CS653: Project Report
Wikipedia-Map

Vaibhav Nagar
14785

Abstract

Wikipedia map is an application for visualizing the connections between the wikipedia pages. User enters wikipedia topics in the application text-box, then server running in back-end, using Haskell, queries these topics to Wikipedia API and get the webpages and parse the webpages to get all the links. In the front-end, a single node is generated for each different entered topic which is then connected by the links present in its wikipedia webpage.

- Back-end: Haskell- API call, web-scraping or parsing, rendering data
- Front-end: Two methods for visualizing: Web-application using HTML5, CSS, JS and interactive GUI using Haskell only.

Why is it useful?
Wikipedia, which is web-based and the most popular free-content encyclopedia, contains links to other topics embedded in the content of a topic. This can be used to compare multiple articles and find their dependencies.
Also, one can test an interesting fact- "Wikipedia:Getting to Philosophy" which says that repeatedly clicking on the first link (non-parenthesized, non-italicized) in the main text of a Wikipedia article, eventually leads to Philosophy article.

Links:

- `https://luke.deentaylor.com/wikipedia/`
- `https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy`

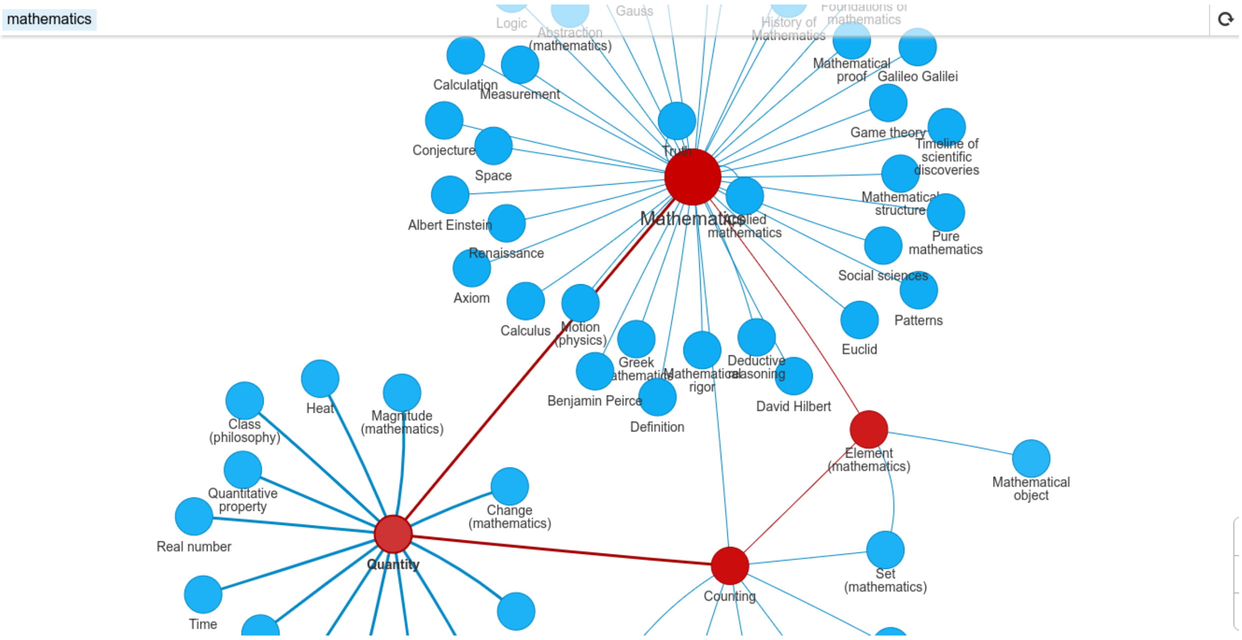

Figure 1: Wikipedia Map of "mathematics" page with the loop by clicking the first link successively

Implementation Details

User can provide wikipedia page names or choose “random” option from the web application (frontend), which then makes HTTP GET request to the my API server running in Haskell using Wai package. The API server provides several endpoints:

- “links” : Requires wiki page name as a query parameter and replies with the list of links in JSON format
- “pagename” : Replies accurate wiki page name from the page name provided in the query parameter. This works only when the given page name is an alias in the MediaWiki API.
- “random” : Replies random accurate wiki page name

The API server queries Wikimedia API to get the wikipedia articles, accurate page name or random page name in the JSON format.
The random page name and accurate page name are extracted by directly parsing the JSON data.
To get links from wikipedia page: Raw HTML text, received from Wikimedia API, is parsed and links are extracted. Citation links, parenthesized and external links are ignored. Parser also provides several other functionalities:

- Get first N wiki links from a wikipedia page
- Get all wiki links from first N paragraphs of a wikipedia page

Deviation from proposal

A CLI based application is also implemented in which user provides the two wiki page names (starting point and ending point) and the application successively traverses through the first non-parenthesized link in the main text of wikipedia page which begins from starting page until it reaches the ending page or the loop is detected.

Link To Code: https://github.com/vaibhavnaagar/wikipedia-map