Intelligent Web Page Information Retrieval and Link Browsing

Documentation

CS 410 Final Project
Shi Yao Liu (shiyao3)
Xinrui Zhu (xinruiz4)
Ryan Cedzo (cedzo2)

Motivation

Modern web browsers' default search capabilities are limited to exact string matches within the current page. In a typical browsing scenario, a user would open links to examine related information. One such platform where this use case is very common is Wikipedia. Since Wikipedia has vast collections of articles, it is very common for articles to reference each other as well as external authoritative sources. As a result, Wikipedia articles often contain vast collections of links. Since they are always scattered throughout the page, it would be very helpful to consolidate all relevant links that appear on the page. Of course, Wikipedia is not the only platform where this problem arises. A feature like this would be useful on many sites.

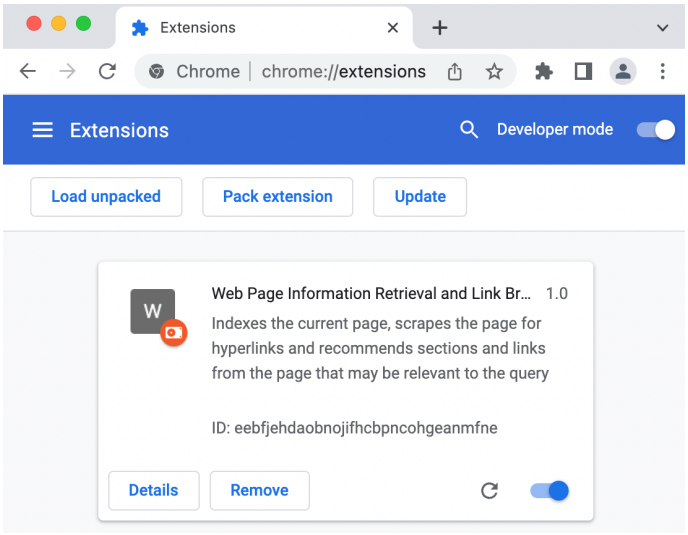
Description

Our team's project is a Chrome extension that collects and presents links that appear on the webpage relevant to a given query. The extension appears as a typical web browser search box. A text field is provided to enter a query, and a button to trigger the search. The extension collects all links from the current page and scrapes the content of the linked pages. BM25 is then used to rank the links based on their corresponding pages. Relevant links are then provided to the user for examination. The project was specifically tailored to Wikipedia since it is the ideal environment for testing such a tool.

Installation

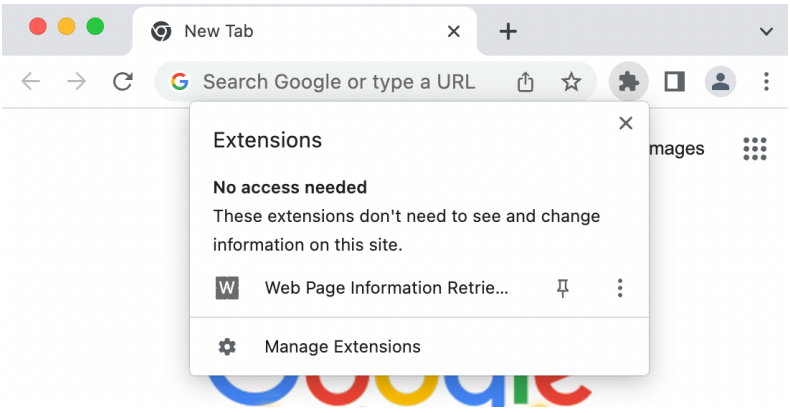
Installation is the same as any chrome extension. Download the source: https://github.com/cedzo2/BrowserExtension.git. Open Chrome and enter `chrome://extensions` in the URL field and press enter.

Enable `Developer Mode` and click `Load Unpacked` to bring up a navigation window. In the navigation window, navigate to the folder containing the source files for the extension. After selecting the folder, the chrome extension will be loaded and ready to be used.

