

Web Page Analyzer

Andrew Pickner

AJ Jones

Jake Henson

Yuxi Liu

Fengyuan Zhang

Webpage Analyzer: What is it?

Enter one or more webpages you want to analyze!

e.g. <https://www.nytimes.com/>



SUBMIT

This application will take in your chosen website and analyze the content of the page(s). This includes, but is not limited to, metadata about your chosen site, (how the content is shown, percentages) as well as specific analysis of the text!

Developed by: AJ Jones, Andrew Pickner, Fengyuan Zhang, Jake Henson, and Yuxi Liu for CSCI 3308, © 2019. All rights reserved.



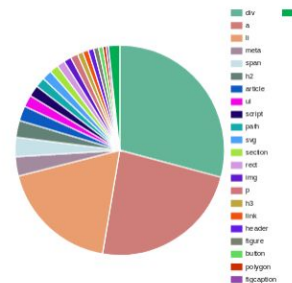
Site Description

The New York Times: Find breaking news, multimedia, reviews & opinion on Washington, business, sports, movies, travel, books, jobs, education, real estate, cars & more at nytimes.com.

Tag Breakdown

div:	304
ac:	245
li:	192
meta:	31
span:	31
h2:	27
article:	24
ul:	19
script:	18
path:	16
svg:	16
section:	14
rect:	13
img:	12
p:	12
h3:	9
link:	9
header:	9
figure:	8
button:	7
polygon:	5
figcaption:	4
style:	2
nav:	2
line:	2
g:	2
head:	1
circle:	1
iframe:	1
title:	1
body:	1
noscript:	1

HTML Tag Pie Chart



Who is it useful for?

Initially, we designed this with data scientists in mind

However, it can be useful for anyone who wants to know more about the sites they visit and the info about them. Displaying metadata and text analysis of a page can also be useful for just about anyone who's curious about the sites they visit!



How does it work?

1. Front-end site to allow the user to enter website URLs
2. API calls to Heroku server using NodeJS
3. Python web-crawler gets data from the requested URL
4. Population of the page with python API call + ChartsJS
5. Changeable views of forms/data with buttons on the Results Page

Tools we used:



VCS Repository: GitHub Repository

5/5 ★★★★★



Project Tracker: Github Projects

4/5 ★★★★★



IDE: Atom

5/5 ★★★★★



Deployment Achiever: Heroku

2/5 ★★☆☆☆



Database: PostgresSQL

2/5 ★★☆☆☆



FrameWork: NodeJS

4/5 ★★★★★



Languages: Python, HTML, CSS, SQL

5/5 ★★★★★



Pug

2/5 ★★☆☆☆

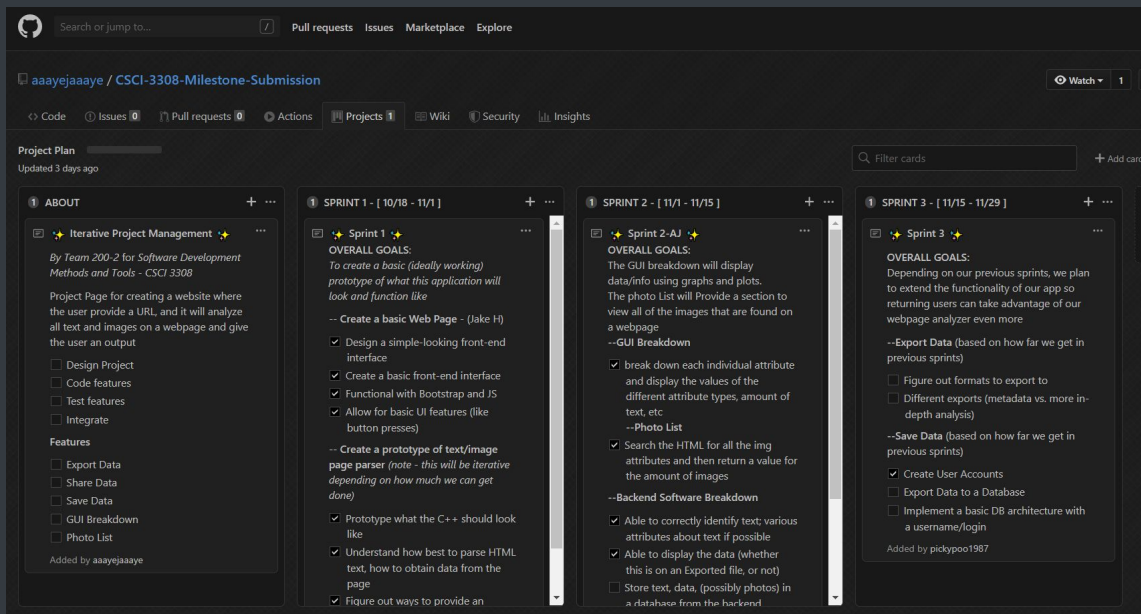
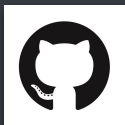


Methodologies: Agile + Iterative

4/5 ★★★★★

Workflow:

- GitHub Projects as a project planner to make sure we kept track of our sprints
- Discord to message each other about the project



Challenges

Setting up the Heroku backend: Heroku had to be created with the correct file structure and importing a ***ton*** of dependencies for python.

Integrating the individual components: Connecting the HTML to the Python script and connecting the database to the webpage.

Rewriting HTML: We ended up switching our HTML to Pug

SQL Database: We had issues creating/populating SQL databases both locally and on Heroku

Middle-Layer Integration: We had difficulty merging all of the different pieces together

Let's see a live demo!

The project:

<https://website-analysis-csci3308.herokuapp.com/>

Our github:

<https://github.com/aaayejaaaye/CSCI-3308-CodeNStuff>