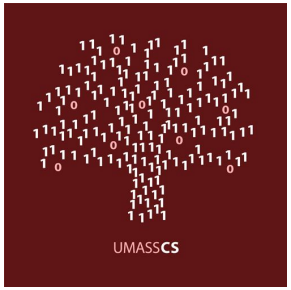


CS653 PROJECT

WEBSITES VS. WEBCRAWLER

By: Divyesh Harit

College of Information & Computer Sciences,
UMass Amherst

Introduction

- Innocuous web crawlers are used by search engines, but can be malicious
- Headless browsers are used for automated testing, but can be used for DDoS attacks
- So how resilient are top websites against these two approaches of access?

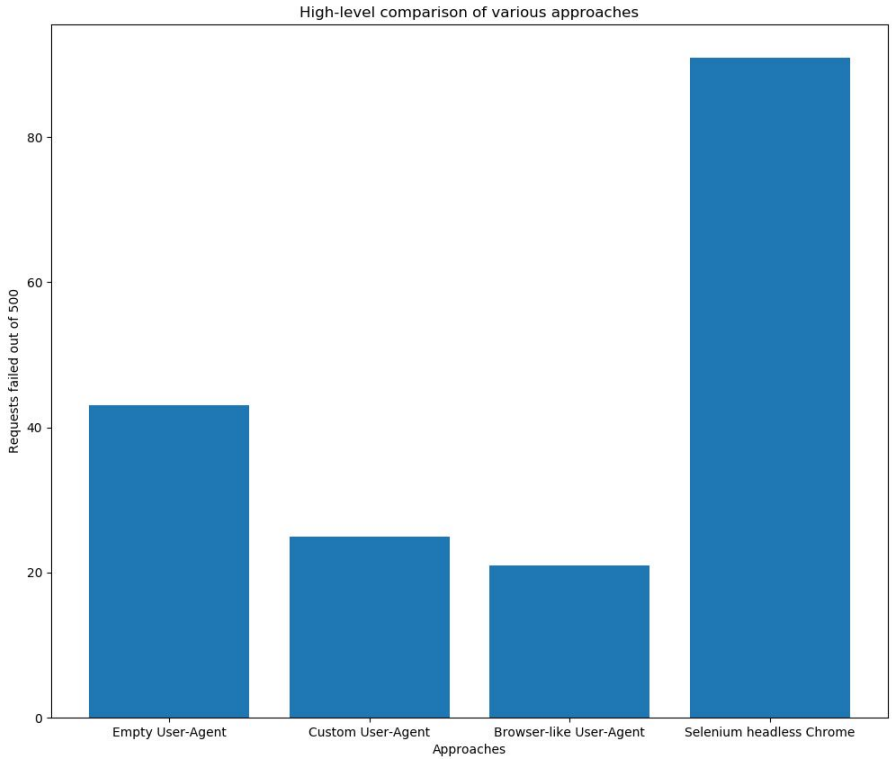
Research Questions

- Which approach performs better? Why?
- Any trend of blocking/non-blocking between the two approaches?
- Which kind of errors are the most frequent? Why?
- What can we do to reduce these errors?

High-level Methodology

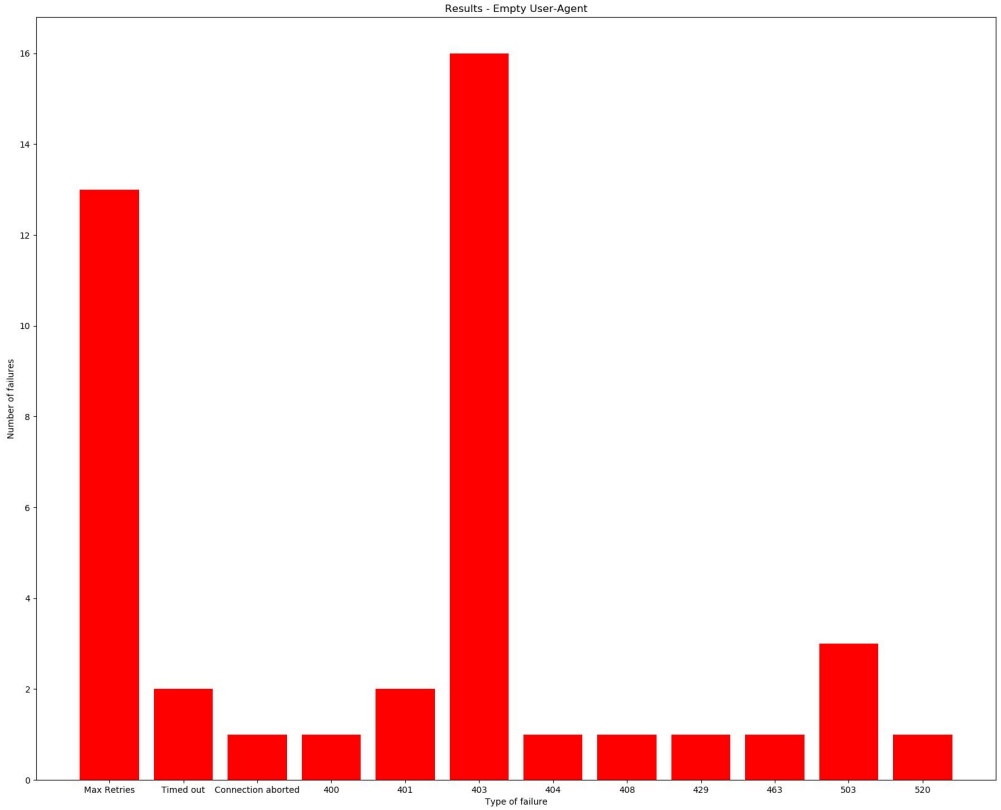
- Extract top 500 websites list from Alexa
- Create a crawler that sends HTTP GETs to these websites using requests
- Vary the “user-agent” field
    i) Empty user-agent
    ii) Custom user-agent
    iii) Browser-like user-agent
- Use Selenium Headless Browser to access the same websites
- Record response, plot and compare/contrast

