

# מדריך Web Crawler:

## מסך ראשי:

The screenshot shows the 'Web Crawler' application window. It contains a 'URL:' field with 'https://www.google.com' entered. Below it are 'Recursion depth:' and 'Number of threads:' fields, both set to '3'. At the bottom are two buttons: 'Start new scan' (grey) and 'Continue the last scan' (green). Red arrows point from Hebrew text boxes to these fields and buttons.

Web Crawler

URL:

Recursion depth:

Number of threads:

לתיבה זאת הכנס  
מספר הדפים  
שתרצה לסרוק בו  
זמנית

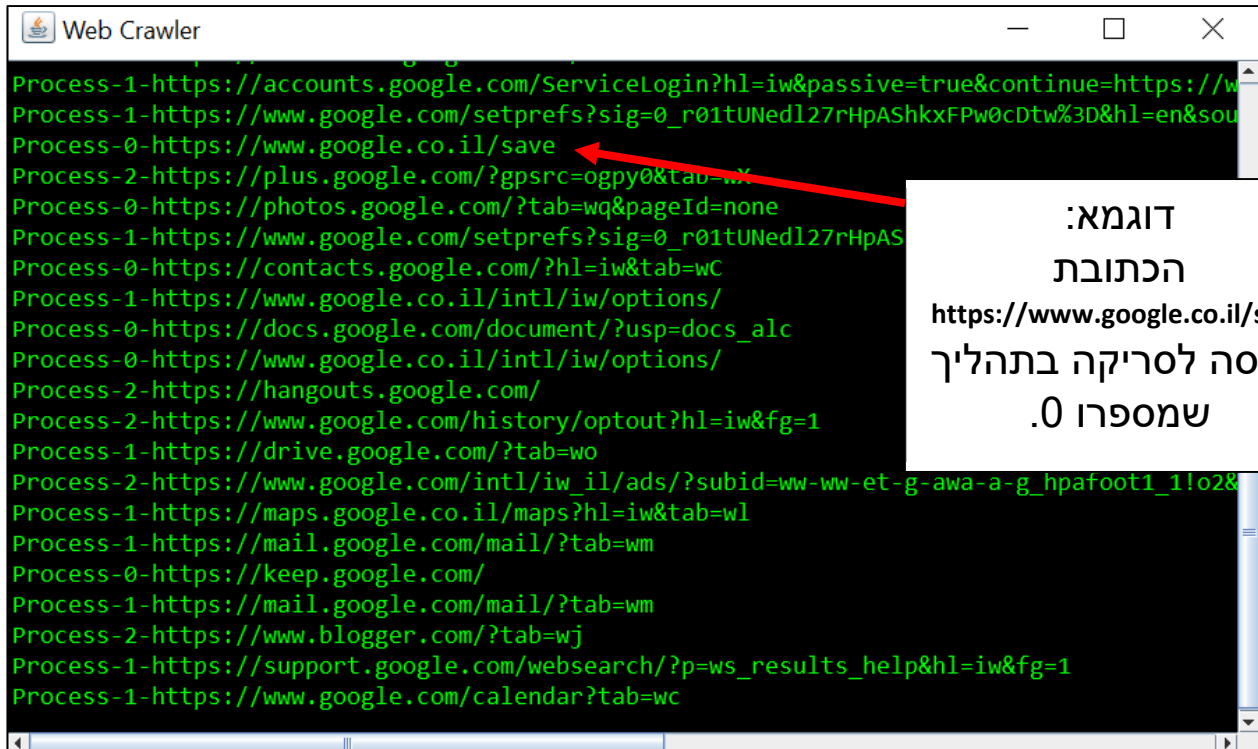
התחל סריקה חדשה  
ע"פ הנתונים  
שהוכנסו למעלה

המשך את הסריקה  
הקודמת ע"פ  
הנתונים שהוכנסו  
בעבר.  
(זמין רק כאשר  
הייתה סריקה  
שהופסקה)

