dataset, how will that affect the way in which people refine their queries? Since these participants could have a better understanding of the data, as previous literature mentioned, confidence could lead participants to the idea that they have selected the best possible dataset for their needs (Suresh, 2014). Therefore, the need to refine the query would be lower.

This leads to the first set of hypotheses:

H1: Content preview in dataset searches will reduce user needs for query refinement.

H1a: Preview of the cells of the dataset will lead to higher levels of certainty of the content;

H1b: Higher levels of certainty in the content will lead to increased confidence in relevance judgment of the dataset results;

H1c: Higher levels of confidence in the relevance judgment will lower the need for query refinement.

Additionally, we know that relevance is an aspect users consider when evaluating datasets, which could lead to a lower need to refine the query. This leads to the second hypothesis:

H2: Higher level of perceived relevance of the search results will lead to lower levels of need for query refinement.

To further examine the relationship between query refinement and dataset search, we know from the literature that users satisfied with results will issue fewer queries. If a user is more satisfied with results and we know they issue fewer queries, then there is a relationship between dataset search experience and query refinement. This leads to the next hypothesis:

H3: Need for query refinement is negatively associated with dataset search experience.

Considering the variables of satisfaction, relevance, and the way in which they are linked to query refinement and dataset search experience we consider the final factor associated with dataset search that will be examined which is content preview. If it is true that content preview facilitates understanding which leads users to a better-perceived relevance of results and satisfaction query refinement is associated negatively with dataset search experience, users that have a better perceived relevance of results and are more satisfied with the results will have a better dataset search experience. This leads to the final hypothesis:

H4: Content preview will lead to enhanced dataset search experience.

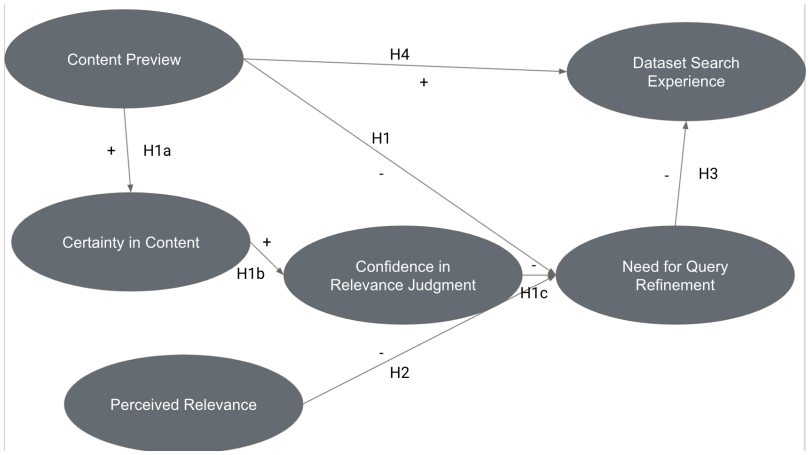


Figure 1. Hypotheses Visualization. Visualizing relationship between hypotheses.

Methods

Participants

An experimental study was conducted with N = 88 participants recruited from Amazon Mechanical Turk. The participants aged from ages 23 years old to 63 years old. The average age of the participants was 39 years old and the median age was 37 years old. 52 (59.1%) people who completed the survey indicated their gender as Male and 36 participants indicated their gender as Female (40.9%). 65 participants identified as White/Caucasian, 15 participants identified as Black or African American, 6 identified as Asian, 2 participants identified as Hispanic, Latinx or Spanish Origin, and 2 participants identified as Native Hawaiin/Pacific Islander. Regarding participants Educational background, 75 participants (85.2%) of participants had indicated having a B.A. or B.S. Degree as their highest level of completed education.

Compensation

Upon completion of the experimental study, participants were paid $2.50 for their participation in the experiment.

Design

Participants were divided into 2 different groups with 4 conditions total. The independent variable was the content level preview of the cells. 44 participants were not shown a content level preview of the cells and 44 participants were shown a content level preview of the cells. In both conditions, participants previewed two different search results, in a randomized order. One group of search results pertained to the query, “Crime in Chicago by neighborhood” and the other set of search results pertained to the query, “COVID-19 in California by county”.

For these conditions, “present” indicates that the feature of dataset content preview was shown to the participant while “absent” indicates that the stimuli shown to the participant did not contain content preview. The number of participants pertaining to the randomized order in which they saw the conditions are presented in the table below. Dependent variables included certainty in content, perceived relevance, confidence in relevance judgment, and need for query refinement, and dataset search experience, which includes a measurement of how satisfied the participant finds the results to be.

Content Preview	Chicago-C2|COVID-C2, and 21 Order of Results	Order of Results
Absent	COVID-C1|Chicago-C1 N = 22	Chicago-C1|COVID-C1 N = 23
Present	COVID-C2|Chicago-C2 N = 21	Chicago-C2|COVID-C2 N = 23

Table 1

Materials

Stimuli. The search result stimuli were constructed on Google Docs and Microsoft PowerPoint. They were not meant to be interactive. In order to mimic a real search result, the phrase "135 datasets were found for the query x". The Website stimuli resembled a web link, with a URL, hyperlink, and a textual description of the dataset. In order to mimic the experience of searching online, words that appeared in the description that was in the query were bolded to highlight the potential relevance of the result. For the dataset prototype, participants were shown the same URL, hyperlink, and description but were also shown a table that mimicked a preview of a dataset. Each dataset was 6 columns and 5 rows. In order to mimic the preview aspect of a dataset, the phrase "5 of 4738", or another arbitrary number, was displayed underneath the dataset. In order to display potentially relevant information, terms that appeared in the query that also appeared in the dataset were highlighted in yellow.

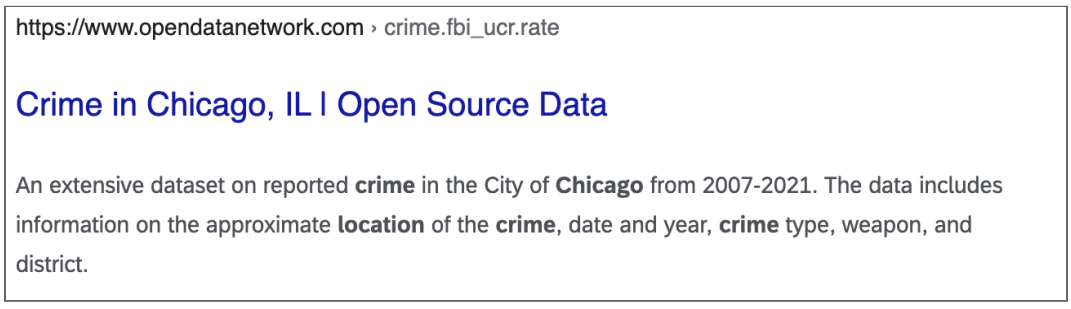
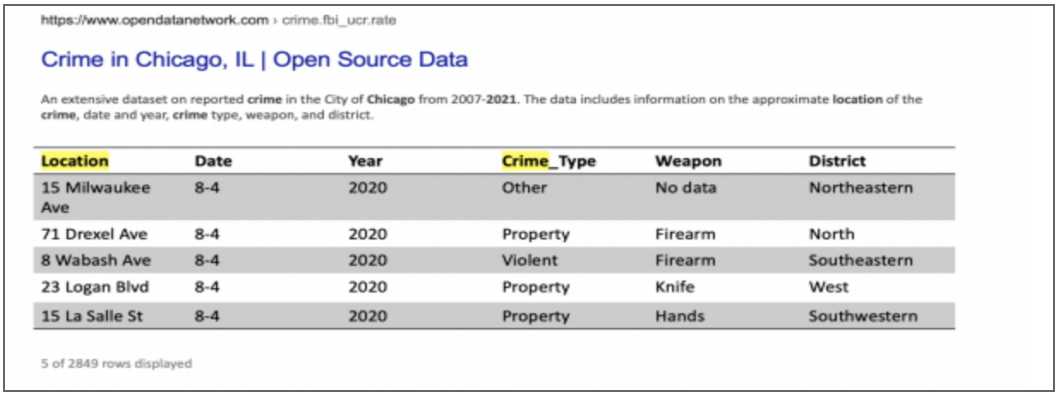


