

NETSPIDER

WEB SCRAPER FOR HOMELAND SECURITY INVESTIGATIONS



Ragy Costa de Jesus, Diego Grisales, Oscar Fox
PROJECT TEAM

TABLE OF CONTENTS

OVERVIEW.....	2
The Purpose	2
The Technology	2
The Impact	2
SYSTEM REQUIREMENTS	3
Processor.....	3
Memory (RAM).....	3
Storage	3
Operating System	3
INSTALLATION	3
Website for Releases	3
Install Application.....	3
Uninstall Application	7
APPLICATION SETTINGS	7
Select Paths	7
WEB SCRAPER SEARCH	8
Select Website	8
Select Location	9
Select Keywords	9
Select Set of Keywords	10
Type Text to Scrape	10
Select All Keywords	11
Find Only Posts with Payment Methods.....	11
Keyword Inclusive	12
EDIT KEYWORDS.....	12
Add Keywords	12
Remove Keyword	13
Add Set.....	13
Remove Set	14
VIEW DATA.....	14
View CSV File	14
View Screenshots	15

USER MANUAL

ABSTRACT



A software application that scrapes specified web pages in search for posts containing keywords that have been linked to human trafficking. Posts will be categorized based on the number of keywords found within the post. A post with more keywords will be prioritized over posts with fewer keywords. The location and payment options will be a factor in determining the priority of the leads. The software will provide HSI agents with leads to human trafficking posts on the internet, which will include all the information in the post along with a screenshot. The software product will streamline operations for discovering people and organizations that are supporting human trafficking. Identifying subjects and victims on open-source platforms. Support investigations of cybercrimes in coordination with computer forensics.

OVERVIEW

Florida Gulf Coast University (FGCU) and Homeland Security Investigations (HSI) are pleased to partner for the purpose of developing a software solution to counter web advertisement of human trafficking. The FGCU senior project team comprised of Ragy Costa de Jesus, Diego Grisales, Oscar Fox achieved goals to streamline operations for discovering people and organizations that are supporting human trafficking with the NetSpider software application.

The Purpose



This software will be important in identifying human trafficking victims and suspects on open-source platforms, as well as supporting investigations of cybercrimes in coordination with computer forensics. Ultimately, the software will aid in combating human trafficking and protecting vulnerable individuals from exploitation.

The Technology



- *Python programming language to develop the program.*
- *Selenium library for web scrapping automation and screenshot capture.*
- *PyQt6 library to implement the user interface.*
- *Pandas library to create CSV files with ad results.*

The Impact



Homeland Security Investigations is investigating different cases at the same time with limited personnel. In addition, each of these cases involves an exhausting amount of time searching on the internet for leads and writing these leads in a file or on paper. However, by creating the HSI Web Scraper the amount of work will be reduced by more than 63% by performing an algorithm to find all the relevant keywords and store these in a file automatically. As a result, the agents will have an already sorted by priority file with all the relevant data for them to check for false-positive or positive leads.

SYSTEM REQUIREMENTS

The minimum and recommended system requirements for the software application.

Processor

1GHz processor or faster

Memory (RAM)

Minimum: 2 GB

Recommended: 4 GB or greater

Storage

500 MB hard disk space or greater

Operating System

Windows: 7 and above

macOS: Monterey (version 12) and above

INSTALLATION

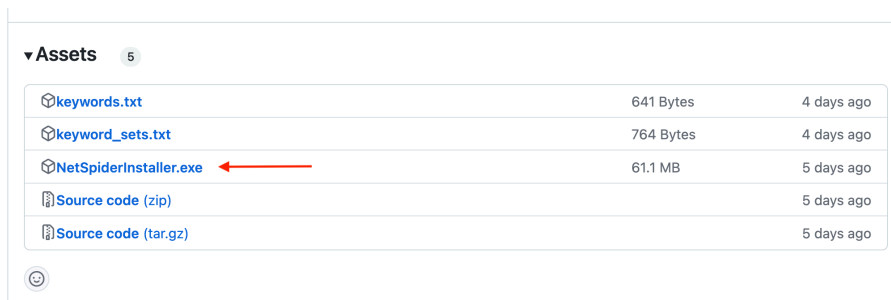
The recommended steps for installing and uninstalling the software application. These steps may vary depending on the host system.

Website for Releases

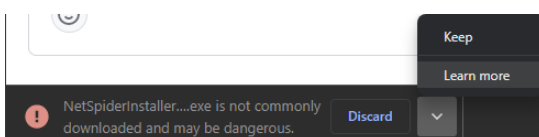
<https://www.github.com/dfgrisales5078/HSI-Web-Scraper/releases>

Install Application

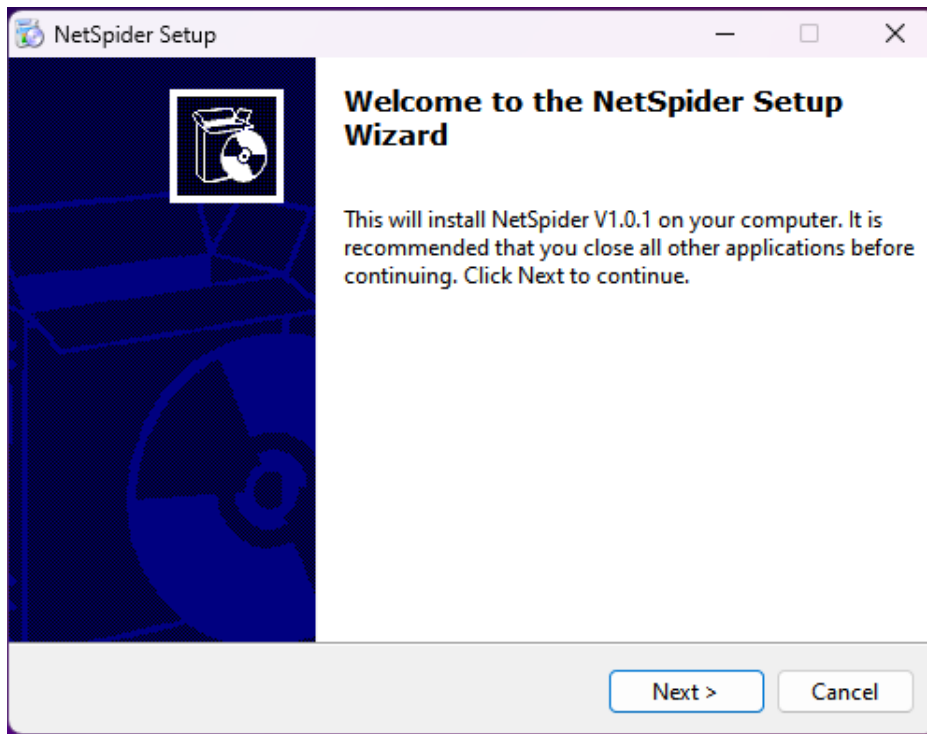
- 1 Navigate to **website for releases** and click **NetSpiderInstaller.exe** for installation.



