Long Le
Kerensa Crump
Brent Freeman

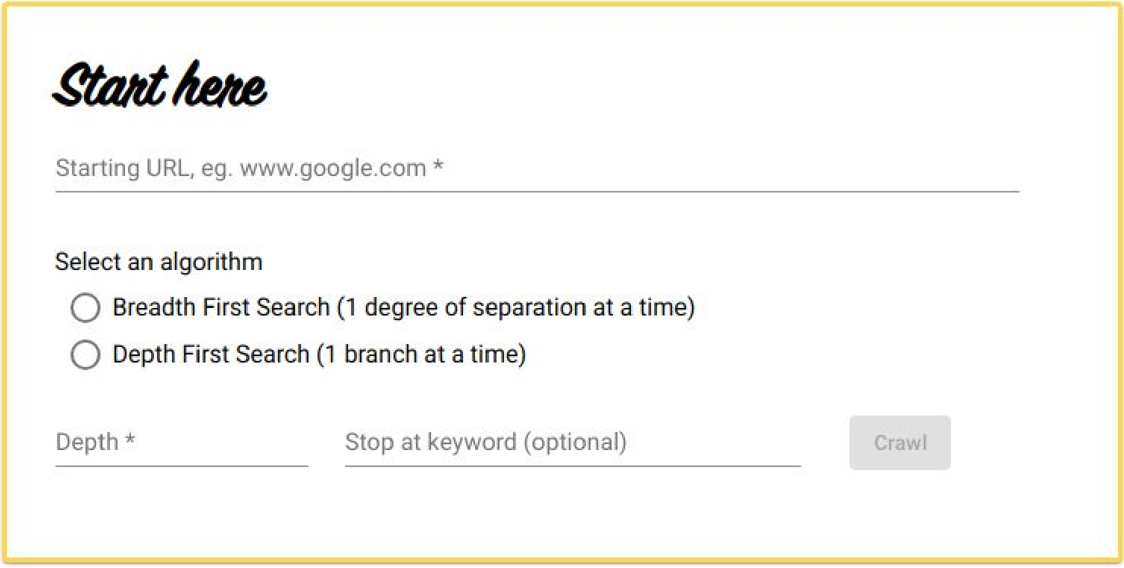
Team Gudja WebCrawler Instructions

The website is deployed on Google Cloud Platform at the following link:

https://gudja-webcrawler.appspot.com/

Please utilize the **Google Chrome** or **Chromium** web browser only. Other web browsers will not display the graphical design correctly and reduce user experience.

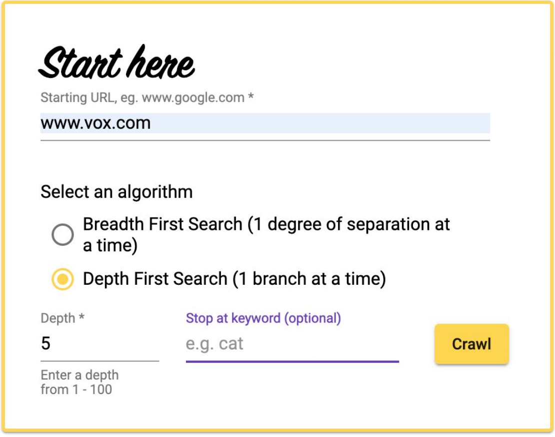
Website Instructions:


Start here
Starting URL, eg. www.google.com *
Select an algorithm
○ Breadth First Search (1 degree of separation at a time)
○ Depth First Search (1 branch at a time)
Depth *
Stop at keyword (optional)
Crawl

1. Enter a starting URL for the web crawler to start.
2. Choose either Breadth First Search or Depth First Search.
3. Choose the depth level of the crawl.
4. Enter a keyword to search for on the web pages. (optional)
5. Click Crawl!
6. Look at the visualization - hover over nodes for URL tooltip.
7. Repeat steps 1 - 6

Form Validation:

Quality input is ensured through proper form validation. Here, the required fields are Starting URL, Algorithm, and Depth. If any of these is not properly entered, the “Crawl” button will remained greyed out and the form will not submit. Hints beneath the input fields and text color changes in and below the input fields guide the user. Grey text is inactive, purple is active, and red is invalid. The depth field only accepts 1-4 for BFS and 1-100 for DFS. This reduces the risk of timing out for each algorithm. Other usability features include drop down menus on the URL and keyword fields as well as incremental arrows on the right side of the depth field. This makes repeated use of the crawler fast, intuitive, and pleasant.