Figure 2. Stimuli Example. Stimuli without content preview.



https://www.opendatanetwork.com › crime.fbi_ucr.rate

Crime in Chicago, IL | Open Source Data

An extensive dataset on reported **crime** in the City of **Chicago** from 2007-**2021**. The data includes information on the approximate **location** of the **crime**, date and year, **crime** type, weapon, and district.

Location	Date	Year	Crime_Type	Weapon	District
15 Milwaukee Ave	8-4	2020	Other	No data	Northeastern
71 Drexel Ave	8-4	2020	Property	Firearm	North
8 Wabash Ave	8-4	2020	Violent	Firearm	Southeastern
23 Logan Blvd	8-4	2020	Property	Knife	West
15 La Salle St	8-4	2020	Property	Hands	Southwestern

5 of 2849 rows displayed

Figure 3. Stimuli Example. Stimuli with content preview.

Survey. In addition to viewing the stimuli, participants completed a pre-test and post-test questionnaire. To measure certainty in content, perceived relevance, confidence in relevance judgement, and need for query refinement participants were asked to answer questions on a 5 point scale measuring certainty, perceived relevance, confidence in relevance judgment and need for query refinement. To measure dataset search experience, participants were asked to rate how much they agreed or disagreed with the following statements. The statements were on a 5 point scale, with 1 being “Strongly Disagree” and 5 being “Strongly Agree”. The questions were as follows:

- The content of the results matches my information needs.
- The search engine provides comprehensive information.
- The search engine provides information that matches my needs.
- I intend to use this system for dataset search if it is available.
- I intend to use this system for work if it is available.

The post-test questionnaire also included an open ended question, “What helped you determine whether the datasets were a good match for your information needs? Please use the space below to write down your thoughts”. Finally, in order to gauge the attention and effectiveness of content preview on recalling results, participants were presented with a set of 3 results. 2 of the results had been shown in the stimuli and the third result had not been seen before. Participants were asked to indicate which result had not been seen before.

The pre-test questionnaire included a series of demographic questions, including a technology fluency scale, and questions about age, gender identity, ethnicity/race, and education level.

Procedure

Participants were directed to the Qualtrics survey through the link Amazon Mechanical Turk. Once completing the informed consent form they read the following instructions:

You will be asked to imagine that you were completing a dataset search task. You will first read a description of your task and then be presented with several search results. Lastly, you will be asked to answer a number of questions regarding this dataset search.

The participants were presented with the stimuli then completed the post-test questionnaire described earlier. Following the completion of the survey, participants were thanked for their participation.

Results

Data analyses, including descriptive statistics, independent sample t-tests, and simple linear regressions were conducted to examine whether the hypotheses were supported.

Descriptive statistics were first conducted to examine the distribution, central tendency, and skewness. Reliability was also assessed using Cronbach's α with variables that were measured using multi-item scales (i.e. Literacy and User Experience). Means, Standard Deviations, medians, skewness, Kurtosis, and Cronbach's α for these continuous variables are reported in Table 2. Results have indicated acceptable levels of reliability of the scales (Cronbach's α close to or above .80).

Variable	Mean	Standard Deviation	Median	Skew	Kurtosis	Cronbach's α
Literacy	3.95	.68	4.08	-1.35	3.58	.79
Confidence	4.02	.66	4	-.02	-.73	
Relevance	3.91	1.05	4	-.65	.14	
ConfidenceinJ	4.11	.9	4	.06	.37	
Satisfaction	3.77	.93	4	-1.23	1.73	

NeedToRefine	3.24	1.3	3	-.26	-1.09	
EasyToRefine	3.85	.98	4	-.80	.58	
Experience	4.01	.69	4	-1.21	2.72	.83
Confidence2	3.84	.9	4	-.83	.89	
Relevance2	3.9	1.15	4	-.39	-.13	
ConfidenceInJ2	4.03	.86	4	-.81	.65	
Satisfaction2	3.7	1.14	4	-.99	.31	
NeedToRefine2	3.45	1.28	4	-.42	-.84	
EasyToRefine2	3.75	1.02	4	-.84	.22	
Experience2	3.88	.92	4	-1.46	3.09	.89

Table 2

In order to examine the effect of content-level preview on outcome variables, as hypothesized in H1, H1a, and H4, independent sample t-tests were conducted. While the complete list of results is included in Appendix 3, statistically significant results are reported as follows:

An independent sample t-test showed a statistically significant effect of content-level preview on perceived relevance of the search results (see Figure 4). Participants who viewed the search results without a content-level preview reported a higher level of perceived relevance (M = 4.18) than those with the preview (M 3.60, SD = 1.45), t (86) = 2.41, $p < 0.05$. An independent sample t-test showed a statistically significant effect of content-level preview on the need to refine the query (see Figure 5). Participants who viewed the search results with a content-level preview had a higher level of needing to refine the query (M = 3.814, SD = 1.27) than those who did not have the preview (M = 3.11), t (86) = -2.67, $p < 0.05$. An independent sample t-test also showed a statistically significant effect of content-level preview on satisfaction of the search results (see Figure 6). Participants who viewed the search results without content-level preview had higher levels of satisfaction with the search results (M = 3.93, SD = 1.13) than those who did have content-level preview (M = 3.46), t (86) = 1.963), $p < 0.05$.

The data for the t-tests that were significant, along with their boxplot, are shown below.

T-Test Result: Effect of Content Preview on Perceived Relevance of Search Results

	t	df	SD	p-value	Mean (Without preview)	Mean (With Preview)
Perceived Relevance	2.41	86	1.45	.01804	4.178	3.604

Table 3

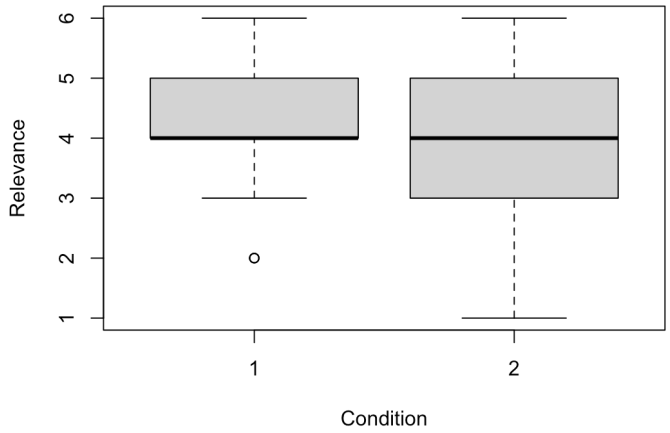


Figure 4. T-test boxplot. Effect of content level preview on perceived relevance of results.