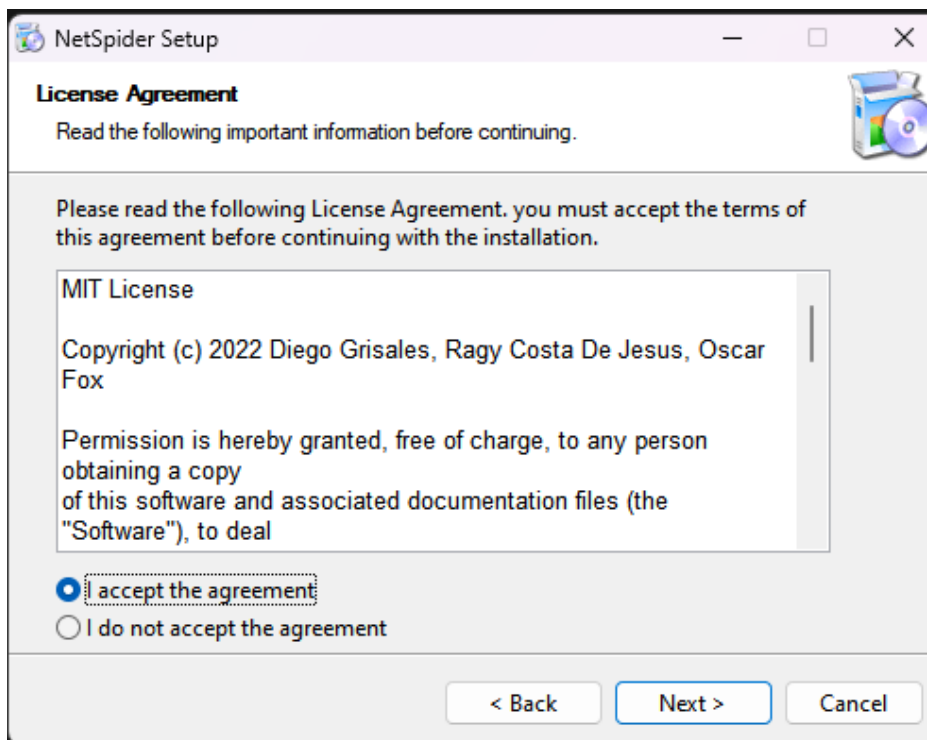
- 2 Click the installation icon in the browser and select to open NetSpider Setup Wizard.
Warning: the browser will say uncommon application download. This is normal because it is not a known system. Click the arrow then click KEEP.



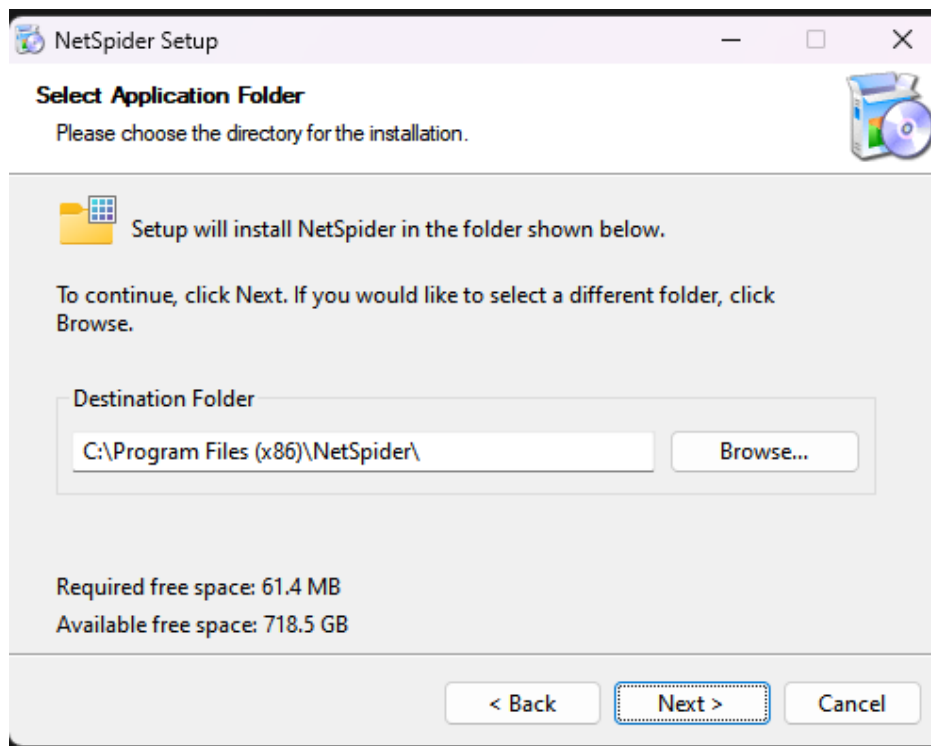
- 3 Welcome section for **NetSpider Setup Wizard**, click the **Next** button to get started with installation.



