Project B: Digital library search engine with Wiki-cards

Overview

The goal of this project is to build a web-based user interface (UI) for abstracts for electronic theses and dissertations (ETDs). The UI should have a landing page containing a text box for users to search for ETDs indexed by Elasticsearch. When a user clicks one item in the search result page (SERP), a summary page should show up and display metadata of an ETD, including its title, authors, year, university, advisor, discipline, and abstract. The keyphrases should be highlighted in the abstract text. These keyphrases were pre-extracted by a keyphrase extractor, which can map a keyphrase to a Wikipedia term. When the user hops its mouse on the highlighted phrases, a small window (hereafter Wiki-card) will pop out and display the first sentence in the keyphrase's explanation in Wikipedia. Users can also click the Wiki-card and go to the keyphrase's Wikipedia page.

The UI will implement other features such as user registration, log-in, reCAPTCHA, spell-check, auto-complete, customized list, star rating, feedback, download, etc. More details will be specified in milestone specifications.

Data

The students will be provided a dataset containing the following parts.

1. Metadata of 500 ETDs, including the title, author, year, academic discipline, advisor, university, and degree in CSV format. Each ETD has a unique ID.
2. The PDFs of these ETDs, named using the unique ID of each ETD.

Infrastructure

The students will build the web applications on a virtual machine in the CCI academic environment. The search function should be powered by Elasticsearch. The user information, metadata and pre-extracted keyphrases should be stored in MySQL.