T-Test Result: Effect of Content Preview on the Need to Refine Search Results

	t	df	SD	p-value	Mean (Without preview)	Mean (With preview)
Need to Refine	-2.67	86	1.27	.009048	3.111	3.814

Table 4

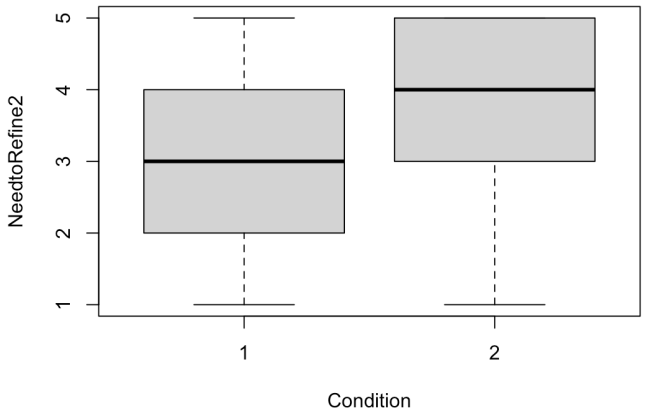


Figure 5. T-test boxplot. Effect of content level preview and the need to refine the query.

T-Test Result: Effect of Content Preview on Satisfaction of Search Results

	t	df	SD	p-value	Mean (Without Preview)	Mean (With Preview)
Satisfaction	1.963	86	1.13	.05278	3.93	3.46

Table 5

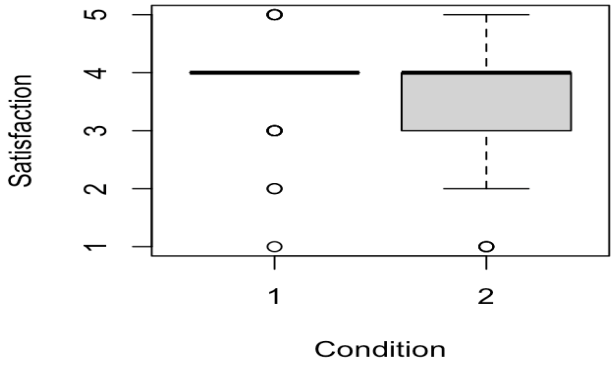


Figure 6. T-test boxplot. Effect of content level preview and satisfaction with search results.

The analysis that was done tested the hypotheses and some were supported. From a literature review done that looked at the ways in which content preview could facilitate understanding, we predicted

that content preview in dataset searches will reduce user needs for query refinement. This hypothesis was correct, as the t-test showed H1 was supported with p-values < .05.

Additionally, the relationship between the condition and the result of the stimuli test was significant. Recognition here is measured with whether participants correctly recognized the interface and/or features that they were exposed to, and can serve as a proxy to their cognitive effort/attention paid to the content. The difference in the means indicates the variance in participants' recognition, which turned out statistically significant only in the second task (Chicago crime rates). Recognition was found higher among participants in the "With Preview" condition than those in the "Without Preview" condition.

In order to test H1b, H1c, H2 and H3, simple linear regression analyses were conducted. A complete list of results is included in Appendix 3, and the statistically significant results are reported below.

A simple linear regression test showed a significant negative effect of Perceived Relevance on Need for Query Refinement, F (1, 86) = 10.47, *Adjusted* R^2 = 0.10, $p < 0.01$. β = -0.30, $p < 0.01$. The result indicates that greater Perceived Relevance is significantly associated with a lower Need for Query Refinement (as shown in Figure 7). It supported H2, which hypothesized a negative relationship between perceived relevance and need for query refinement.

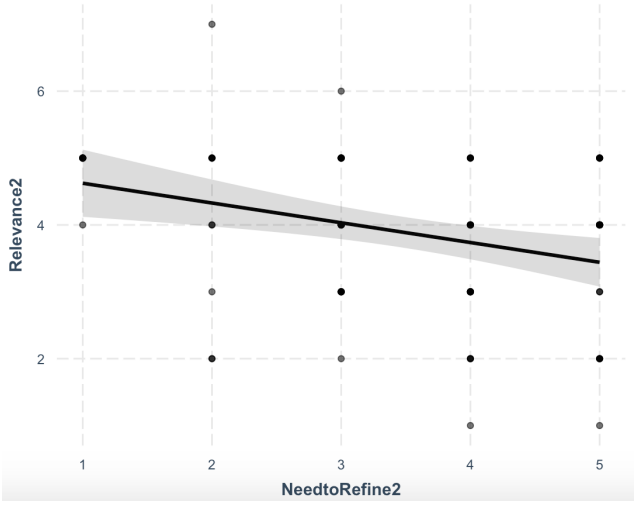


Figure 7. Linear Regression Model. Perceived Relevance & Need for Query Refinement.

A simple linear regression test showed a significant positive effect on Certainty in Content & Confidence in Relevance Judgement, F (1, 86) = 3.88, *Adjusted* R^2 = 0.03, p < 0.01; β = 0.28, p < 0.05. The result indicates that greater Certainty in Content is significantly associated with a higher Confidence in Relevance Judgement (as shown in Figure 8). It supported H1b, which hypothesized that higher levels of certainty in the content will lead to increased confidence in relevance judgment of the dataset results.

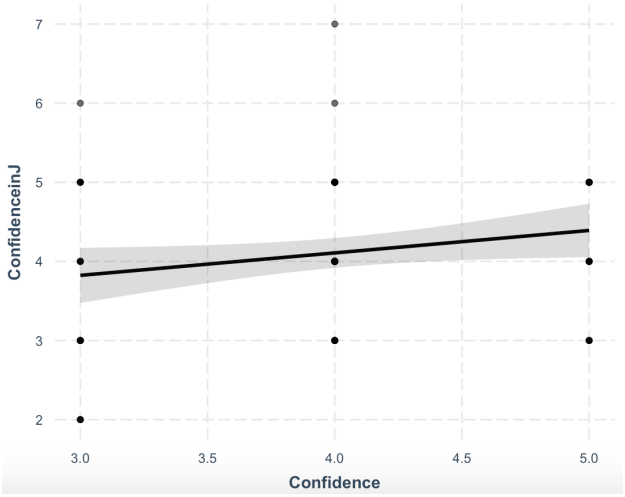


Figure 8. Linear Regression Model. Certainty in Content & Confidence in Relevance Judgement.

A simple linear regression test showed a significant negative effect on Need to Refine & Dataset Search Experience, F(1,86) = 8.67, *Adjusted* R^2 = 0.08, p < 0.01; β = -0.19, p < 0.01. The result indicates that a greater Need to Refine is significantly associated with a worse Dataset Search Experience (as shown in Figure 9). It supported H3, which hypothesized that need for query refinement is negatively associated with dataset search experience.