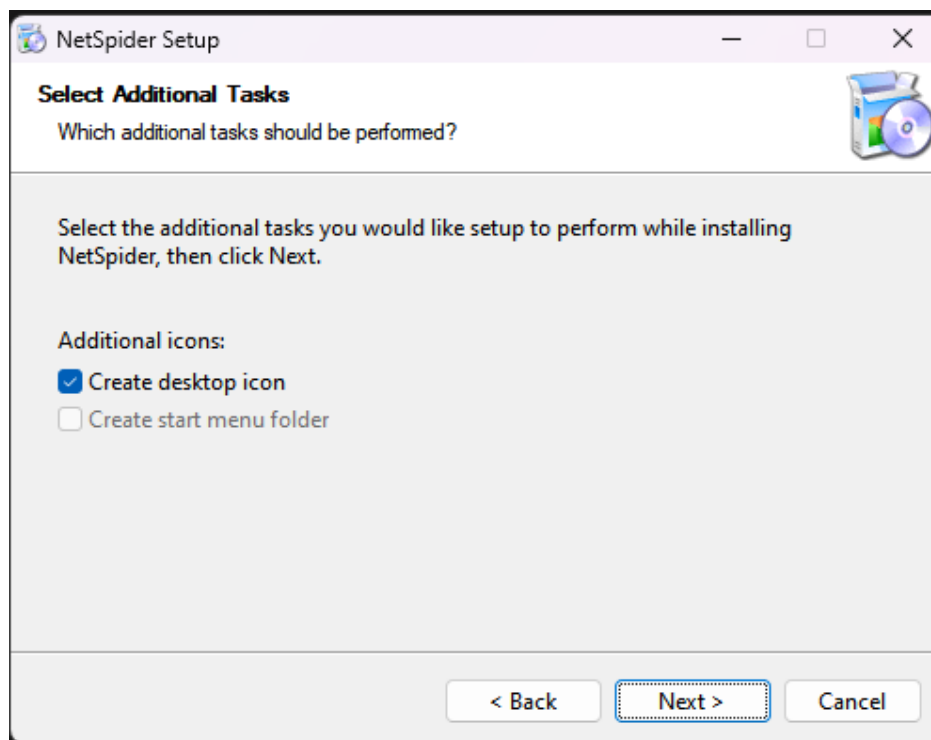
- 4 In the NetSpider Setup window read the **License Agreement** and click "**I accept the agreement**" if you agree to the terms. After, click the **Next** button to continue.



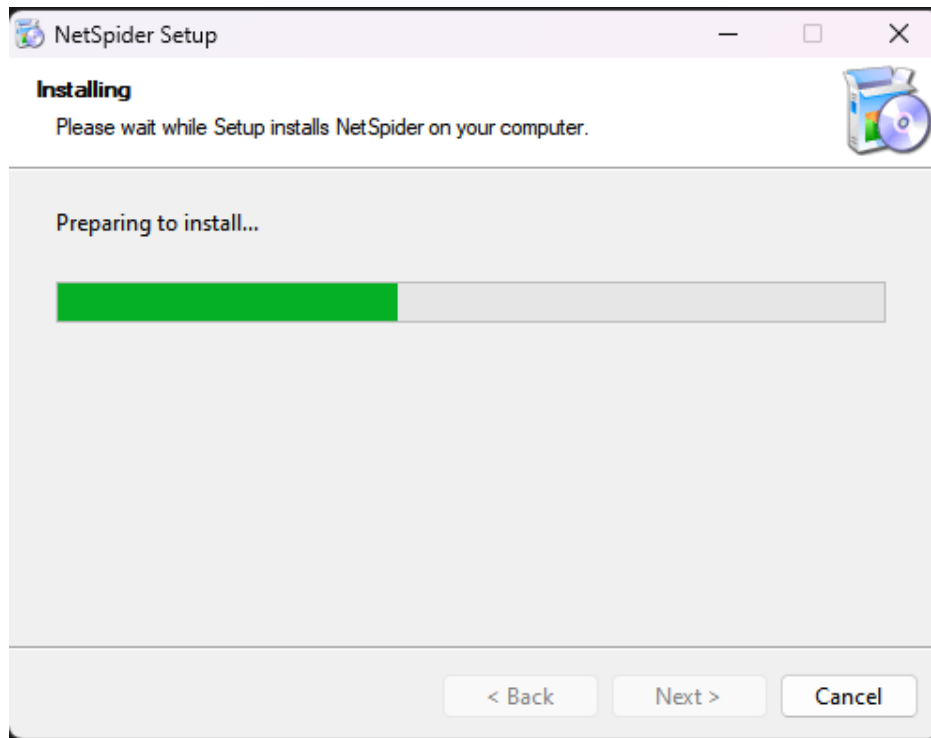
- 5 Within the **Select Application Folder** section, **browse** and choose location to install NetSpider application. After, click the **Next** button to continue.



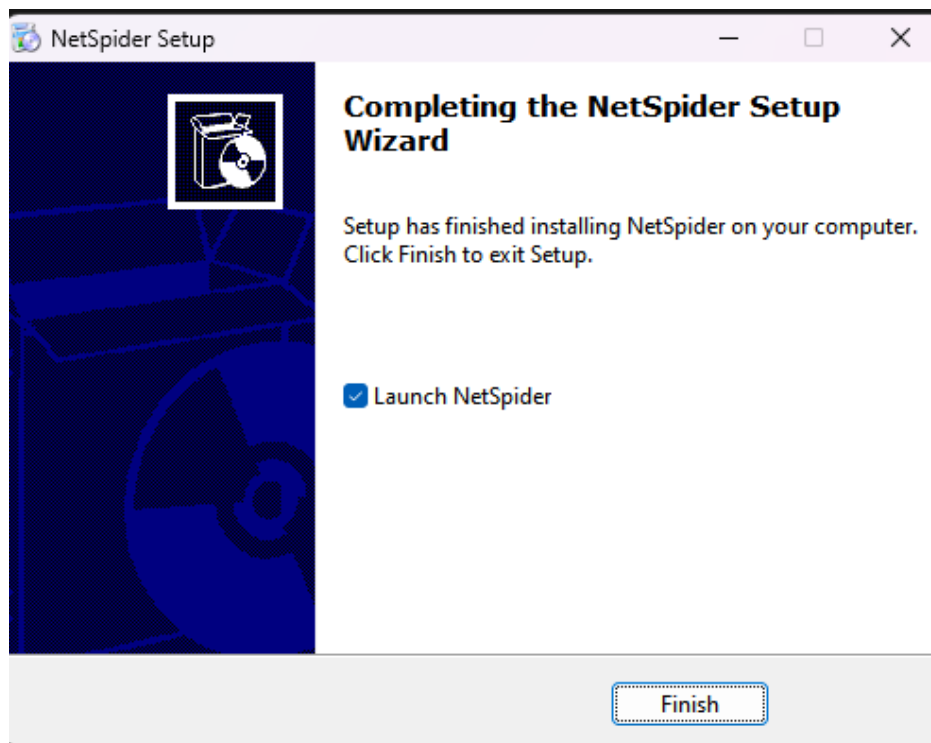
- 6 Within the **Select Additional Tasks** section, choose additional icons for shortcuts to NetSpider application.



