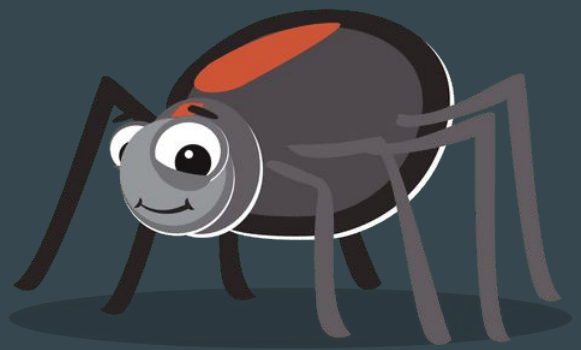
Writing Web Crawlers

Rebecca Bui

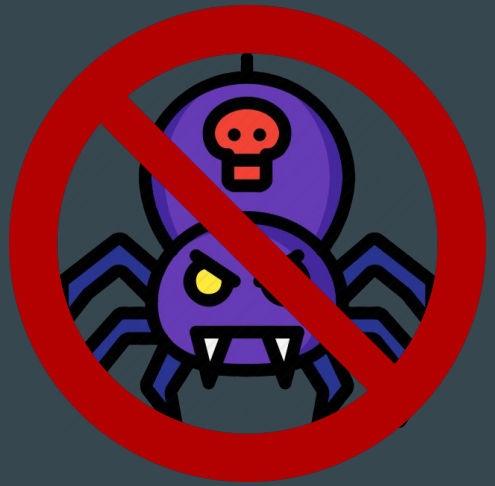


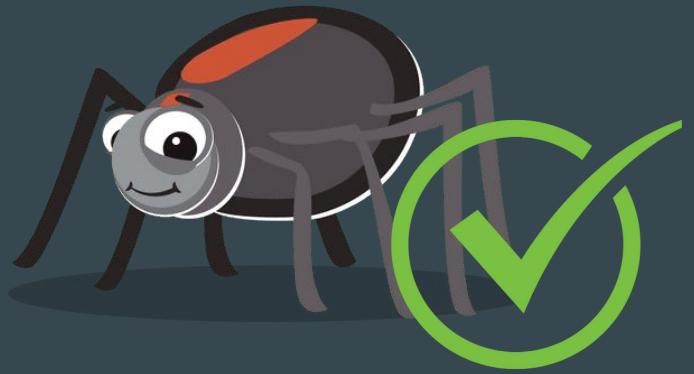
What is a web crawler?

- Program/ Script made to manipulate web data from URLs in a recursive manor
- Usually involves some sort of metadata about the pages

Not all Spiders are Equal:

1. Respect the robots.txt file
2. Never degrade the performance of a website
3. Include your contact information as a UserAgent





Presentation Goals:

1. To be a Good Spider
2. To Not Get Blacklisted
3. To Write a Simple Functional Web Crawler