CS 839 Project Stage-2

Group Number: **23**
Chao Hsiang Chung(cchung49@wisc.edu)
Shang-Yen Yeh(syeh6@wisc.edu)
Junxia Zhu(jzhu334@wisc.edu)

Web Source:

1. **Book Depository**
   The first Web source we choose is Book Depository(https://www.bookdepository.com), which is a UK-based online bookseller with a large catalogue offered.
2. **GoodReads**
   The second Web source we choose is Goodreads(https://www.goodreads.com), which is a "social cataloguing" website that allows individuals to search its database of books, annotations, and reviews.

Extraction Method:

1. **Book Depository**
   a. Go on several Book Depository's lists to pull books' URLs by BeautifulSoup and output URLs into a file.
   b. Explore HTML layout/template of source websites, then find the consistent "HEAD," and each "LEFT" & "RIGHT" pairs tags for all attributes we are interested in extraction.
   c. Use the file generated from step (a) as data resources and extract features by tags found in step (b).
   d. Repeated doing every URL by step (c).
   e. Output the result table.
2. **GoodReads:**
   a. Go to the "What We've Read So Far in 2018" list on the website.
   b. We observe the template/HTML tags of the source website. When we detect a new book title, we capture the URL which directed to the book detail page.
   c. After we go to the individual book information page, we recognize the locations of our desired info and then get those data for every book.
   d. If we meet a book lack of certain kind of information, we put unknown into the cell.
   e. The way we extract the data is to use BeautifulSoup to parse the website and then transform the HTML format into string format. After this, we

construct a dictionary contains the left string and the right string besides our target. At last, we get the substring between those two strings, which will be the information we want.

Entity of Choice:

We choose to extract **'books'** from above two websites. Because both of these two websites provide users with book information such as Name, Author, format, pages, language and ISBN, etc. The two tables we extract represents tuples storing a unique book id and above attributes for a specific book respectively from our two web sources. Our table schema is listed as follows:
(ID, Name, Author, Format, Page, Language, ISBN_10, ISBN_13)

- **Book Depository**: 7467 tuples
- **GoodReads:** 6385 tuples

Open Source tools

1. Beautiful Soups: It is the most commonly used open-source library for building website crawler. It can extract the DOM tree easily from a website and then allow users to locate HTML tags in the website with some built-in functions.
2. Requests: It is an HTTP library for Python, the goal of the library is to make HTTP requests simpler and more human-friendly.
3. Pandas: It is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It is used a lot in data science/machine learning problems to deal with data cleaning/exploration/transformation/integration tasks.