Web Crawling and Community Review to Prevent Misleading Links

CSU Bakersfield
School of Natural Sciences, Mathematics, and Engineering

Adam Arreguin, Adrian Gutierrez, Joel Staggs, Kenny Taylor

Project Summary

Our Goal is to promote honest web practices and spot manipulative "clickbait" webpages and warn users of dishonesty. To accomplish this, we've developed an interconnected software infrastructure consisting of several components.

Automated Web Scraping Server

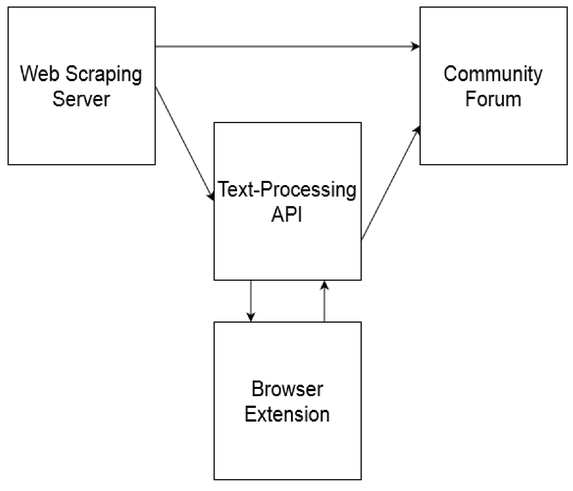
Users can retrieve or request articles scraped by the server by searching the domain name of an article's source. The articles are ranked by a page rank algorithm that returns a reputation score by taking the number of links an article contains and determining a score based on the quality of the links; this is done through several iterations to give articles a much more precise score.

How it Works

The **text-classification API** uses sentiment analysis to detect clickbait using article titles, titles are gathered from an installable **Web Browser Extension**, or through our **Automated Web Scraping Server**. Users can use our browser extension or **Community Forum** to read or comment on article reviews.

Text-Classification API

Our text-classification API is responsible for analyzing web article titles, and sending the results to the requesting service, formatting the response according. The API's core algorithm is a Multinomial Naïve Bayesian classifier. This program will classify articles based on probabilities calculated for each word within the article's title. This program uses previously classified articles to make its assumptions, and its results are stored and distributed to other project components when requested.



A simplified overview of our project

Our browser extension's pop-up information box.

Browser Extension

Hover over a link to receive an inline information box that will provide information about our assessment of the article. Right click on a link and select "Explore Link" from the context menu to view further details, post and read comments in real time using WebSocket comment server, or preview the site from within the sidebar.

Community Forum

Users can sign into our forum to view the article's we've archived and scored, request new articles to be scraped by our automated web scraping server, subscribe to other users to track their comments, and subscribe to domains to see when new articles are reviewed.

Conclusion

Our project has given us invaluable experience in maintaining, collaborating on, and implementing features on a large, distributed project with many interworking parts.