Start here

Starting URL, eg. www.google.com *

www.vox.com

Select an algorithm

○ Breadth First Search (1 degree of separation at a time)

● Depth First Search (1 branch at a time)

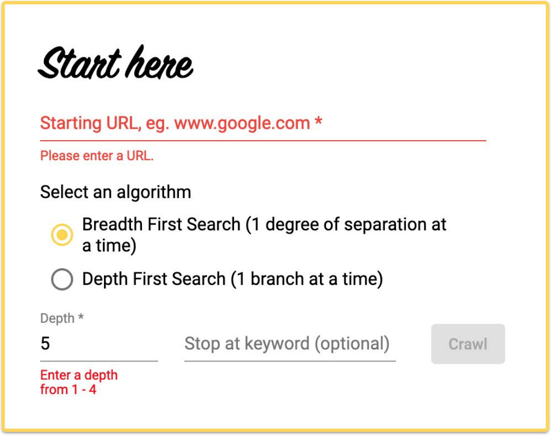
Depth *

5

Enter a depth from 1 - 100

Stop at keyword (optional)

e.g. cat



Start here

Starting URL, eg. www.google.com *

Please enter a URL.

Select an algorithm

● Breadth First Search (1 degree of separation at a time)

○ Depth First Search (1 branch at a time)

Depth *

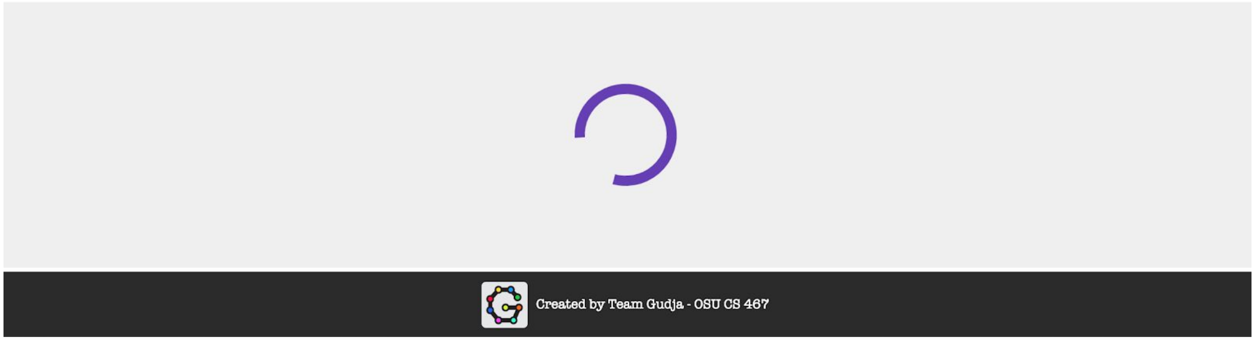
5

Enter a depth from 1 - 4

Stop at keyword (optional)