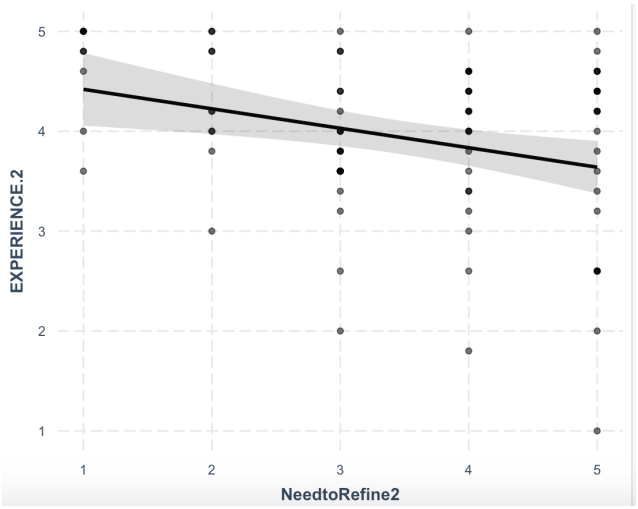


Figure 9. Linear Regression Model. Need to Refine & Dataset Search Experience.

The table below contains data on the regression analyses that were performed but not found to be significant.

Relationship	*F*	*Adjusted* R^2	β
Confidence in Relevant Judgement & Need for Query Refinement	(1,86)	0.008337	-0.09721
Perceived Relevance & Need for Query Refinement	(1,86)	-0.006931	-0.05467

Table 6

While literature suggested that a preview of the cells could facilitate understanding, it was not true that content preview leads to higher levels of certainty in the content. The t-test showed H1 was not supported with p-values both > .05. However, as examined in the literature review, we know that research has indicated there is a link between confidence and relevance when evaluating datasets. H1b predicted higher levels of certainty in the content will lead to increased confidence in relevance judgment of the dataset results and the regression analysis showed that H1b was supported. While literature suggested that confidence in the results was linked to relevance, it was not true that confidence in the relevance judgment was linked to query refinement. H1c predicted that higher levels of confidence in the relevance judgment will lower the need for query refinement and the regression analysis showed H1c was not

supported. This could be for a variety of reasons. First, perhaps it is the content preview that facilitates understanding of results that leads users to feel the need to refine the query. As Koesten stated earlier, content level preview could lead to a better understanding of results.

When looking at the relationship between relevance and the need to refine the query, we hypothesized from the literature that higher levels of perceived relevance of the search results would lead to lower levels of need for query refinement. This was based on literature that examined confidence and relevance when searching for datasets that argued those factors were important when searching for datasets. The regression analysis showed that this relationship between relevance and query refinement was significant. When examining variables from previous literature that were used in dataset search, we know satisfaction is an important aspect. Since users that are satisfied with their results issue fewer queries, we hypothesize that the need for query refinement is negatively associated with dataset search experience. The regression analysis that was on this relationship showed this was supported. However, the connection between the literature that attempted to link content preview and dataset search experience as a whole was not supported. While the relationship between various factors in the dataset search I examined such as relevance, confidence, and the need to refine queries were significant, it was not shown overall that content preview enhanced dataset search experience, as the t-test showed H4 was not supported. The visual below highlights the findings between the hypotheses, with a dotted line indicating no significance and a solid line with an accompanying significant value to indicate the hypothesis was correct.

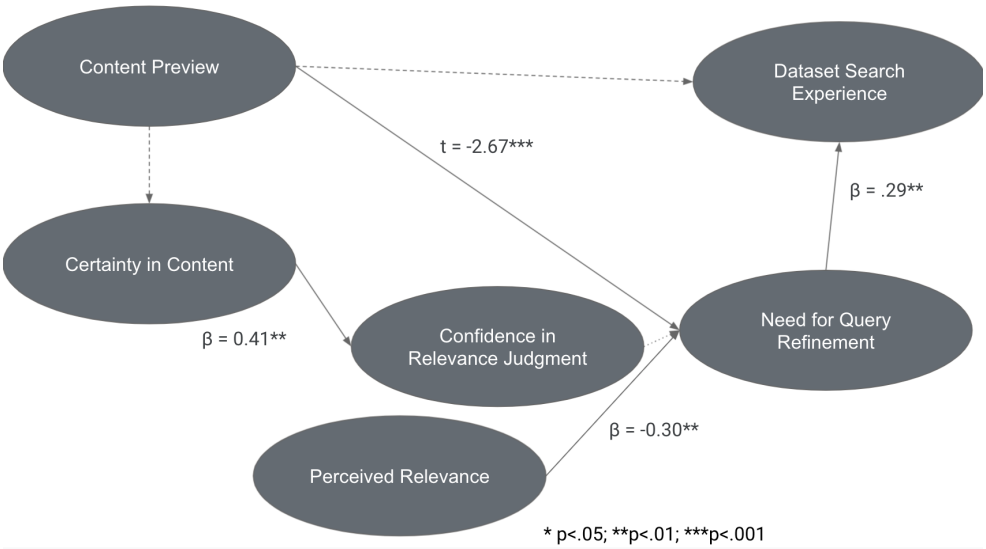


Figure 10. Hypotheses visualization. Hypotheses with significant results.

Discussion

While it is true that content preview reduces the need for query refinement, it was not true that it leads to more confidence in results. However, the analyses that were done did find that there was a significant relationship between certainty in content and confidence in the relevance judgement assigned. This supports the literature review that incorporated the definition of confidence with relevance. If a user finds a document, or in this case, a dataset to be highly relevant, then it appears to be true that they have high confidence in what they are looking for and viewing. Additionally, the analyses found that there was a significant relationship between relevance and the need to refine the query. This supports the literature that was reviewed that stated that users that find documents to be highly relevant are more satisfied with their search results. The literature also proved that satisfaction with results leads to fewer query refinements.

Since we know that content preview and perceived relevance are the two factors that have a significant relationship with the need for query refinement, we can look at the relationships that exist between the other factors and why that may be. While content preview did not have a significant relationship with certainty in content, certainty in content was significant regarding the confidence level of relevance judgment. The need for query refinement is also significantly associated with dataset search experience. The significance between these relationships could be that content preview does in fact lead users to need to refine the query, as they have a higher perceived relevance of the content they are looking at. Therefore, users might know they must refine the query to get better results for their dataset needs.

There was also a significant negative correlation between the need to refine the query and dataset search experience. As the literature stated, users issue fewer queries when satisfied with the results being presented to them. The need to refine the query influences dataset search experience as the user is not satisfied with the results being presented to them. While this experimental study was not fully functional

in providing users with search results, it is still interesting to note that the need to refine the query (not just the action of refining a query) contributes negatively to a user's dataset search experience.

In order to facilitate a more positive dataset search experience, we conclude that content preview leads to certainty in content. Confidence is an important factor that contributes to dataset search experience. We also know that relevance contributes to the need to refine the query and the need for query refinement is negatively associated with dataset search experience. Therefore, we can hypothesize that content preview reduces the need for query refinement which leads to a more positive overall dataset search experience.

Limitations & Future Work

In order to have a better understanding of how users might be searching for datasets, it would be interesting and helpful to conduct the study with an interactive and functional dataset search engine. If users were able to interact with a fully functional dataset search engine, they would be able to enter their own query, rather than have a preset query that was in this study. This would also allow users to receive tailored results specific to the search query that they entered. Another future direction of the study would be expanding the study to participants outside of Amazon mTurk to allow for more variations of participants.