Results: Comparison of approaches



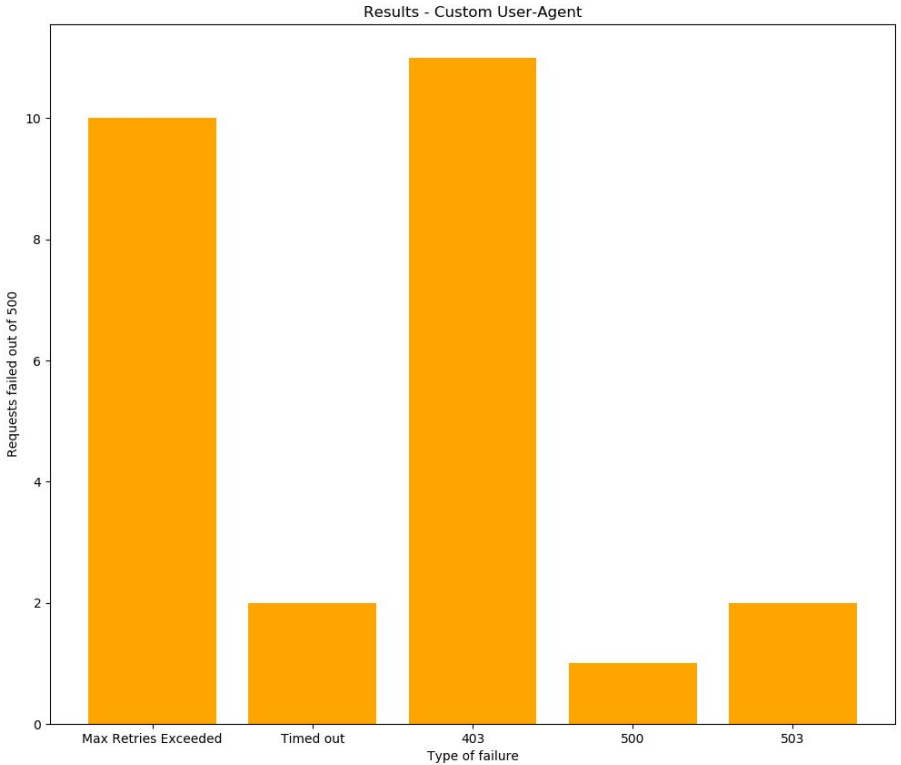
- Empty user agent performs the worst out of the variations of 1st approach
- Simulating browser user-agent performs very well
- Our custom user-agent is neither here nor there
- Selenium headless chrome performs very poorly, failing for ~⅕ websites
- Selenium is super slow. Takes 3-4 hrs for the total experiment.
- Each of the 1st approach variations take ~10 mins.