- 7 Wait for Installation to complete for the NetSpider application.

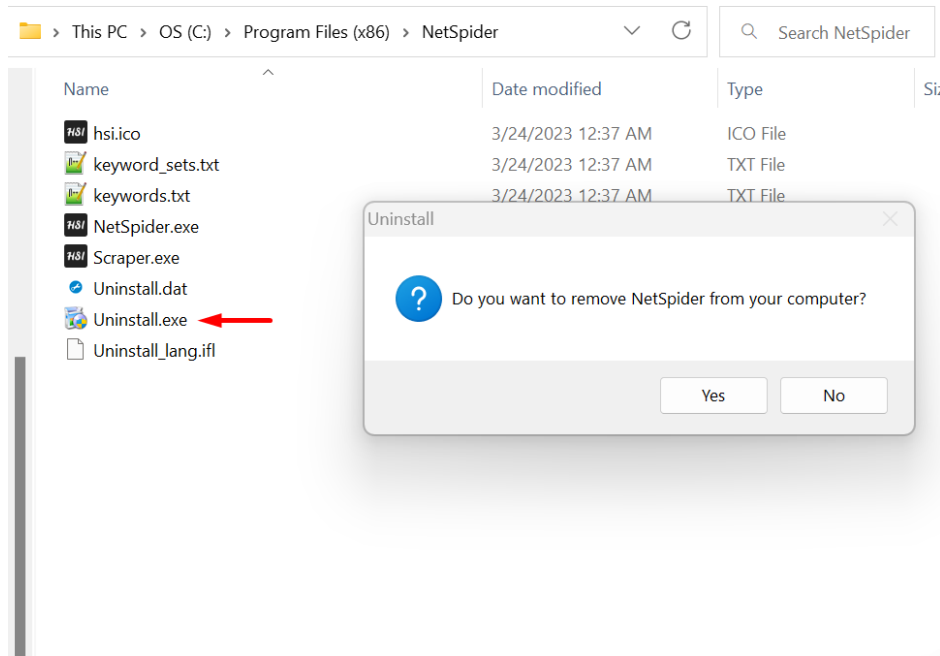


- 8 Completing the NetSpider Setup, click **Launch NetSpider** checkbox and **Finish** button.



Uninstall Application

- 1 To uninstall the application, go to the file path the application was installed (**By Default:** C: > Program Files (x...) > NetSpider) and click **Uninstall.exe**, then confirm you want to remove NetSpider from the computer.



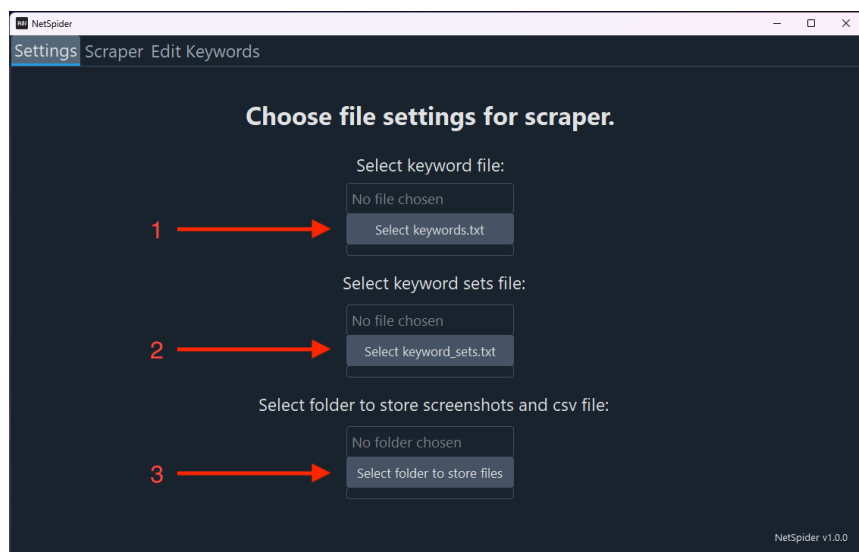
APPLICATION SETTINGS

The **Settings** tab is intended to handle the file paths for the web scraper to support the management of keywords/sets, as well as providing a location for the results (csv and screenshots) to be exported.

Select Paths

- 1 Select the **keyword.txt** file to include your keywords.
- 2 Select the **keyword_sets.txt** file to include your sets of keywords.
- 3 Select the folder where you want the results exported, to include csv file and screenshots.


* Each person may have their own folder.



WEB SCRAPER SEARCH

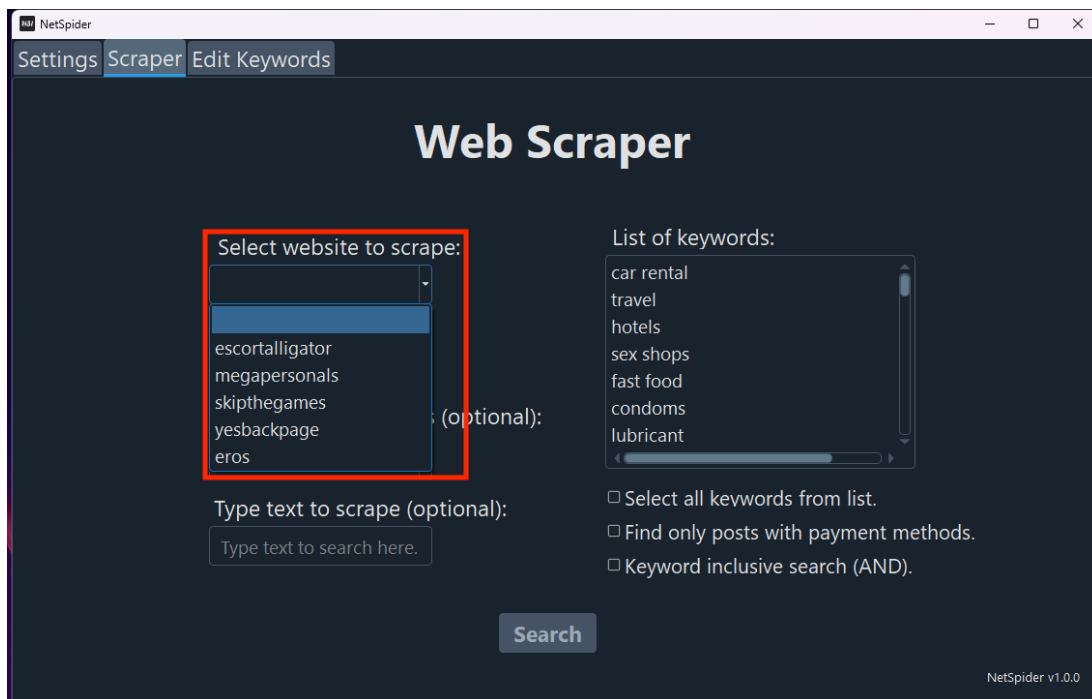
The **Scraper** tab is the main functionality of the application. This section is intended to grab data from open-source platforms based on search criteria.