Conclusion

The tests ran did not find a direct effect between content preview and enhanced dataset search experience. However, it was true that content preview was significantly linked to the need to refine the query. It was true that the perceived relevance of results led to query refinement, and it was also true that the need to refine the query was associated with a negative dataset search experience. Content preview reduces the need for query refinement which could overall lead to a better dataset search experience. In order for non-expert users of all domains to navigate searching for data in the expanding online world of data, having a dataset search engine that contains this content preview accompanying the results has proven to be important.

Acknowledgement: This report is based upon work supported by the National Science Foundation under Grant No. 1816325.

References

Hassan, Ahmed, and Ryen W. White. “Personalized Models of Search Satisfaction.” *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM '13*, 2013, doi:10.1145/2505515.2505681.

Koesten, Laura M., et al. “The Trials and Tribulations of Working with Structured Data.” *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, doi:10.1145/3025453.3025838.

Koesten, Laura, et al. “Everything You Always Wanted to Know about a Dataset: Studies in Data Summarisation.” *International Journal of Human-Computer Studies*, vol. 135, 2020, p. 102367., doi:10.1016/j.ijhcs.2019.10.004.

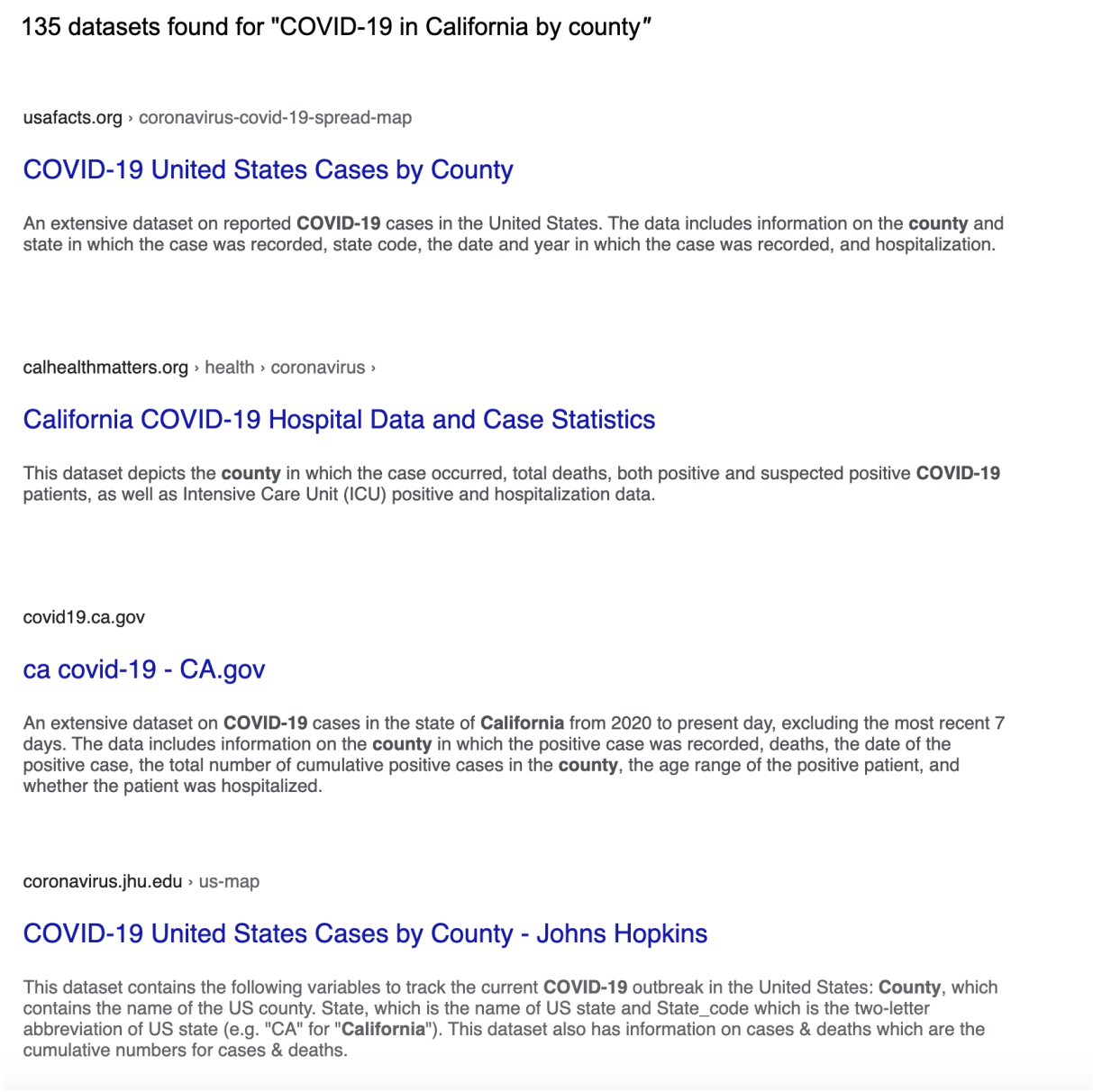
Marchionini, G., et al. “Accessing Government Statistical Information.” *Computer*, vol. 38, no. 12, 2005, pp. 52–61., doi:10.1109/mc.2005.393.

Ooi, Jessie, et al. “A Survey of Query Expansion, Query Suggestion and Query Refinement Techniques.” *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, 2015, doi:10.1109/icsecs.2015.7333094.

Suresh, Ralla. “ADVANCED APPROACH IN QUERY REFINEMENT USING REFINEMENT FILTERS FROM KNOWLEDGE BASE.” *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, 2 Apr. 2014.

Appendices

Appendix 1: Stimuli



Group 1 - Without content preview (COVID Data)

135 datasets found for "*Crime in Chicago by neighborhood*"

https://www.opendatanetwork.com › crime.fbi_ucr.rate

Crime in Chicago, IL | Open Source Data

An extensive dataset on reported **crime** in the City of **Chicago** from 2007-2021. The data includes information on the approximate **location** of the **crime**, date and year, **crime** type, weapon, and district.

https://home.chicagopolice.org › Statistics & Data

Statistical Reports | Chicago Police Department

Summary data on murders that occurred in **Chicago** by district, type of **location**, clearance, day month, and time, motive, and method.

https://www.icpsr.umich.edu › web › ICPSR › studies

Chicago Crimes, 2001-2018 - ICPSR - University of Michigan

This dataset contains information on **crimes** committed **in Chicago** including the crime ID number, the date (**2011**-2019) and address of the crime, the type of crime, whether an arrest was made (Yes/No) and the community area code.

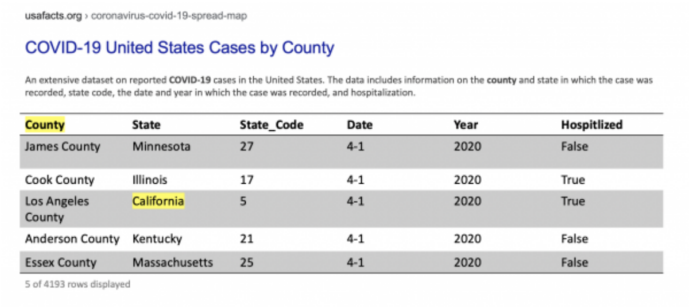
https://data.cityofchicago.org › Crimes-2001-to-Present

Crimes - Chicago Data Portal - City of Chicago

This dataset reflects reported incidents of **crime** that occurred **in** the City of **Chicago** from 2001 to present, minus the most recent seven days. Data is extracted from the **Chicago** Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of **crime** victims, addresses are shown at the Latitude and Longitude level.

