team m22

project coordinator/leader Michael Neill Hartman mnh5@illinois.edu

Progress Report

custom scraper / information aggregator

The progress report should give us an idea of how you're implementing your proposal. It should answer 3 main questions:

1) Which tasks have been completed?

I have generated prototypes for a few typical use-cases:

- I can now query if items are in stock at amazon or newegg given the url of the item
- Partial: I can now pull the top 10 news headlines from google news
- Partial: I can now pull crypto prices and exchange rates from coinmarketcap
- Partial: I can now list all images contained on a site, with some sites
- Partial: I have started prototyping the configuration file format
- Partial: I have started prototyping the HTML/Javascript interface

2) Which tasks are pending?

- pulling crypto prices and exchange rates from sites other than coinmarketcap
- pulling the top 10 news headlines for news sites other than google news
- General list all images contained on a site
- Finalized configuration file format
- Working HTML/Javascript interface

3) Are you facing any challenges?

Realizing that many sites interpret an HTTP GET request from a browser differently than a plain HTTP GET request and identifying how to make them the same took some time.

I have been having a hard time attempting to make generic implementations across multiple sites. It is not very difficult to generate a scraping and parsing method for each site, but more difficult to generate one that works across dissimilar sites. While I don't need to

create fully generic implementations to fulfill my use cases, it is only natural to look for ways to do so while prototyping.

Some sites use content delivery networks with very odd resource names. This causes a lot of problems trying to identify all the images on a site.

Amazon suggests similar items when an item is out of stock. This makes it more difficult to determine if an item is in stock or not. This took some time to resolve.

Many of the sites which have crypto prices and exchange rates make parsing the values extremely difficult, but offer an api. I have only found one site that I could reasonably create a method to parse the values from.