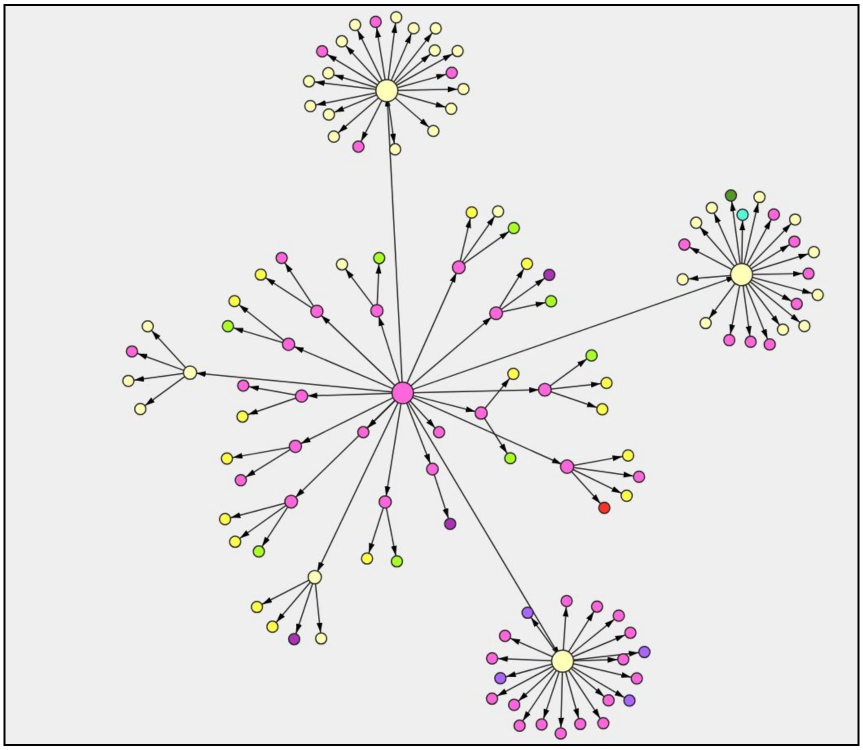
Waiting Spinner:

This waiting spinner animation is immediately displayed after clicking “Crawl” and disappears as soon as the data is received and the crawler graphics are rendered. This seamless transition gives the user useful feedback that their request has been submitted and is processing. The spinner will timeout after 10 minutes if data has not been received from the server and will present the error message “Error: timeout”.

