# Crawler 1: Infosecinstitute

URL: <https://resources.infosecinstitute.com/topics/capture-the-flag/>

Notizen:

* Extraction-Tag in der HTML
  + Commands stehen zwischen den Zeichen ‘<< ‘ und ‘ >>‘
* Duplikate werden verhindert, indem:
  + jeder individuelle Challenge Name und
  + jede individuelle Challenge URL in Liste gespeichert werden, falls diese so noch nicht in der Liste enthalten sind

🡪 Nur unique Challenges werden für die Extraktion betrachtet

* Filterkriterien, um nur CTF-Challenges zu extrahieren:
  + URL muss ‘walkthrough‘ enthalen
  + URL muss zusätzlich entweder ‘ctf‘ oder ‘capture-the-flag‘ enthalten

🡪 Nur CTF Challenges werden für die Extraktion betrachtet

# Crawler 2: Hackingarticles

URL: <https://www.hackingarticles.in/ctf-challenges-walkthrough/>

Notizen:

* Sicherstellen, dass nur relevante Inhalte für die Extraktion betrachtet werden:
  + Nur folgende Webseiten betrachten:
    - URL enthält ‘hackthebox‘
    - URL enthält ‘tryhackme‘
    - URL enthält ‘vulnhub-walkthrough‘
    - URL enthält ‘hack-the-box’
    - URL enthält ’ctf-challenge’
* Commands sind je nach „URL“ unterschiedlich im HTML gekennzeichnet
  + Bei HackTheBox und TryHackMe in ‘span‘
  + Bei Vulnhub-walkthrough, hack-the-box und ctf-challenge in einer ’pre class’