Comp533 project description

Frank Wu - 260580792
Cheong Chin Wang - 260807317

Overview
For this project, our group choose to develop a new concern of a webpage crawler which will visit Web sites and reads their pages in order to extract page content to local computer.

Feature model:

- **Page getter:** The function of a webcrawler to obtain source code of a webpage by sending HTTP request to a certain Url address. This is a include feature which every webcrawler must have.
- **Authentication handling:** Authentication handling is an optional feature which will handle authentication required by the website. This feature has two subfeatures which will handle login and robot detection. For the login handling user will be asked to enter login information. However for the robot detection the software should be able to distinguish the robot detection and get permission to obtain a session to continue viewing the page content.
- **Page content processing:** After getting raw source code from webpage, page content will be processed to extract content which user desired. User will be asked to enter input depends on whether the processed content is text or other media.(image, audio, video) For media, media link will be extracted and the actual content of media will be then downloaded to local computer.

Goal and impact:

- **Authentication robustness** This goal measures how diffcult the crawler get stuck on authentication pages like login or robot check page. Only login and robot checking handling features influence this goal. login feature has value 5 and robot check has value 10 because we think robot detection is technically more diffcult to handle than login.
- **Data type diversity** This goal measures the diversity of extracted data. Text and media processing can affect this goal where text has value 3 and media has value 10 as media is much more diversified(include image,vedio,etc) than pure text.
- **Reduce running time** This goal measure how much time will be reduced. Media processing has value -5 while robot detection handling has value -10 on this goal because the media downloading takes times and robot detection handling may need machine learning algorithm for handling some situations such as captcha.

- **User convenience** This goal measures the ease of use of web crawler in terms of how much operations need to be performed by user. Login handling has value -10 because users need to input user name and password for login.