Automated Web Analysis and Archivation

Author: Ing. Tomáš Kocman
Supervisor: Ing. Libor Polčák Ph.D.

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Introduction

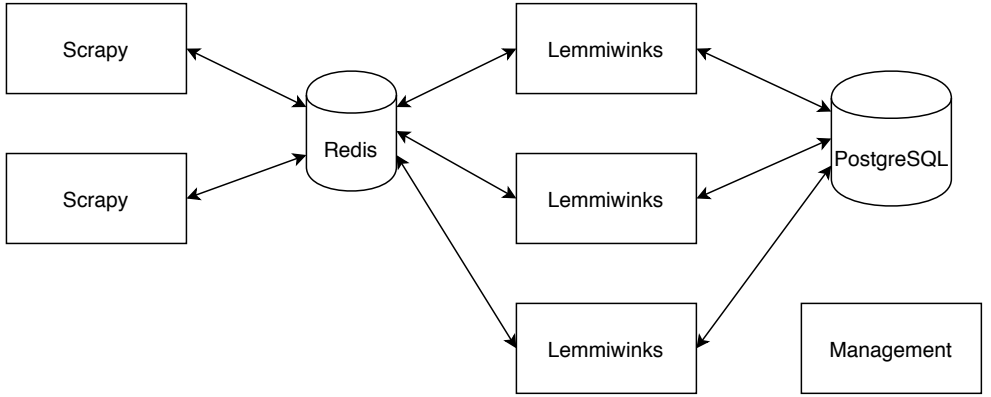
This thesis allows you to automate the archiving of web pages. But only those that meet certain rules. The data on them conforms to the defined regular expressions. The result of this work is a platform that can be configured in such a way to search and archive web pages according to different strategies. With the platform, you can easily visit all pages on the site automatically and if the site meets the defined rules, the platform will back up them. In the area of cybercrime, for example, investigators know the websites where the perpetrator was conducting crime. Then they can use the platform to find evidence.

Investigation of cybercrime

Cybercrime cannot be considered a mere hypothetical phenomenon. Attacks are taking place on an increasing scale not only to ordinary Internet users, or individuals but also to entire states (for example, to manipulate the political situation). Today, however, much of the attacks and incidents, often described in media and presented as cybercrime, can be described as the use of cyberspace by terrorists. So far, terrorist organizations are unlikely to have the capacity to carry out cyberattacks with serious implications. However, it is not difficult to purchase these capacities in the form of a service. For example, the Islamic State was able to carry out cyberattacks (but not sophisticated) that other terrorist organizations were unable to carry out for a long time.

Evidence collection platform

Scrapy is a library for collecting and analyzing web data. Allows dynamic configuration of spiders and pipelines through which downloaded data flows and at that point can be performed different operations.



Redis is a distributed queue that operates only within memory independently of other platform components. There is only one database type list that serves as a queue. Scrapy enqueues data and Lemmiwinks consumes data from the queue. All HTML documents together with metadata are queued.

Lemmiwinks is a web archiver that stores output archives in the MAFF format. The Redis process retrieves entire HTML documents and recursively archives them. Within the platform, there may be more instances of the Lemmiwnks process as the Redis queue can fill much faster than Lemmiwinks can consume data.

PostgreSQL is a persistent database that assigns paths to their saved MAFF archives. The database schema is enriched with metadata related to the archived page.

Management serves as an API for the entire platform. Through this REST API, it is possible to manipulate individual parts of the platform. For example, start or end Scrapy or Lemmiwinks processes and retrieve a MAFF location according to the query.