Select Website

 Selecting the website determines what **open-source platform** the web scraper will gather data from for further analysis.

Available data sources:

- Escortalligator - escortalligator.com.listcrawler.eu
- Megapersonals - megapersonals.eu
- Skipthegames - skipthegames.com
- Yesbackpage - www.yesbackpage.com
- Eros - www.eros.com



The screenshot shows the NetSpider application window with the 'Scraper' tab selected. The interface has a dark theme. At the top, there are three tabs: 'Settings', 'Scraper', and 'Edit Keywords'. The main heading 'Web Scraper' is centered. Below it, there are two main sections. The left section is titled 'Select website to scrape:' and contains a dropdown menu with a list of websites: escortalligator, megapersonals, skipthegames, yesbackpage, and eros. The right section is titled 'List of keywords:' and contains a list of keywords: car rental, travel, hotels, sex shops, fast food, condoms, and lubricant. Below the keywords list, there are three checkboxes: 'Select all keywords from list.', 'Find only posts with payment methods.', and 'Keyword inclusive search (AND)'. At the bottom, there is a 'Search' button. The version number 'NetSpider v1.0.0' is displayed in the bottom right corner.

NetSpider

Settings Scraper Edit Keywords

Web Scraper

Select website to scrape:

- escortalligator
- megapersonals
- skipthegames
- yesbackpage
- eros

(optional):

Type text to scrape (optional):

Type text to search here.

List of keywords:

- car rental
- travel
- hotels
- sex shops
- fast food
- condoms
- lubricant

☐ Select all keywords from list.

☐ Find only posts with payment methods.

☐ Keyword inclusive search (AND).

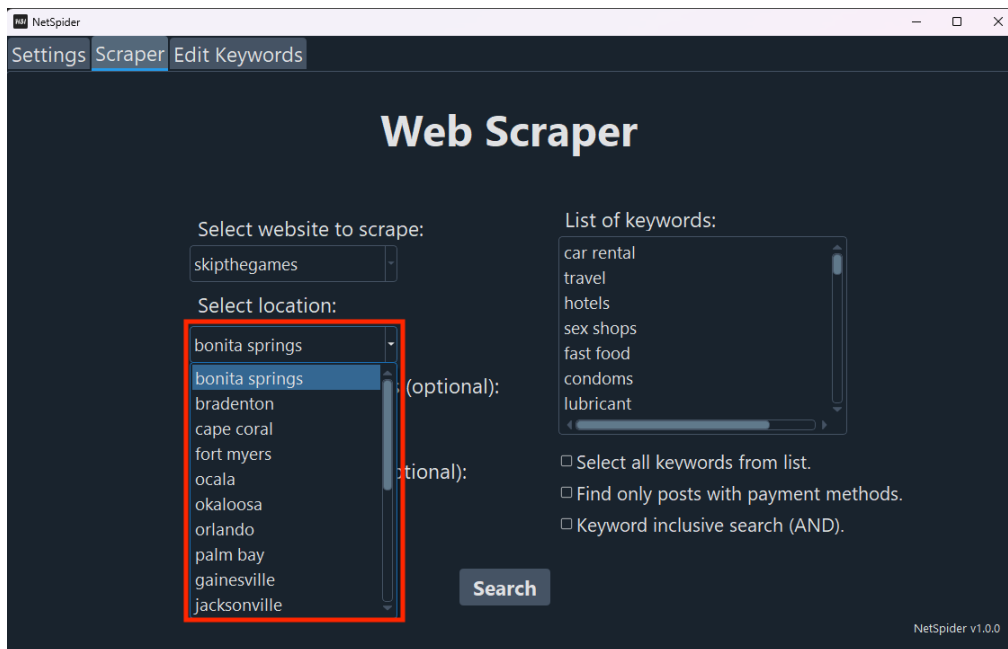
Search

NetSpider v1.0.0

Select Location

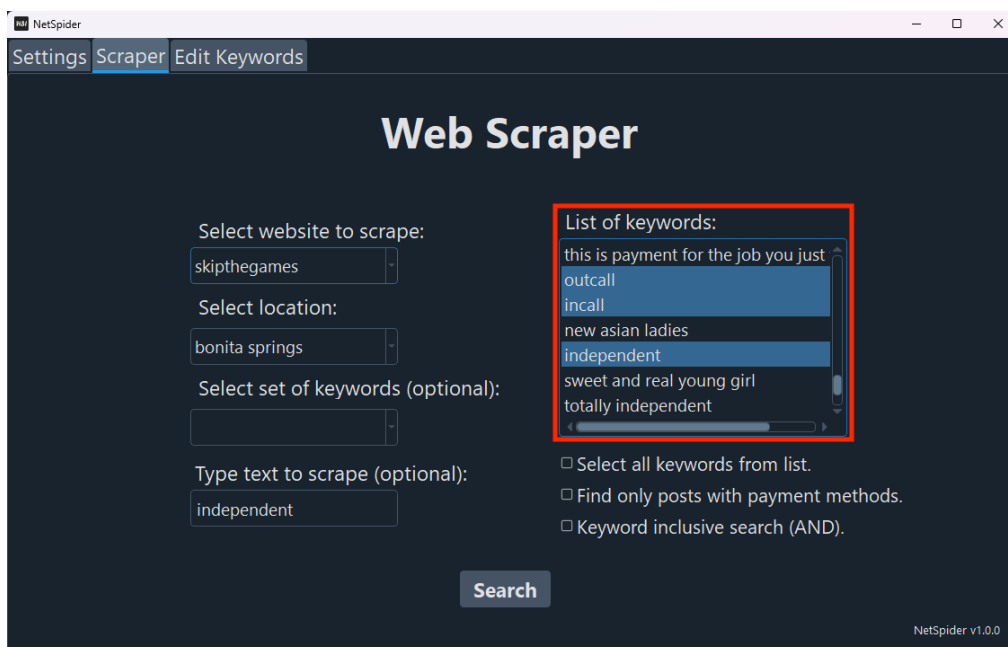
📍 Selecting the location determines what **city** in Florida the web scarper will gather data from for further analysis.