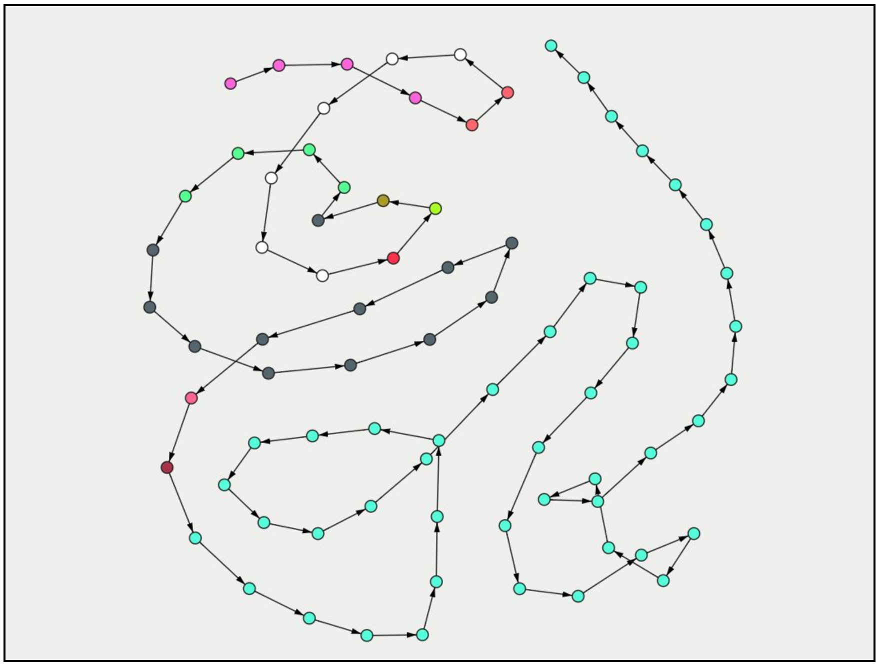


Created by Team Gudja - OSU CS 467

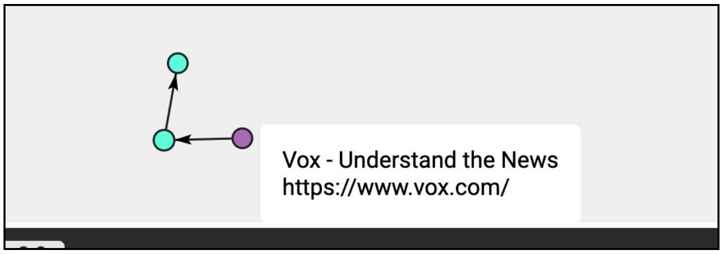
Breadth First Search Result:
BFS results generally follow a fireworks shaped pattern, creating clusters of nodes that expand out from the starting URL by the number of links equal to the depth.



Depth First Search Result:
DFS results strictly follow a single branch pattern.

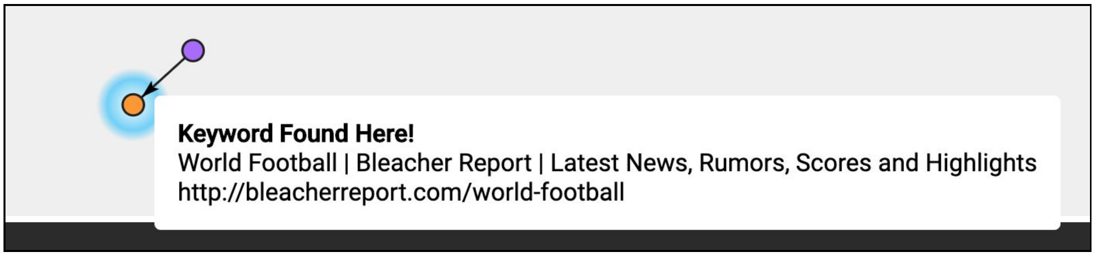


Tooltip: shows up when hovering over a node.
Click to open the link in a new tab



Vox - Understand the News
https://www.vox.com/

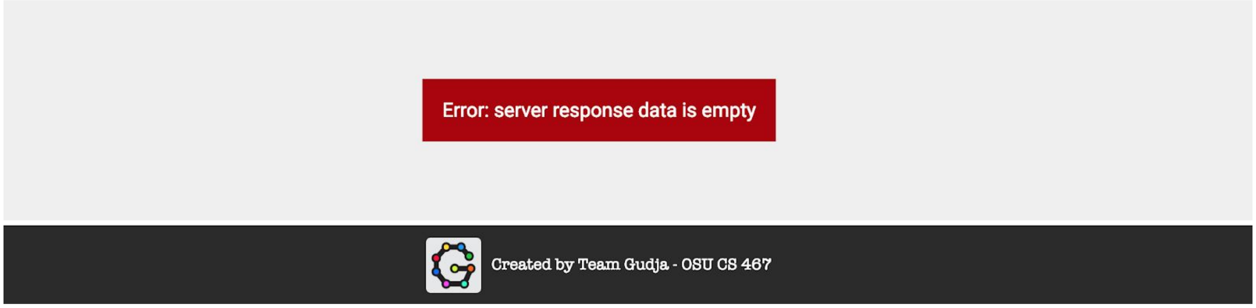
Found Keyword (halted program and highlighted URL):



Keyword Found Here!
World Football | Bleacher Report | Latest News, Rumors, Scores and Highlights
http://bleacherreport.com/world-football