Group 1 - Without content preview (Chicago Data)

135 datasets found for "COVID-19 in California by county"



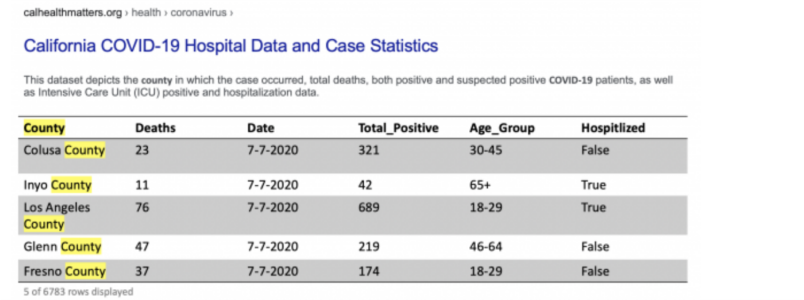
usafacts.org › coronavirus-covid-19-spread-map

COVID-19 United States Cases by County

An extensive dataset on reported **COVID-19** cases in the United States. The data includes information on the **county** and state in which the case was recorded, state code, the date and year in which the case was recorded, and hospitalization.

County	State	State_Code	Date	Year	Hospitlized
James County	Minnesota	27	4-1	2020	False
Cook County	Illinois	17	4-1	2020	True
Los Angeles County	California	5	4-1	2020	True
Anderson County	Kentucky	21	4-1	2020	False
Essex County	Massachusetts	25	4-1	2020	False

5 of 4193 rows displayed



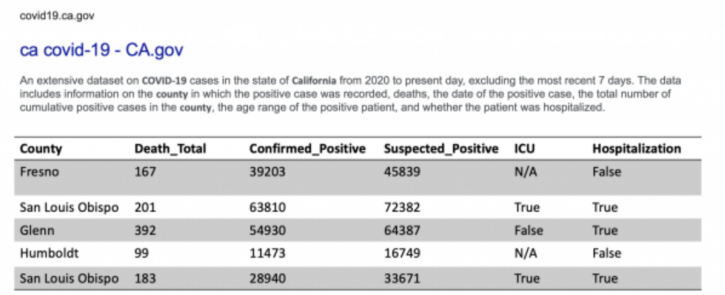
calhealthmatters.org › health › coronavirus ›

California COVID-19 Hospital Data and Case Statistics

This dataset depicts the **county** in which the case occurred, total deaths, both positive and suspected positive **COVID-19** patients, as well as Intensive Care Unit (ICU) positive and hospitalization data.

County	Deaths	Date	Total_Positive	Age_Group	Hospitlized
Colusa County	23	7-7-2020	321	30-45	False
Inyo County	11	7-7-2020	42	65+	True
Los Angeles County	76	7-7-2020	689	18-29	True
Glenn County	47	7-7-2020	219	46-64	False
Fresno County	37	7-7-2020	174	18-29	False

5 of 6783 rows displayed



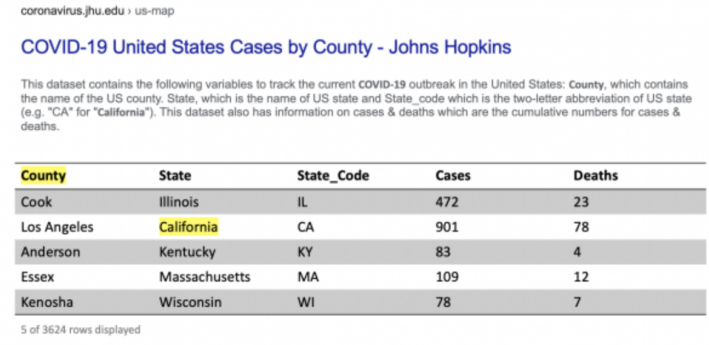
covid19.ca.gov

ca covid-19 - CA.gov

An extensive dataset on **COVID-19** cases in the state of **California** from 2020 to present day, excluding the most recent 7 days. The data includes information on the **county** in which the positive case was recorded, deaths, the date of the positive case, the total number of cumulative positive cases in the **county**, the age range of the positive patient, and whether the patient was hospitalized.

County	Death_Total	Confirmed_Positive	Suspected_Positive	ICU	Hospitalization
Fresno	167	39203	45839	N/A	False
San Louis Obispo	201	63810	72382	True	True
Glenn	392	54930	64387	False	True
Humboldt	99	11473	16749	N/A	False
San Louis Obispo	183	28940	33671	True	True

5 of 8256 rows displayed



coronavirus.jhu.edu › us-map

COVID-19 United States Cases by County - Johns Hopkins

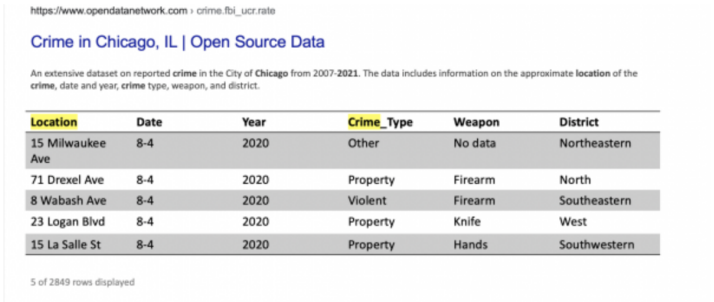
This dataset contains the following variables to track the current **COVID-19** outbreak in the United States: **County**, which contains the name of the US county. State, which is the name of US state and State_code which is the two-letter abbreviation of US state (e.g. "CA" for "**California**"). This dataset also has information on cases & deaths which are the cumulative numbers for cases & deaths.

County	State	State_Code	Cases	Deaths
Cook	Illinois	IL	472	23
Los Angeles	California	CA	901	78
Anderson	Kentucky	KY	83	4
Essex	Massachusetts	MA	109	12
Kenosha	Wisconsin	WI	78	7

5 of 3624 rows displayed

Group 2 - With content preview (COVID Data)

135 datasets found for "*Crime in Chicago by neighborhood*"



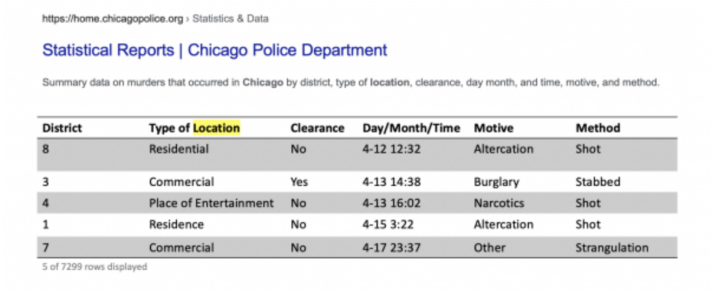
https://www.opendatanetwork.com › crime.fbi_ucr.rate

Crime in Chicago, IL | Open Source Data

An extensive dataset on reported **crime** in the City of **Chicago** from 2007-**2021**. The data includes information on the approximate **location** of the **crime**, date and year, **crime** type, weapon, and district.

