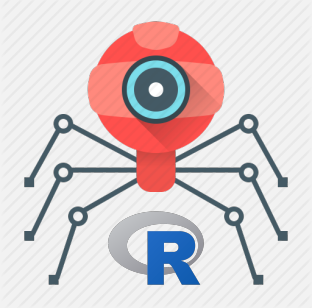
CS636 Data Analytics with R Program
Prof. Zhi Wei

NJIT New Jersey Institute of Technology



CS636 Data Analytics with R Program

Project 1

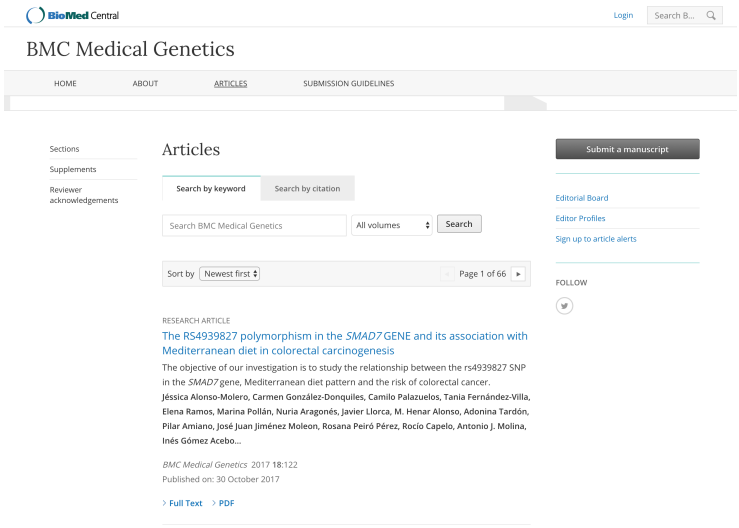
By
Group 6

Duyen Ngyen | Bicheng Xiao | Shubham Gulia

Scope of the project

Use R program to extract required information from specified website

Main page:
https://bmcmedgenet.biomedcentral.com
Total articles: 1642
Required information:
10 fields:
DOI, Title, Authors, Author Affiliations, Corresponding Author, Corresponding Author's Email, Publication Date, Abstract, Keywords, Full Text (Textual format)



Contribution of group members

Bicheng Xiao	main program: (loadAllArticles.R) Functions(util.R): • analysisArticle • extractAuthors • extractAffliation	Testing and debugging Project report(PDF file)
Duyen Ngyen	Function(util.R): • loadArticleList	Testing and debugging Readme.txt
Shubham Gulia	Function(util.R): • extract • extracAttribute	Testing and debugging ReadTheResult.R

Contribution

Challenges

unfamiliar with R packages, like XML	Check the online document and samples
extract author, corresponding author, corresponding author's email	• Carefully check the HTML of the article page and find the xpath • Use a 2nd time extraction to get more detailed or related data

CS636 Data Analytics with R Program
Prof. Zhi Wei

NJIT New Jersey Institute of Technology

Thank you