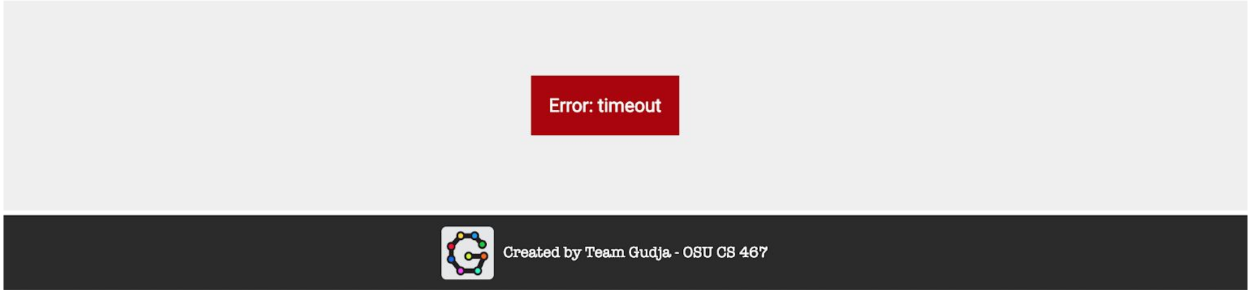
Error Messages:

On occasion during extensive crawls, the GCP cloud functions will malfunction and send back empty data, which cannot be rendered into a graphic. In this case, the error is displayed such as the image below. These errors can be caused by a few problems, notably timing out during a crawl. For both algorithms, and especially Breadth First Search, the Cloud functions attempt to make many http calls to the websites, if one of these http calls gets “hung up” it can cause a timeout which the Cloud Function can usually catch along with any other exceptions thrown by the Python function and will send an empty JSON. However, if the Google Cloud Function, on the Google business side has an issue, it should return but we have found instances where it does not return and does not timeout, which leaves the spinner spinning until 10 minutes is reached on the Angular timer and a timeout message is displayed. For further discussion on Google Cloud Functions see section III.iv and section V.i.