Location	Date	Year	Crime_Type	Weapon	District
15 Milwaukee Ave	8-4	2020	Other	No data	Northeastern
71 Drexel Ave	8-4	2020	Property	Firearm	North
8 Wabash Ave	8-4	2020	Violent	Firearm	Southeastern
23 Logan Blvd	8-4	2020	Property	Knife	West
15 La Salle St	8-4	2020	Property	Hands	Southwestern

5 of 2849 rows displayed



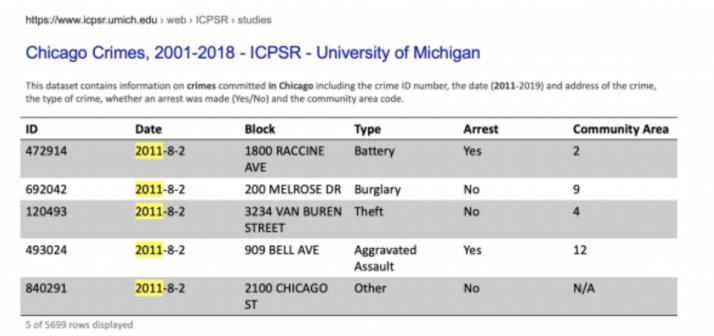
https://home.chicagopolice.org › Statistics & Data

Statistical Reports | Chicago Police Department

Summary data on murders that occurred in **Chicago** by district, type of **location**, clearance, day month, and time, motive, and method.

District	Type of Location	Clearance	Day/Month/Time	Motive	Method
8	Residential	No	4-12 12:32	Altercation	Shot
3	Commercial	Yes	4-13 14:38	Burglary	Stabbed
4	Place of Entertainment	No	4-13 16:02	Narcotics	Shot
1	Residence	No	4-15 3:22	Altercation	Shot
7	Commercial	No	4-17 23:37	Other	Strangulation

5 of 7299 rows displayed



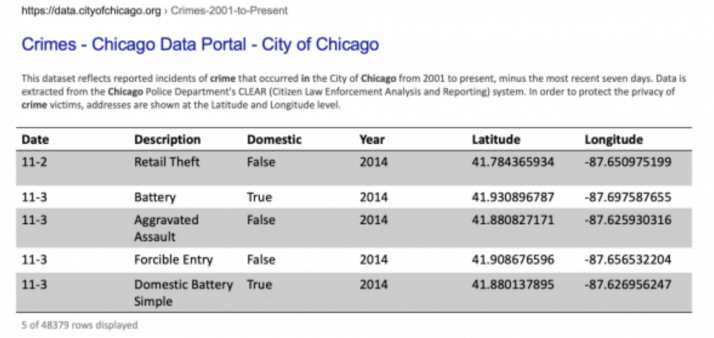
https://www.icpsr.umich.edu › web › ICPSR › studies

Chicago Crimes, 2001-2018 - ICPSR - University of Michigan

This dataset contains information on **crimes** committed **in Chicago** including the crime ID number, the date (**2011**-2019) and address of the crime, the type of crime, whether an arrest was made (Yes/No) and the community area code.

ID	Date	Block	Type	Arrest	Community Area
472914	2011-8-2	1800 RACCINE AVE	Battery	Yes	2
692042	2011-8-2	200 MELROSE DR	Burglary	No	9
120493	2011-8-2	3234 VAN BUREN STREET	Theft	No	4
493024	2011-8-2	909 BELL AVE	Aggravated Assault	Yes	12
840291	2011-8-2	2100 CHICAGO ST	Other	No	N/A

5 of 5699 rows displayed



https://data.cityofchicago.org › Crimes-2001-to-Present

Crimes - Chicago Data Portal - City of Chicago

This dataset reflects reported incidents of **crime** that occurred **in** the City of **Chicago** from 2001 to present, minus the most recent seven days. Data is extracted from the **Chicago** Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of **crime** victims, addresses are shown at the Latitude and Longitude level.

Date	Description	Domestic	Year	Latitude	Longitude
11-2	Retail Theft	False	2014	41.784365934	-87.650975199
11-3	Battery	True	2014	41.930896787	-87.697587655
11-3	Aggravated Assault	False	2014	41.880827171	-87.625930316
11-3	Forcible Entry	False	2014	41.908676596	-87.656532204
11-3	Domestic Battery Simple	True	2014	41.880137895	-87.626956247

5 of 48379 rows displayed

Group 2 - With content preview (Chicago Data)

Appendix 2: Questionnaire

Pre-Stimuli Questionnaire

How strongly do you agree or disagree with the following statements?

I often search for data online (e.g., using Google, Amazon Databases, governmental data repositories, etc.).

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I often work with data (e.g., collect, organize, analyze, and/or visualize data).

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I am familiar with terms and ideas related to statistics.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I am familiar with terms and ideas related to graphical and tabular displays.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I am confident in my abilities to search and collect useful data.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I am confident in my abilities to review and assess data.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

Post-Stimuli Questionnaire

Below please find one of the search results that you've seen before. You will be asked to evaluate this search result.

https://data.cityofchicago.org › Crimes-2001-to-Present

Crimes - Chicago Data Portal - City of Chicago

This dataset reflects reported incidents of **crime** that occurred **in** the City of **Chicago** from 2001 to present, minus the most recent seven days. Data is extracted from the **Chicago** Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of **crime** victims, addresses are shown at the Latitude and Longitude level.

How relevant do you think this search result is to your query: "Crime in Chicago by location?" On a scale of 1-5 with 1 being not relevant at all and 5 being very relevant

1 (Not Relevant)	2 (Slightly Relevant)	3 (Somewhat Relevant)	4 (Fairly Relevant)	5 (Extremely Relevant)
○	○	○	○	○

In the previous question, you indicated the search result as ${q://QID136/ChoiceGroup/SelectedChoices} to the query "Crime in Chicago by neighborhood".

How confident are you in your judgement of relevance?

1 (Not confident at all)	2 (Slightly Confident)	3 (Somewhat Confident)	4 (Fairly Confident)	5 (Very confident)
○	○	○	○	○

When searching for information online, if we are not completely satisfied with the search results, we may refine the search query (modify our search terms) and search again to meet our search needs.

How satisfied are you with the search results that were presented earlier?

1 (Not Satisfied At All)	2 (Slightly Satisfied)	3 (Somewhat Satisfied)	4 (Fairly Satisfied)	5 (Completely Satisfied)
○	○	○	○	○

How strongly do you feel the need to refine this query?

1 (Not At All)	2 (Slightly)	3 (Somewhat)	4 (Very)	5 (Extremely Strongly)
○	○	○	○	○

How easy would it be to modify the query?

1 (Not Easy)	2 (Slightly Easy)	3 (Somewhat Easy)	4 (Fairly Easy)	5 (Extremely Easy)
○	○	○	○	○

Think about the search results presented to you earlier. Please rate how much you agree or disagree with the following statements:

The content of the results matches my information needs.

