Creating Your Own Link Collection

Workflow from Screaming Frog SEO to Gephi

This manual explains a method for creating a collection of hyperlinks from the live web with the program Screaming Frog SEO, and how to prepare it for further analysis, e.g. with the Gephi software.

The procedures are meant as a help for obtaining exhaustive data when using the free version of the program. With the paid version unlimited and more exhaustive (configured) crawls will be possible.

1. Create Your Own Link Collection

There are many ways to create a link collection. The web scraping program Screaming Frog SEO is one of them. A link to the program and a video with further introduction to the functions mentioned in this manual may be found here: https://cc.au.dk/en/cdmm/tools-and-tutorials/data-collection/screaming-frog-seo.

The procedure described here is handheld; it will be necessary to create new seed lists with web addresses found by the program, and reprocess them with the program in order to obtain more addresses via the new seed lists. The number of iterations necessary will depend on the research question(s) and the amount of data needed in order to facilitate a meaningful analysis.

1.1. Format of the Link List

The link list to be analysed in Gephi will have the following minimum format (e.g.):

source,target

dr.dk,jp.dk

Source is a web address where the link starts , target is a web address which the source address points to. In the example there has thus been found a single link from dr.dk to jp.dk.

The link list may also contain other information, e.g. number of links (will typically be a column count), or specific characteristics of the individual web pages (e.g. the anchor text for the link, page length, or similar), but in that case the data has to be split into a node table and an edge table (the example above is from an edge table).

1.2. Creating a Link List with Screaming Frog SEO

Use the following procedure:

(1) Make a starting list of URLs that your archiving process should start from. The URLs must include the protocol in order to be recognized as web pages, e.g. “https://www.dr.dk/” as opposed to just “www.dr.dk” or “dr.dk”. The list must contain one URL per line. Several file types are supported, but an easy way is to start with a simple .txt format. Name your starting list appropriately for your workflow, e.g. seeds_step1.
Tip: If your starting list will be based on pages you have actively found in web searches (as opposed to a pre-existing list of pages) you can auto generate your list, if you are using Mozilla Firefox for searching/browsing. You must first install the Firefox add-on “URLs List”. By clicking the button for the add-on in the Firefox menu bar you can directly copy/paste the URLs from all open tabs in a Firefox window. This list will have the correct format if pasted into a simple .txt file. The add-on and a detailed description may be found here: https://addons.mozilla.org/en-US/firefox/addon/urls-list/
(2) Open Screaming Frog SEO.
(3) In the top menu bar click ’Mode’ and select ’List’.
(4) Upload your starting URLs by uploading seeds_step1 (the button ’Upload’ in the top menu, ’From a file’). If your uploaded file seems faulty choose ’Enter Manually’ instead of ’From a file’ and copy/paste your starting list.
(5) Screaming Frog SEO will now find all the links that the URLs in the seed list point to.
(6) Select the ”Internal” tab, click/mark all seeds (the web addresses in the top window) – they can be marked with “mark all”. Now click ’Outlinks’ in the bottom window, and all the outgoing links will appear.
(7) Click ’Export’ at the top of the bottom window. (The button may be hidden unless the Screaming Frog SEO window is maximised).
(8) Download the output file. It is recommended to select the file type “Excel Workbook (.xlxs)). Name the download file appropriately for your workflow, e.g. result_step1.
The exported file contains a lot of interesting information such as anchor text, link type (Navigation, Content…) etc.; but here the columns of interest will be “From” and “To”.
(9) Open the downloaded file result_step1 and copy the ”To” column. Insert the content from the ”To” colums into a new Excel spreadsheet and remove duplicates. (Select the column and open the Data tab at the top of the ribbon. Find the Data Tools menu, and click Remove Duplicates. Press the OK button on the pop-up to remove duplicates from your data set.) Remember to save the new Excel file as an .xlsx file or you risk losing your data. Since this will be your next seed list you can name it accordingly, e.g. seeds_step2.
(10) Click ’Clear’ in Screaming Frog SEO (at the top where you also found the “Upload” button). Now upload seeds_step2 and repeat steps (3)-(9) in as many iterations as you need.
Important: The free version of Screaming Frog SEO has a maximum limit of 500 URLs, wherefore seed-lists with more than 500 URLs must be divided into shorter lists. It is recommended to limit your lists to 499 URLs in order to avoid hitting a ”free use maximum”.

The data set will quickly become quite large, and you will get far with just two iterations.

You may also consider exporting from the top window. This spreadsheet contains a lot of interesting information such as page length, word count, etc. This information can be useful to combine with a nodes table before importing to Gephi.

1.3. Gather, Clean, and Prepare Data

Collect all result_-lists in one file, e.g. using Excel. This file needs two columns, source and target. You will now have a complete list of all the links going from the web pages found in your iterations.

The data must now be cleaned (also supported in Excel):

- Reduce to domain addresses only (Data to text, separate by /). It may also be relevant to remove ”www” and similar.
- Remove duplicates in both columns. (See description at step (9) in the iterations process above).

2. Gephi

You should now have a spreadsheet file with two columns; source and target, and there should be only one copy of each link.

A copy of this file needs to be saved as a .csv file (save as “CSV (comma delimited)”) since this is a supported file format in Gephi.

The .csv file can now be imported in Gephi for further analysis of the hyperlink network.