Error: server response data is empty

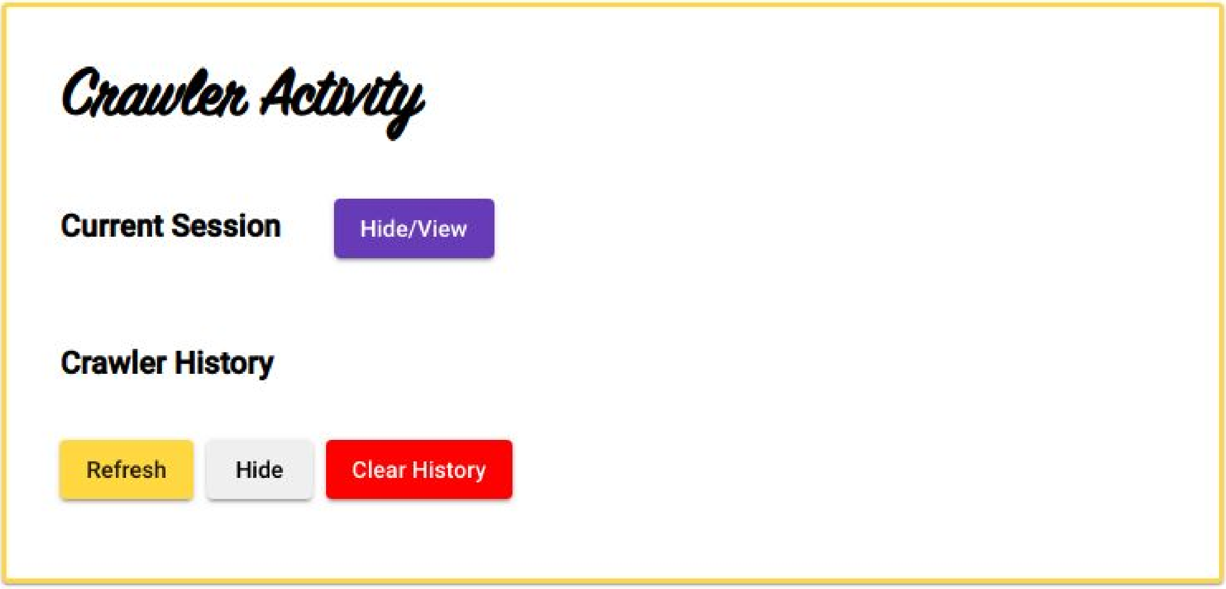
Created by Team Gudja - OSU CS 467



Error: timeout

Created by Team Gudja - OSU CS 467

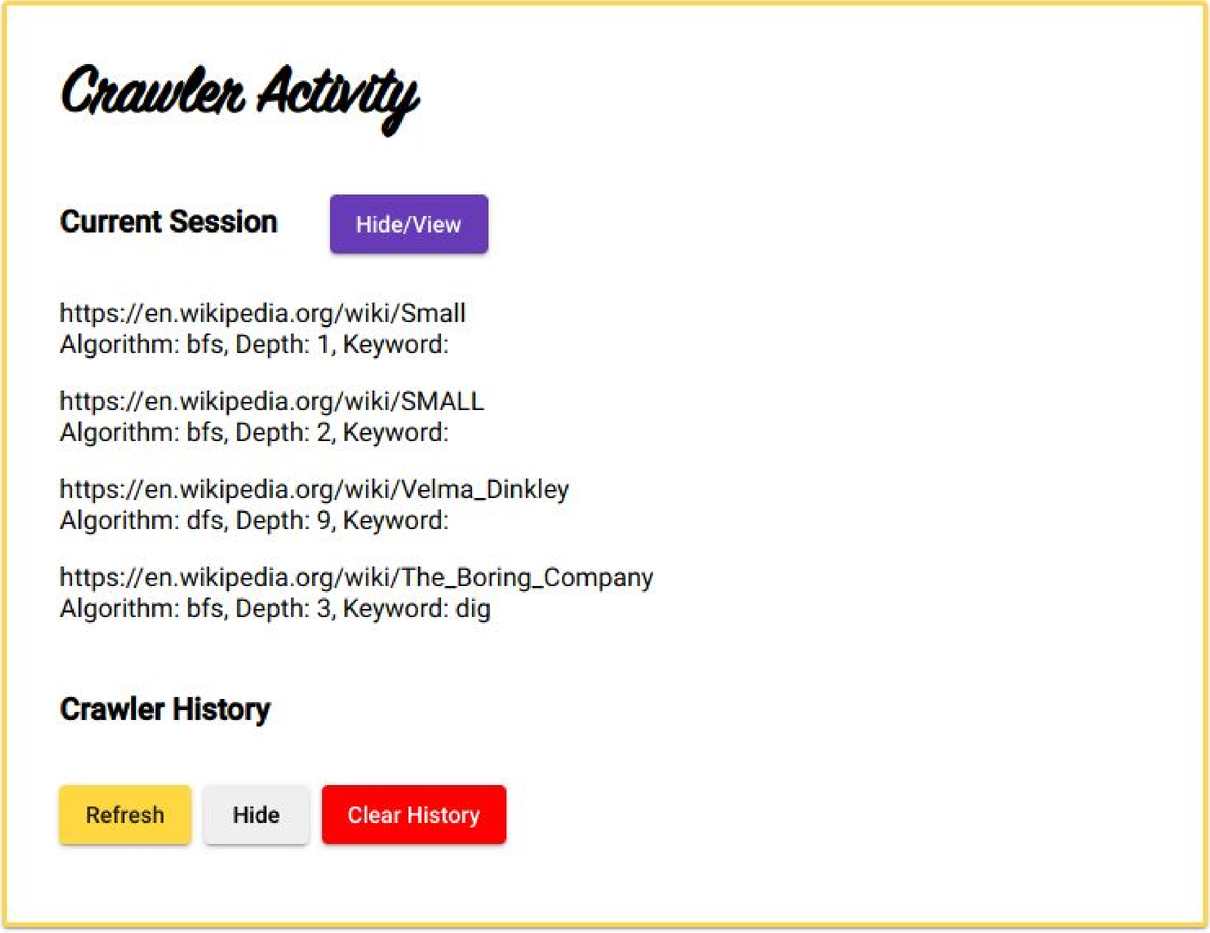
Session & History:



The Crawler Activity section keeps track of a user’s URL current session activity all completed crawl history, even after closing the browser. Submitted form data will be saved to Angular “posts” data model which is immediately updated in the Current Session section. The urlHistory and keywordFoundURL cookies are sent from the server with a successful response and contain the updated url and keyword data. This completed crawl data is updated in the Crawler History section and is accessible after closing the browser. Click the Refresh button to reload and display the previously entered URLs. On the other hand, the Clear History button allows the user to delete all of the information on the cookie.

