

Team GAASK members are: Angela Mei (amei5; captain), Adrienne Caby (acaby2), Sruthi Jaganathan (sruthij2), and Kristie Wong (kw32), and Grant Yang (guoyong2). The topic our group has chosen is intelligent browsing. Modern text retrieval methodologies are a key theme discussed throughout CS410, and the team will utilize these methods to build a tool for a real-world application.

Team GAASK plans to create a new Chrome extension that will allow users to search for secondhand items and receive a consolidated output from popular secondhand marketplaces. Currently, when a user searches for used items through a search engine (ie. Google,) the top search results link to results pages on individual websites. The user then has to continue browsing through those results to find the item. Our extension would make this process more efficient, by consolidating the results from various websites into a single view.

To build this extension, Team GAASK plans to utilize text retrieval methods, tokenization, ranking/relevance functions, and recommender systems. The text retrieval model would need to implement TF-IDF weighting for the ranking function to prioritize items that match the descriptive words of the query item. The recommender system would utilize content-based filtering, and recommend items with similar features that the user has searched for. The ranking function would also use multi-level relevance judgements to place items with the highest relevance score at the top of the list. The dataset we would use is from the existing listings on certain secondhand seller websites, including eBay, Craigslist, Poshmark, and depop. We could apply the MapReduce technique to process the data from the multiple sites in parallel and return a set of items from each website.

In order to demonstrate the functionality, the team plans to test the extension with a variety of search terms. The expected result would be a set of listings for the relevant items on the most popular second hand websites. For each search, the results will be evaluated based on how relevant the listings are to the input and the order that the listings are presented. Top listings should match the item in the query and the results should be presented in order of relevance to the user.

The main programming languages we intend to use are Javascript and Python. The Chrome extension will primarily use Javascript, HTML, and CSS for web development. For the web scraper, we will utilize Python and its vast collection of libraries such as Scrapy and BeautifulSoup. Additionally, Python's libraries can also support data extraction, XPath, debugging., etc. for our development.

The team estimates that it will take approximately 120 hours to implement a working version of our extension. The main tasks are broken down into 1) building a Chrome extension (30 hours), 2) implementing the functionality including web crawling, web scraping, filtering, ranking the results, etc. (50 hours), 3) formatting views and UI design (20 hours), 4) QA/user testing (10 hours), and 5) preparing the presentation/demo (10 hours).