*** The cities displayed are specific to the website selected.**



Select Keywords

📋 By selecting the keywords, the results are refined to exclusively comprise of data that contains the chosen keywords.



Select Set of Keywords

- By selecting a set of keywords, the results are refined to exclusively comprise of data that contains the keywords within that set.

The screenshot shows the NetSpider Web Scraper application window. The 'Scraper' tab is selected. The interface includes the following elements:

- Select website to scrape:** A dropdown menu with 'skipthegames' selected.
- Select location:** A dropdown menu with 'bonita springs' selected.
- Select set of keywords (optional):** A dropdown menu that is open, showing a list of keyword sets: 'child trafficking trends', 'child traffi...t code words', 'adult trafficking', and 'working set'. This dropdown is highlighted with a red rectangle.
- List of keywords:** A scrollable list containing: 'car rental', 'travel', 'hotels', 'sex shops', 'fast food', 'condoms', and 'lubricant'.
- Search options:** Three checkboxes: 'Select all keywords from list.', 'Find only posts with payment methods.', and 'Keyword inclusive search (AND)'.
- Search button:** A button labeled 'Search'.
- Footer:** 'NetSpider v1.0.0' in the bottom right corner.

Type Text to Scrape

- By typing desired text, the results are refined to exclusively comprise of data that contains the text input. This feature may be used to search for specific phone numbers, email addresses, social media usernames, phrases, etc. Note: The contents of the text will be searched for in its entirety.

The screenshot shows the NetSpider Web Scraper application window, similar to the previous one, but with the 'Type text to scrape (optional):' field highlighted with a red rectangle. The interface elements are:

- Select website to scrape:** A dropdown menu with 'skipthegames' selected.
- Select location:** A dropdown menu with 'bonita springs' selected.
- Select set of keywords (optional):** A dropdown menu that is currently empty.
- Type text to scrape (optional):** A text input field containing the word 'independent'.
- List of keywords:** A scrollable list containing: 'car rental', 'travel', 'hotels', 'sex shops', 'fast food', 'condoms', and 'lubricant'.
- Search options:** Three checkboxes: 'Select all keywords from list.', 'Find only posts with payment methods.', and 'Keyword inclusive search (AND)'.
- Search button:** A button labeled 'Search'.
- Footer:** 'NetSpider v1.0.0' in the bottom right corner.

Select All Keywords



By selecting all keywords from list, results are refined to exclusively comprise of data that contains all the keywords in the list.

NetSpider

Settings Scrapper Edit Keywords

Web Scraper

Select website to scrape:
skipthegames

Select location:
bonita springs

Select set of keywords (optional):

Type text to scrape (optional):
independent

List of keywords:
new asian ladies
independent
sweet and real young girl
totally independent
agency
serve many cities
carcall

☒ Select all keywords from list.
☐ Find only posts with payment methods.
☐ Keyword inclusive search (AND).

Search

NetSpider v1.0.0

Find Only Posts with Payment Methods



By selecting find only posts with payment methods, the results are refined to exclusively comprise of data that contains methods of payment.

NetSpider

Settings Scrapper Edit Keywords

Web Scraper

Select website to scrape:
skipthegames

Select location:
bonita springs

Select set of keywords (optional):

Type text to scrape (optional):
Type text to search here.

List of keywords:
this is payment for the job you just
outcall
incall
new asian ladies
independent
sweet and real young girl
totally independent

☐ Select all keywords from list.
☒ Find only posts with payment methods.
☐ Keyword inclusive search (AND).

Search

NetSpider v1.0.0

Keyword Inclusive



By selecting keywords inclusive search (AND), the results are refined to exclusively comprise of data that contains ALL the keywords selected.

EDIT KEYWORDS


The **Edit Keywords** tab is for management of the terms and phrases stored in text files (keywords.txt and keyword_sets.txt) so that they may be used as search criteria.

Add Keywords

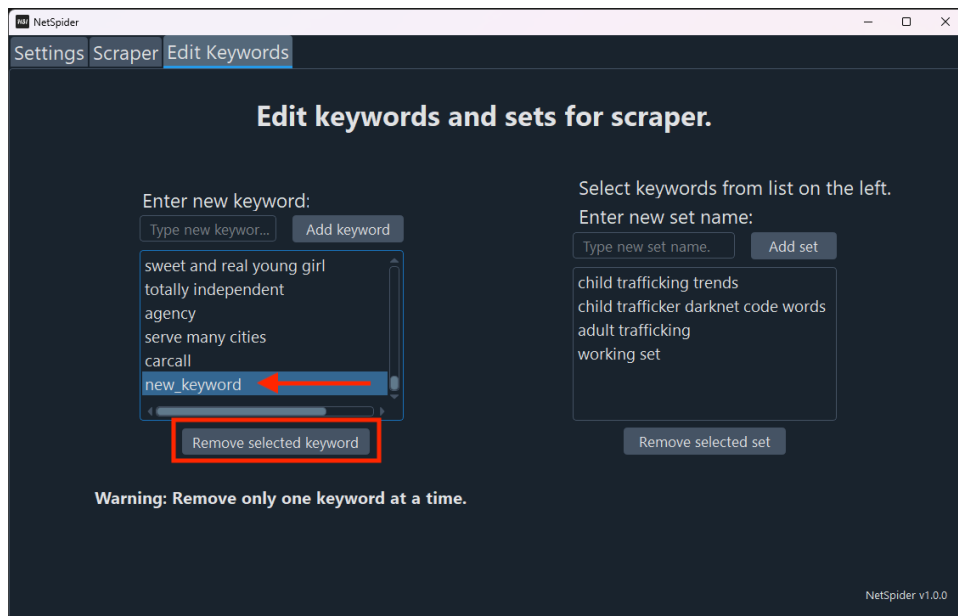


By selecting keywords inclusive search (AND), the results are refined to exclusively comprise of data that contains ALL the keywords selected.


