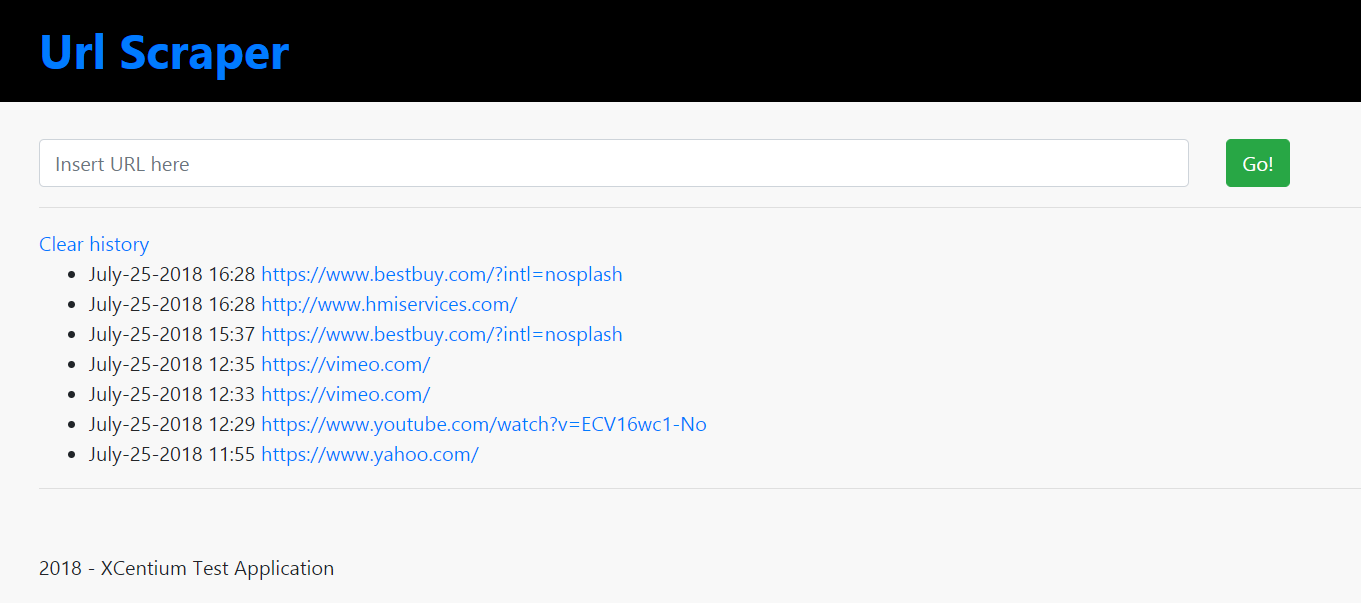
Url Scraping Tool



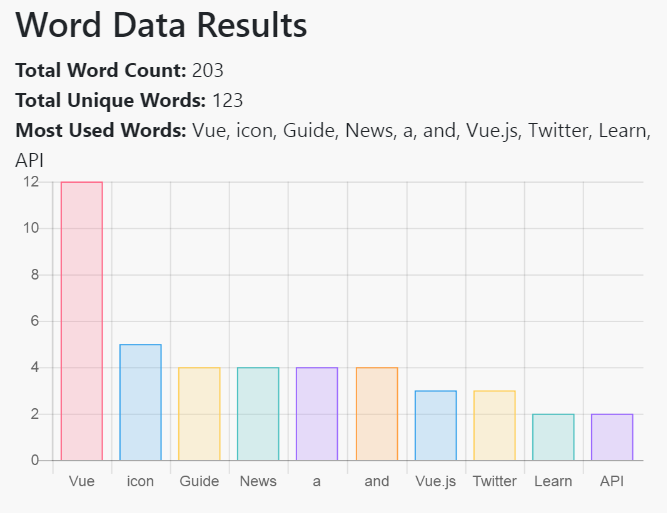
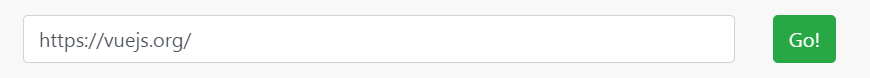
# Features

* Image Retrieval
* Text Retrieval
* Recorded History
* Execution Failure Logging

# How It Works

From the homepage you’ll see a single field for user input and below that you’ll find links to display the results of previous executions.

* To run a new scrape
  1. Start at the homepage and enter a valid url into the textfield.
  2. Click Go! You’ll be redirected to the results page.
* To view a previously executed scrape
  1. Start at the homepage and choose one of the bullet-pointed links from the history section. *Note: There may be none shown if you haven’t yet run any scrapes OR if you’ve recently cleared the history.*
  2. You’ll be redirected to the results page. View the data there.
* To clear history data
  1. From the homepage, click the Clear history button
* To get back to the home page
  1. Click the Url Scraper title at the top



# History Log Location

All successful scrapes are serialized into their json respresentations and recorded into the **ScrapeHistory.txt** file. *Note that the file may not actually exist on the server yet if no successful scrapes have been recorded OR if the history has been cleared through the dashboard.*

# Execution Failure Logs

In the event of failure for any reason, The Url Scraping Tool keeps execution logs on the server for reporting and debugging. Logs can be found at the root folder of the site in a File named “**LogFile.txt**”. *Note that the file may not actually exist on the server yet if no errors have been recorded.*

# Development Notes

.Net Framework 4.7  
JQuery 3.3.1  
Bootstrap 4.1.2  
Chart.js 2.7.2

**Tested On**:   
Chrome v67.0.3396.99,   
FireFox v61.0.1,   
Internet Explorer v11.09600.19078

# Known Issues

* There have been a few times where the Word Data Results grid doesn't load. A simple refresh fixes it but higher consistency would be preferable. Best assumption is that it has to do with the timing of JS script loads.
* </form> tag is detected as a text node. Special filters have been put in place to remove these from results but this is a sign that there may be more irregularities in text results. While this is not a visible defect to the end user, the algorithm requires further
* Gif images with “data:<guid>” do not load in the carousel.
  + Ex: execute url “https://www.techwalla.com/”

# Possible Next-Phase Enhancements

* Increase speed and efficiency of scraping algorithm.
* Log to database using EF for more dynamic tracking.
  1. Attempt log to DB first
  2. If DB not available (service outage, out of rotation, etc..), log to temp server file
  3. On first availability, parse temp log file and store in DB
* Maintain webpage state for future possibility that we may want to scroll through pages.