

## CS636 Data Analytics with R Programming Mid Term Project

By

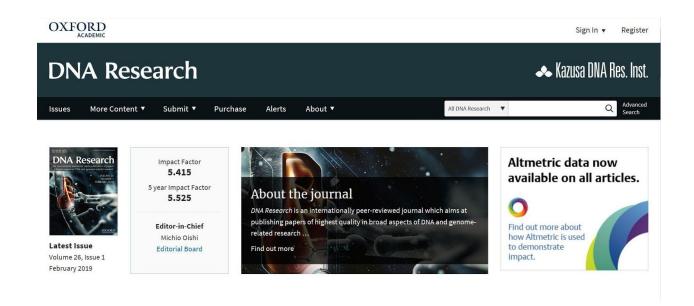
Gabriel Gonalez (geg7)

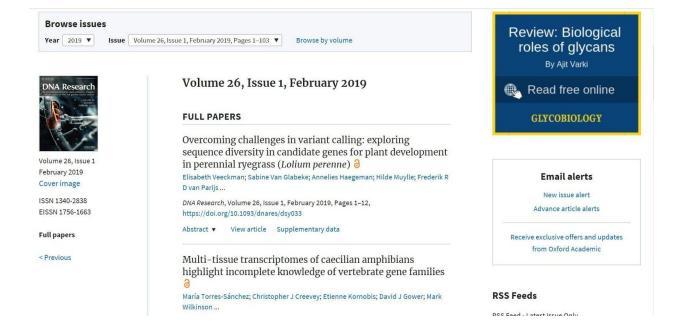
Kelvinkumar Gandhi (ksg27)

Vikramsingh Sunil Nimbalkar (vsn22)

## **Source Journal:**

https://academic.oup.com/dnaresearch





Given a journal, your R code should be capable of fetching html pages of all articles automatically. For each article, you are required to extract the following 10 fields: **DOI**, **Title**, **Authors**, **Author Affiliations**, **Corresponding Author**, **Corresponding Author's Email**, **Publication Date**, **Abstract**, **Keywords**, **Full Text**.

## ☐ Contribution of Group Members:

Vikramsingh Nimbalkar	<ul> <li>Developed R script to store a results into memory with the help of File Handling</li> <li>Executed final script with help of Exception Handling in R.</li> <li>Development and Testing and Debugging</li> </ul>	dna_research_midterm_ proj.R
Kelvinumar Gandhi	<ul> <li>Developed primary R scraping script</li> <li>Development and Performance Optimization</li> </ul>	dna_research_midterm_ proj.R
Gabriel Gonzalez	<ul> <li>Generated Regular         Expressions to match         various patterns in text         from web     </li> <li>Developed R script to         extract predefined 10         fields from scraped html         articles.</li> <li>Development, Testing         and Debugging</li> </ul>	• dna_research_midterm_ proj.R

## ☐ Challenges:

- First challenge in this project was huge number of articles. There are more than 20 years of articles in journal. And articles are not properly organized on journal webpage. So scraping each of those articles was very challenging for us.
- Second challenge in this project was unfamiliarity with some of R packages. Because we have first time use some of R packages like rvest, XML, etc to scrap the webpage and extracting required information.