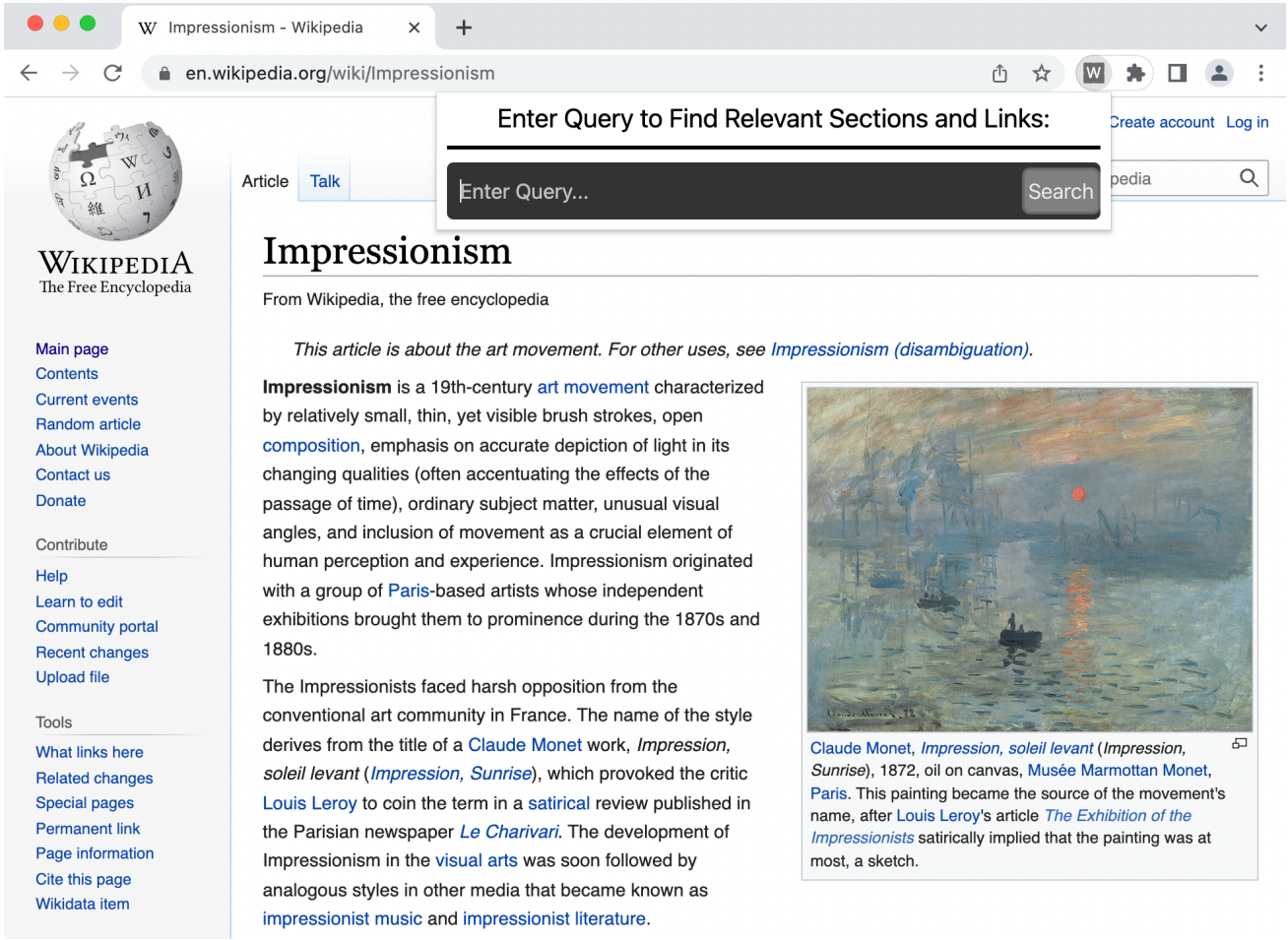


Usage

To open the extension, select the `Extensions` icon on the top-right corner of the window. The name of the extension is “Web Page Information Retrieval and Link Browsing”.



When the extension is opened, a search box appears. As you would with any search box, type in the query into the text field. Clicking the button triggers the search, and relevant links are presented to the user in a pop-up.



Implementation

When search is triggered, the extension first scrapes the webpage for links. Once all links have been collected, the text body is extracted from each. Each page is assigned a document ID, indexed, and IDF computations are performed. Using the user-defined query, BM25 is then executed on each of the pages to generate a score, discarding pages whose scores are zero. The results are then sorted and presented to the user.