1 (Strongly Disagree)	2 (Somewhat Disagree)	3 (Neither Agree nor Disagree)	4 (Somewhat Agree)	5 (Strongly Agree)
○	○	○	○	○

The search engine provides comprehensive information.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4(Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

This question serves to check if our participant is paying close attention. Please select "Strongly Disagree" for this question.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

The search engine provides information that matches my needs.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly Agree)
○	○	○	○	○

I intend to use this system for dataset search if it is available.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	5 (Strongly agree)
○	○	○	○	○

I intend to use this system for work if it is available.

1 (Strongly Disagree)	2 (Somewhat disagree)	3 (Neither agree nor disagree)	4 (Somewhat agree)	Click to write Choice 5
○	○	○	○	○

What helped you determine whether the datasets were a good match for your information needs? Please use the space below to write down your thoughts:

Which of the following is NOT a search result that was previewed earlier?

- ◯ Result 1 - Crime in Chicago, IL | Open Source Data
- ◯ Result 2 - Statistical Reports | Chicago Police Department
- ◯ Result 3 - Neighborhood Watch - Chicago Crime Rates and Statistics

Result 1:

https://www.opendatanetwork.com › crime.fbi_ucr.rate

Crime in Chicago, IL | Open Source Data

An extensive dataset on reported **crime** in the City of **Chicago** from 2007-2021. The data includes information on the approximate **location** of the **crime**, date and year, **crime** type, weapon, and district.

Result 2:

https://home.chicagopolice.org › Statistics & Data

Statistical Reports | Chicago Police Department

Summary data on murders that occurred in **Chicago** by district, type of **location**, clearance, day month, and time, motive, and method.

Result 3:

https://neighborhoodwatch.com› Statistics & Data

Neighborhood Watch - Chicago Crime Rates and Statistics

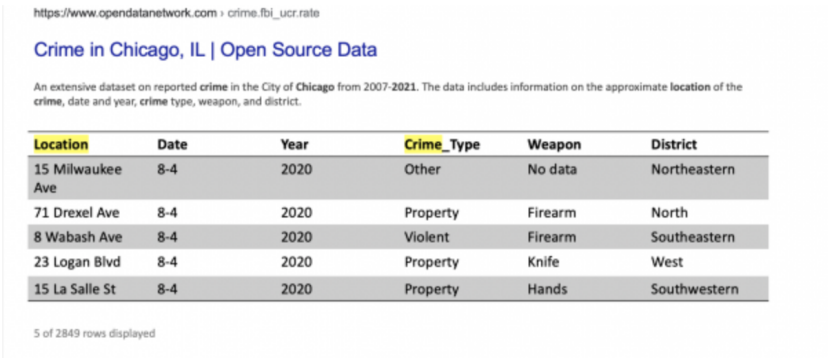
2021 **crime** and murder rates for **Chicago**, IL. This dataset contains information on a **neighborhood** level on various crime that has occurred in **Chicago** over the past several years.

Without preview

Which of the following was NOT a search result previewed earlier?

- ◯ Result 1 - Crime in Chicago, IL | Open Source Data
- ◯ Result 2 - Statistical Reports | Chicago Police Department
- ◯ Result 3 - Neighborhood Watch - Chicago Crime Rates and Statistics

Result 1:



https://www.opendatanetwork.com › crime.fbi_ucr.rate

Crime in Chicago, IL | Open Source Data

An extensive dataset on reported **crime** in the City of **Chicago** from 2007-**2021**. The data includes information on the approximate **location** of the **crime**, date and year, **crime** type, weapon, and district.

Location	Date	Year	Crime_Type	Weapon	District
15 Milwaukee Ave	8-4	2020	Other	No data	Northeastern
71 Drexel Ave	8-4	2020	Property	Firearm	North
8 Wabash Ave	8-4	2020	Violent	Firearm	Southeastern
23 Logan Blvd	8-4	2020	Property	Knife	West
15 La Salle St	8-4	2020	Property	Hands	Southwestern

5 of 2849 rows displayed

Result 2:

https://home.chicagopolice.org › Statistics & Data

Statistical Reports | Chicago Police Department

Summary data on murders that occurred in **Chicago** by district, type of **location**, clearance, day month, and time, motive, and method.

District	Type of Location	Clearance	Day/Month/Time	Motive	Method
8	Residential	No	4-12 12:32	Altercation	Shot
3	Commercial	Yes	4-13 14:38	Burglary	Stabbed
4	Place of Entertainment	No	4-13 16:02	Narcotics	Shot
1	Residence	No	4-15 3:22	Altercation	Shot
7	Commercial	No	4-17 23:37	Other	Strangulation

5 of 7299 rows displayed

Result 3:

https://neighborhoodwatch.com› Statistics & Data

Neighborhood Watch - Chicago Crime Rates and Statistics

2021 **crime** and murder rates for **Chicago**, IL. This dataset contains information on a neighborhood level on various **crime** that has occurred in **Chicago** over the past several years.

Neighborhood	Date	Year	Crime_Type	Weapon	Arrest
Stony Island	1-9	2021	Other	No data	No
Dunning	1-12	2021	Property	Firearm	Yes
Pullman	1-14	2021	Violent	Firearm	No
East Side	1-16	2021	Property	Knife	No
Burnside	1-23	2021	Property	Hands	Yes

5 of 2849 rows displayed

With preview

Demographics

Please specify your gender

- ◯ Male
- ◯ Female
- ◯ Prefer to self identify:
- ◯ Prefer not to specify

What is your age?

What is your Ethnicity/Race? Select all that apply to you.

- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Hispanic, Latinx, or Spanish Origin
- ☐ Middle Eastern or North African
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ White/Caucasian
- ☐ Some other race, ethnicity, or origin (please specify):
- ☐ I prefer not to answer

What is your highest education level?

- ◯ Less than a high school diploma
- ◯ High school diploma or equivalent (e.g. GED)
- ◯ Some college, no degree
- ◯ Associate degree (e.g. AA, AS)
- ◯ Bachelor's Degree (e.g. BA, BS)
- ◯ Master's Degree (e.g. MA, MS, MEd)
- ◯ Professional Degree (e.g. MD, DDS, DVM)
- ◯ Doctorate (e.g. PhD, EdD)
- ◯ Other: please specify

Appendix 3: Data

T-test Results

Relationship * *significant*	t	df	p-value	Mean (Group 1)	Mean (Group 2)

Condition & Relevance*	2.41	86	.01804	4.12	3.6
Condition & Need to Refine2*	-2.67	86	.009048	3.11	3.81
Condition & Satisfaction2*	1.96	86	.05278	3.93	3.46
Condition & Result Test*	-2.15	86	.03394	.35	.48
Condition & Certainty in Content	0.96	86	0.3393	4.1	3.95
Condition & Experience	0.68	86	.50	4	3.96

Simple Linear Regression Analysis Results

Relationship **significant*	*F*	*Adjusted* R^2	β
Perceived Relevance & Need for Query Refinement*	(1,86)	.10	-.30
Certainty in Content & Confidence in Relevance Judgement*	(1, 86)	.03	.28
Need for Query Refinement & Dataset Search Experience*	(1,86)	.08	-.19
Confidence in Relevant Judgement & Need for Query Refinement	(1,86)	0.008337	-0.09721
Perceived Relevance & Need for Query Refinement	(1,86)	-0.006931	-0.05467