Plan de Becarios en Seguridad Informática
Proyecto Módulo 4 - Manual de Usuario

Rafel Diez Gutierrez González
Andrea Itzel González Vargas

August 16, 2018

El presente proyecto realiza un *scraping* de los buscadores web Bing, Google, DuckDuckGo y Board Reader, con la opcion de hacer busquedas en Pastebin. El lenguaje de programacion utilizado es Python 3.

1 Instalacion de dependencias

Para hacer la instalacion de las dependencias se debe de correr el script `scrappy/dependencias.sh`, tambien existe la posibilidad de ejecutar el comando

```
$ python3 scrappy/setup.py install
```

lo cual iniciara la instalacion de dependencias.

2 Ejecucion

Para ejecutar el programa se debe de utilizar el siguiente comando:

```
$ python3 scrappy.py {<busqueda> [opciones] | [busqueda] {f --filetype |
         s --site | h --help | p --ip | u --inurl | m --mail} [opciones]
```

Donde las opciones tienen el siguiente significado:

```
-h, --help            Muestra mensajes de ayuda
-v, --verbose         Indica que se utilice el modo verboso.
-n NUM_RES, --num-res=NUM_RES
                      Numero de resultados por busqueda
-b BUSCADORES, --buscadores=BUSCADORES
                      Se especifica el buscador a utilizar
-N, --no-params       Excluye los parametros GET, haciendo unica cada
                      busqueda
-r, --regex           Se pueden usar Expresiones Regulares
-d, --domains         Indica que se impriman unicamente los dominios en los
                      reportes
```

```
-U USER_AGENTS, --user-agents=USER_AGENTS
                      Archivo con agentes de usuario a utilizar, separados
                      por un salto de linea
-P PROXIES, --proxies=PROXIES
                      Archivo con proxies a utilizar, separados por un salto
                      de linea
-F FORMATO, --formato=FORMATO
                      Especifica el formato de salida
-i INTERVALO, --intervalo=INTERVALO
                      Se especifica el intervalo de tiempo por busqueda
-p IP, --ip=IP        Se especifica la(s) IP(s) de busqueda, separadas por
                      una coma
-m MAIL, --mail=MAIL  Busca correos electronicos en los dominios
                      especificados, separados por una coma
-f FILETYPE, --filetype=FILETYPE
                      Se busca por los tipos de archivo especificados,
                      separados por una coma
-s SITE, --site=SITE  Se busca por los sitios web especificados, separados
                      por una coma
-e EXCLUDE, --exclude=EXCLUDE
                      Se excluyen los resultados que contengan las palabras
                      indicadas, separadas por una coma
-w EXACT_WORD, --exact-word=EXACT_WORD
                      Se buscan las palabras indicadas de manera exacta,
                      cada una va separada por una coma
-I INCLUDE, --include=INCLUDE
                      Se incluyen los resultados que contengan esa palabra
-u INURL, --inurl=INURL
                      Se buscan las palabras dentro de la url, separadas por
                      comas
-o OUTPUT, --output=OUTPUT
                      Nombre de los archivos de reporte.
-t, --tor             Indica que se haga uso de tor para hacer las
                      peticiones (junto con otros proxies, si se utiliza la
                      opcion -P).
```

El parametro `<busqueda>` es la busqueda principal a ser realizada, si esta tiene la forma `palabra1 + palabra2`, se indica que se haga la busqueda independiente del termino `palabra1` y el termino `palabra2`, es decir, el simbolo `+` es equivalente a un `OR`.

En el caso de utilizar la opcion `-r`, el parametro `<busqueda>` sera tomado como una expresion regular, cuyas expansiones comprenderan los terminos a ser buscados independientemente.

3 Ejemplo de ejecucion

Al ejecutar el siguiente comando se hace la busqueda del termino `gatos`, utilizando los buscadores Bing y Google. Unicamente se obtienen busquedas de archivos PDF en el dominio `unam.mx`, obte-

niendose aproximadamente 15 resultados por busqueda. Las peticiones se realizan anonimamente a traves de Tor. El reporte de resultados sera generado en tres formatos, `xml`, `html` y `txt`, los cuales tendran el nombre `busquedas.[html|xml|txt]`. Al estarse utilizando el modo verboso, se mostraran mensajes durante la ejecucion del programa.

```
$ python3 scrappy.py -f pdf -b bing,google "gatos" -F xml,html,txt \
      -s unam.mx -n 15 -t -o busquedas -v
```