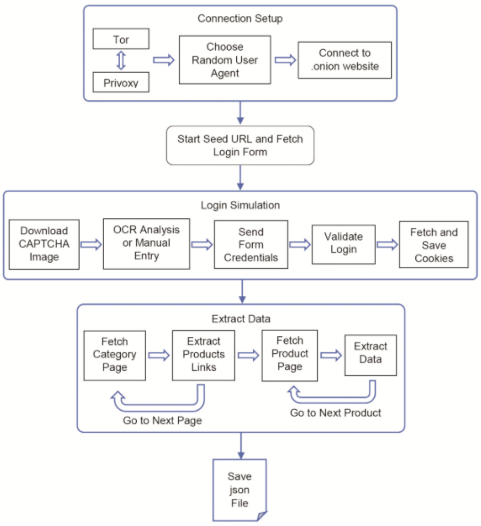
DOCUMENT MADE WHILE PREPARING PROJECT

Suggested System Components We can summarize the basic components of any dark web crawling system as follows: Crawling Space: The crawler starts from a list of websites of illicit activities. 1. We can depend on a number of sources like security resources - if available - (i.e. resources published by governments, or official nongovernmental organizations) , or electronic resources like indexes of Tor network. Links to dark websites can be attainable even on the surface web or by using generic search engines (like Google). The crawling space can expand by adding new links that the crawler finds on the retrieved webpages, which leads to other pages and so on. 2. Website Preprocessing: This includes processing access obstacles in case it needs membership registration, login validation, fetching and saving session cookies. 3. Storage and Analysis: Storing and analysing the retrieved webpages. Crawler needs to access a permanent storage space to save extracted webpages before analysing and processing them, and that can be achieved by two methods, according to the capabilities of the used equipment: connecting a database, or using a simple file storage system where pages are saved like separated files. Crawler is set by means of particular parameters - according to the servers and networks capabilities, and the blocking mechanisms used on some sites - to insure ease of access to the targeted sites, in addition to human assisted approach , and employing dynamic proxies. Such parameters are number of crawls for each site, number of connections to the site, allowed period for downloading a content, speed and timeout of a connection, and others. 4. Our Experimental System Used for Crawling and Data Extraction We have developed a system, Darky, using Scrapy1 (a programming library written in Python), as we provided it with a connection to dark websites on Tor network through Tor software integrated with Privoxy (a software for Virtual Private Network (VPN)) to insure the most possible security and anonymity of the crawler against those sites, by relocating the IP address. After establishing the Tor-Privoxy connection, we operate the crawler starting from the website URL, and it processes the login interface with credentials that we have created earlier on the website for the purpose. We designed the crawler especially for the website under study, according to its structure and the hyperlinks structure among its pages, i.e. we must customize a different crawler design for each website for different interfaces handling methods and different HTML structures.

Figure Illustrates system architecture:


Connection Setup
Tor
Privoxy
Choose Random User Agent
Connect to .onion website
Start Seed URL and Fetch Login Form
Login Simulation
Download CAPTCHA Image
OCR Analysis or Manual Entry
Send Form Credentials
Validate Login
Fetch and Save Cookies
Extract Data
Fetch Category Page
Extract Products Links
Fetch Product Page
Extract Data
Go to Next Page
Go to Next Product
Save json File