Remove Keyword

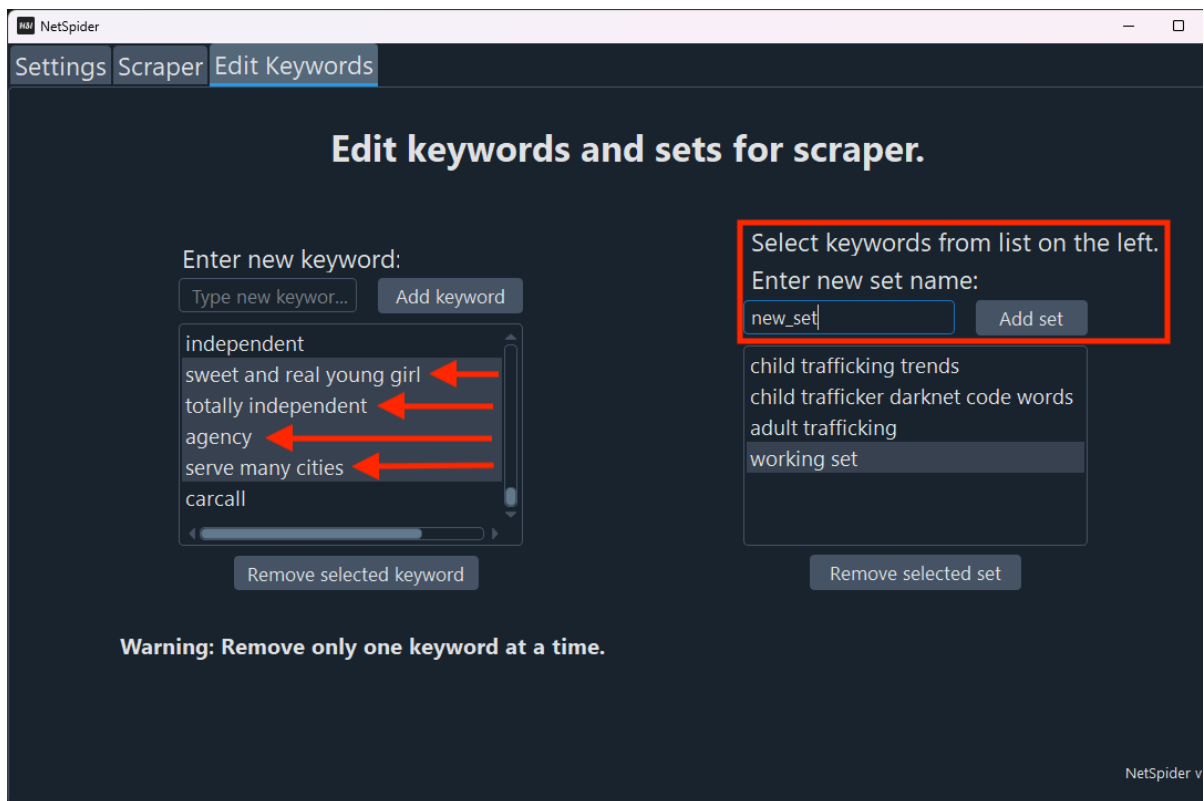
 By clicking remove selected keyword, the last term selected in the keyword list will be removed.

*** Only one keyword can be removed at a time.**




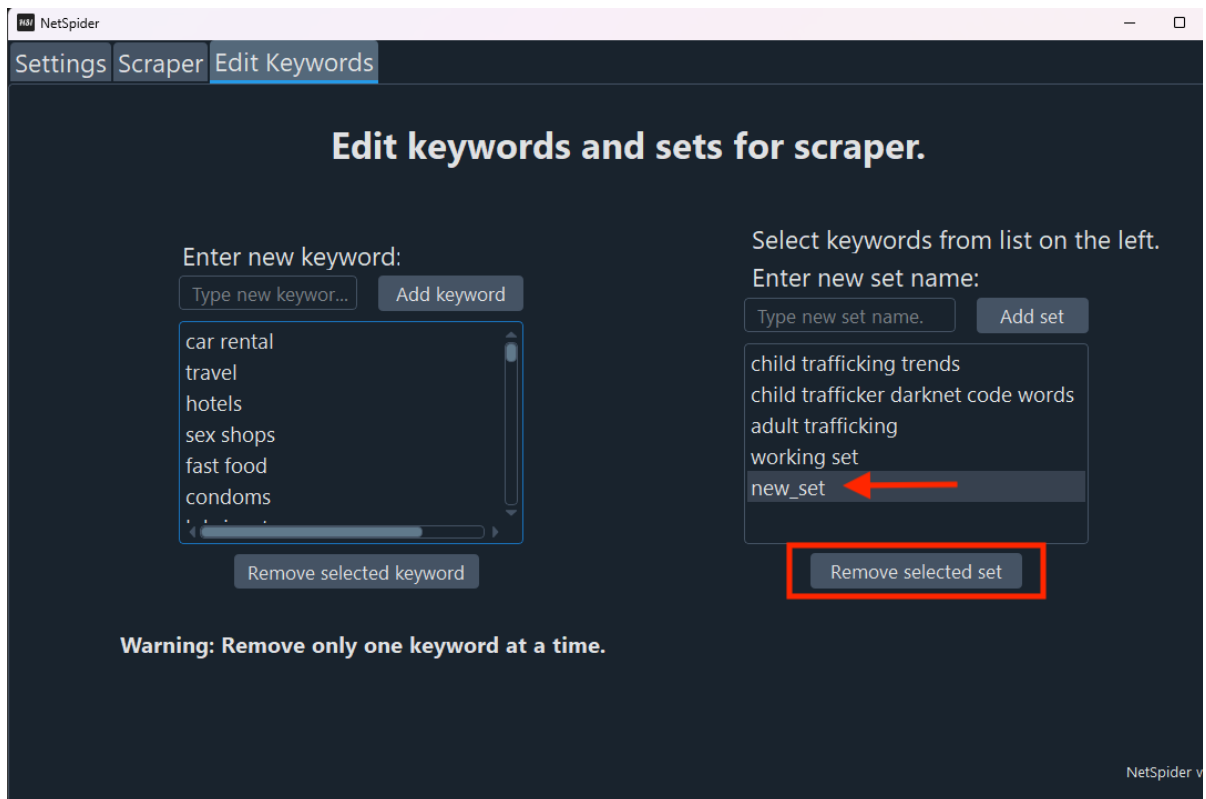
Add Set

 By clicking add set, the terms selected in the keywords list will be contained in the new set.