The Hide/View and Hide buttons are used for functional reasons as well as enhancing aesthetics. Hiding the session history information or the previous sessions’ history allows the user to see more of the graphical results of the crawl without having to scroll further down.

Current Session Activity:



Crawler Activity

Current Session

https://en.wikipedia.org/wiki/Small
Algorithm: bfs, Depth: 1, Keyword:

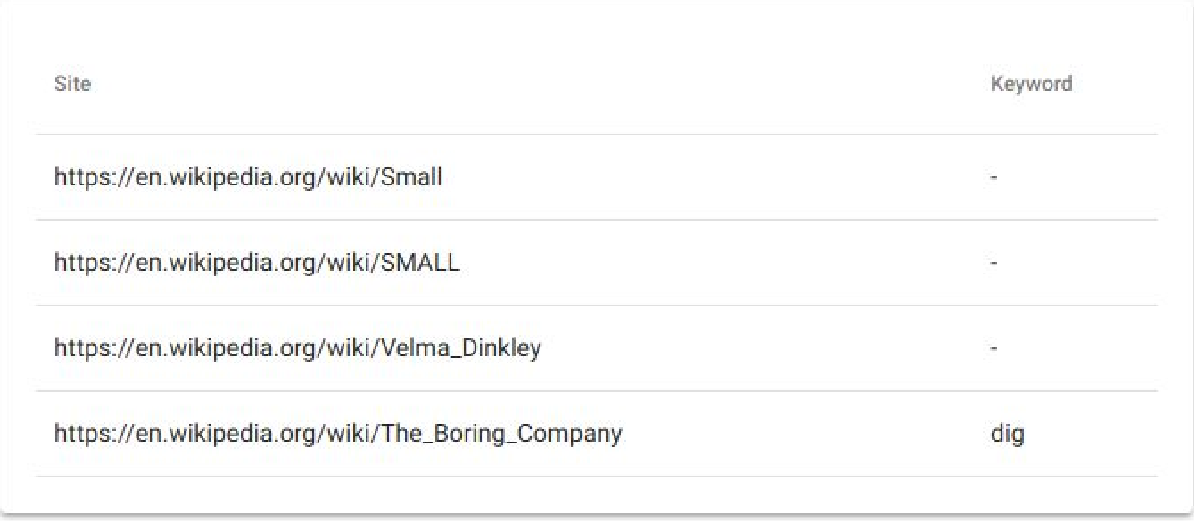
https://en.wikipedia.org/wiki/SMALL
Algorithm: bfs, Depth: 2, Keyword:

https://en.wikipedia.org/wiki/Velma_Dinkley
Algorithm: dfs, Depth: 9, Keyword:

https://en.wikipedia.org/wiki/The_Boring_Company
Algorithm: bfs, Depth: 3, Keyword: dig

Crawler History

History Recalled from Cookie:



Site	Keyword
https://en.wikipedia.org/wiki/Small	-
https://en.wikipedia.org/wiki/SMALL	-
https://en.wikipedia.org/wiki/Velma_Dinkley	-
https://en.wikipedia.org/wiki/The_Boring_Company	dig

Interesting Websites to crawl:

1. https://en.wikipedia.org/wiki/Main_Page
2. https://www.stackexchange.com
3. https://www.vox.com - Depth of 2 BFS

Download instructions:

View GitHub Repo: GudjaWebCrawler

1. Clone master branch to desired directory
2. Navigate to new directory root
3. Run npm install
4. Run node index.js
5. Website will be viewable at port 3000