Results: Empty User Agent



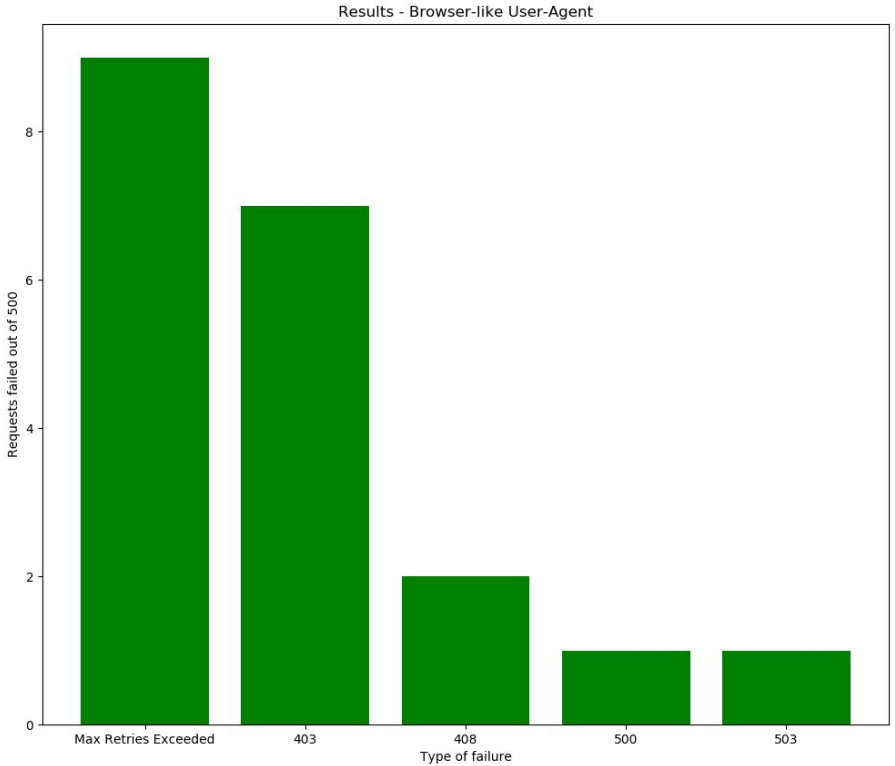
- HUGE variety of errors!
- One weird 463 [Invalid Media Name]: roblox.com
- Max Retries Exceeded: Failed to establish connection even after 5 attempts
- Mostly 403 and Max Retries Exceeded
- Highest proportion of 403 out of all approaches

Results: Custom User Agent



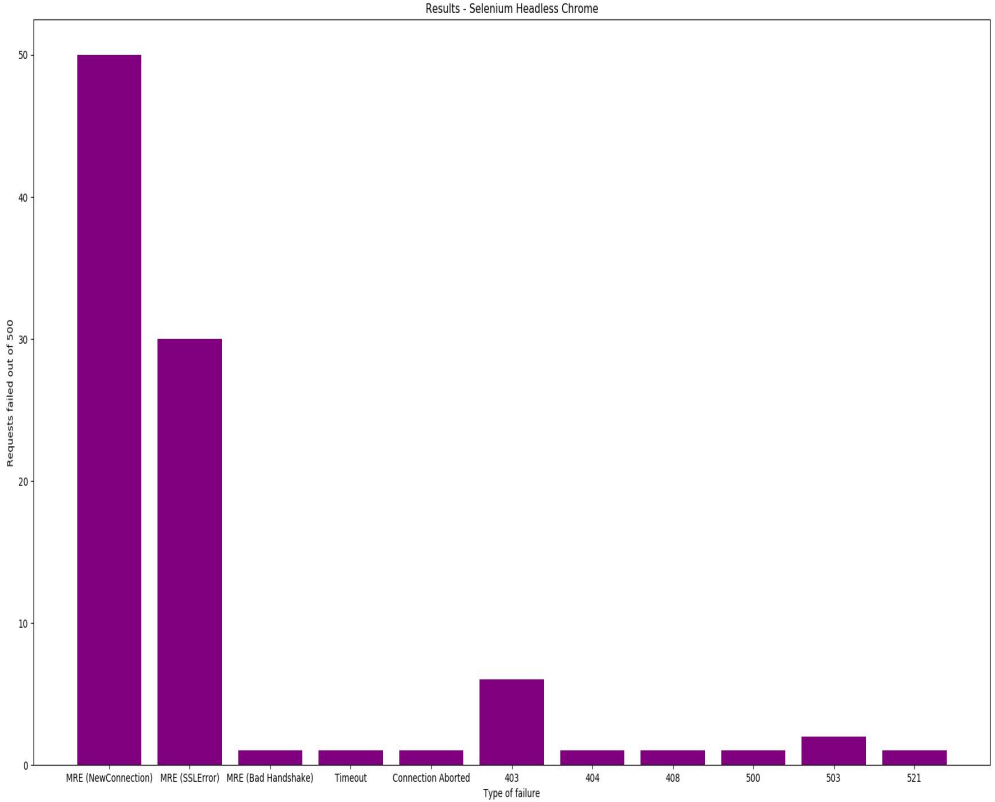
- “DivBot / https://sites.google.com/view/web-crawler-cs653”
- Better performance than Empty user agent
- Less types of errors
- Here too, 403 and Max Retries Exceeded dominate

Results: Browser-Like User Agent



- “Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3202.94 Safari/537.36”
- Best results
- ..although still not as good as an actual browser
- Only crawler approach with more Max Retries than 403s

Results: Selenium Headless Chrome



- Way too many Max Retries - we even got some variety this time!
- Makes up ~90% of all errors
- All sites that give a 403 here also gave a 403 for other approaches
- Turns out they’re just unavailable, even from a browser

Possible future work

- Try and avoid Max retries:
  - Sleep between requests
  - Use Chinese/Russian IPs to send requests
- Explore HTML of responses
- Look at headers of responses

Acknowledgements

- Prof. Phillipa Gill for invaluable feedback on progress report
- Keen Sung for helping out with Selenium
- The Internet, without which all this would not be possible :)