לתיבה זאת הכנס  
את כתובת  
ה Root URL  
(כולל פרוטוקול)  
אותה נרצה לסרוק

לתיבה זאת הכנס  
את עומק הרקורסיה  
של הסריקה

## מסך ה Terminal Log :



```
Process-1-https://accounts.google.com/ServiceLogin?hl=iw&passive=true&continue=https://w
Process-1-https://www.google.com/setprefs?sig=0_r01tUNedl27rHpAShkxFPw0cDtw%3D&hl=en&sou
Process-0-https://www.google.co.il/save
Process-2-https://plus.google.com/?gpsrc=ogpy0&tab=wl
Process-0-https://photos.google.com/?tab=wq&pageId=none
Process-1-https://www.google.com/setprefs?sig=0_r01tUNedl27rHpAS
Process-0-https://contacts.google.com/?hl=iw&tab=wc
Process-1-https://www.google.co.il/intl/iw/options/
Process-0-https://docs.google.com/document/?usp=docs_alc
Process-0-https://www.google.co.il/intl/iw/options/
Process-2-https://hangouts.google.com/
Process-2-https://www.google.com/history/optout?hl=iw&fg=1
Process-1-https://drive.google.com/?tab=wo
Process-2-https://www.google.com/intl/iw_il/ads/?subid=ww-ww-et-g-awa-a-g_hpafoot1_1!o2&
Process-1-https://maps.google.co.il/maps?hl=iw&tab=w1
Process-1-https://mail.google.com/mail/?tab=wm
Process-0-https://keep.google.com/
Process-1-https://mail.google.com/mail/?tab=wm
Process-2-https://www.blogger.com/?tab=wj
Process-1-https://support.google.com/websearch/?p=ws_results_help&hl=iw&fg=1
Process-1-https://www.google.com/calendar?tab=wc
```

דוגמא:

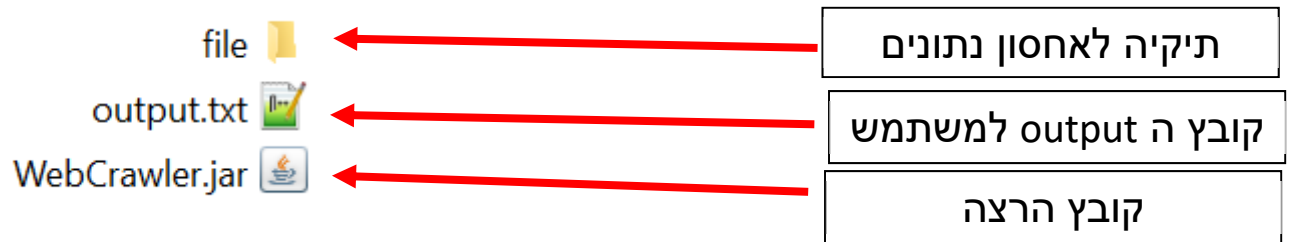
הכתובת

<https://www.google.co.il/save>

נכנסה לסריקה בתהליך

שמספרו 0.

## תיקיות וקבצים:



בתיקיית file->page קיימות כל התיקיות המכילות את הדפסים שנסרקו וכתובות ה URL שיצאו מהם ע"י מבנה מסודר.

marthastewartweddings.com	📁
midwestliving.com	📁
more.com	📁
myrecipes.com	📁
mywedding.com	📁
nytm.org	📁
parenting.com	📁
parents.com	📁
people.com	📁
pinterest.com	📁
rachaelraymag.com	📁
realsimple.com	📁
serpadres.com	📁
siempremujer.com	📁
subscription-assets.timeinc.com	📁
techcrunch.com	📁
theverge.com	📁
thisisinsider.com	📁
travelandleisure.com	📁
twitter.com	📁
usatoday.com	📁
venturebeat.com	📁

## קובץ ה OUTPUT:

Deep: 3 || Ratio: 0.8052631578947383  
<https://www.instyle.com/look-of-the-day/2018-11-09#3391141>  
Deep: 3 || Ratio: 0.7909836065573739  
<https://www.instyle.com/>  
Deep: 3 || Ratio: 0.8052631578947383  
<https://www.instyle.com/look-of-the-day/2018-11-09#3391135>  
Deep: 3 || Ratio: 0.7909836065573739  
<https://www.instyle.com/>  
Deep: 3 || Ratio: 0.852017937219733  
<https://www.instyle.com/reviews-coverage/music>  
Deep: 3 || Ratio: 0.6600000000000004  
<http://www.businessinsider.my>  
Deep: 3 || Ratio: 0.7219251336898412  
<https://www.instyle.com/tag/who-won-fashion-today>  
Deep: 3 || Ratio: 0.7386934673366853  
<https://www.instyle.com/news>  
Deep: 3 || Ratio: 0.7909836065573739  
<https://www.instyle.com>  
Deep: 3 || Ratio: 0.9421052631578967  
<http://www.mywedding.com/>  
Deep: 3 || Ratio: 0.0  
<http://www.allpeoplequilt.com/>  
Deep: 3 || Ratio: 0.5519480519480516  
<https://www.instyle.com/lifestyle/tampon-tax-meaning-real-cost>  
Deep: 3 || Ratio: 0.0  
<http://www.midwestliving.com/>  
Deep: 3 || Ratio: 0.8674698795180763  
<https://www.instyle.com/videos>  
Deep: 3 || Ratio: 0.8037735849056603  
<https://www.instyle.com/celebrity>

דוגמא לכתובת:  
<https://www.instyle.com>  
עומק הרקורסיה שלה 3,  
יחס הכתובות שמוכלות  
בה בעלות תחום זהה הוא  
~0.8