Remove Set

 By clicking remove selected set, the chosen keyword set in the list will be removed.



VIEW DATA

Once the data is collected from open-source platforms and analyzed using search criteria, this data is exported to a CSV file. This file can be opened using spreadsheet software such as Excel and formatted for better readability. Additionally, unique identifiers are assigned to the screenshots, which can be used as a reference while viewing the data in the spreadsheet.

View CSV File

- 1 After opening a spreadsheet software (i.e., Excel), click the **data** tab, **Get Data/Query**, and **From CSV**.

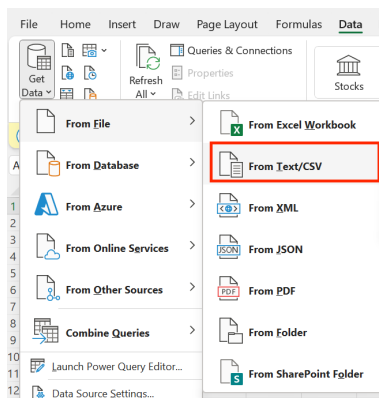


Figure 1 - Excel 2023

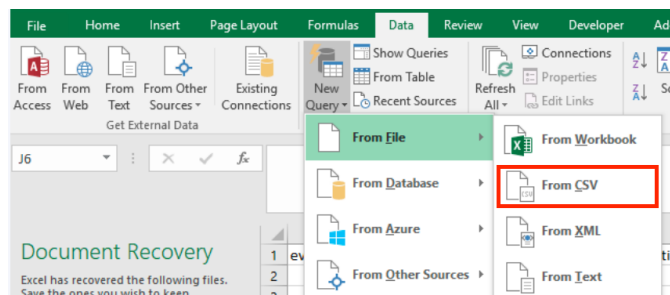
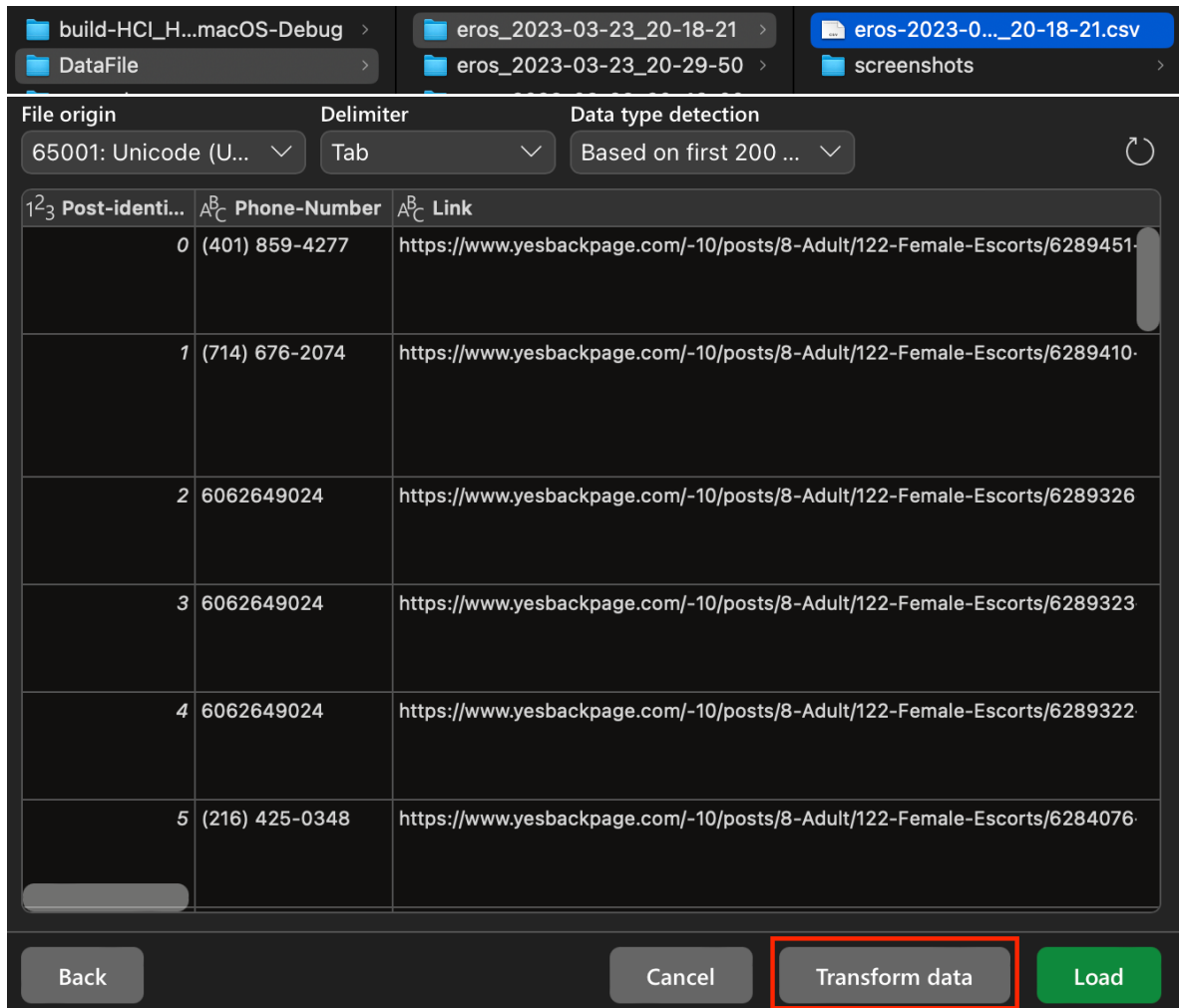


Figure 2 - Excel 2016


- 2 Next, navigate to the chosen folder that contains the scarper data results (if unsure, copy file path from the Settings tab within the NetSpider application). Then select the **website_timestamped** folder that correlates to the desired search. After, select the CSV file and click **transform data** for it to format correctly.

Format:

- File origin: Unicode (UTF-8)
- Delimiter: Tab



View Screenshots

-  Navigate to the chosen folder that contains the scraper data results (if unsure, copy file path from the Settings tab within the NetSpider application). Then select the **website_timestamped** folder that correlates to the desired search. After, select the **screenshots** folder which contains all the screenshots with the unique identifiers for that specific search.

