Web Scraping Project - R Programming

Shawn Cicoria

3/16/2020

Introduction

This webscraping R program and scripts retrieves the content for a year for articles from the **Genetics and Molecular Biology** journal that is hosted at: National Center for Biotechnology Information

Note: this journal is also hosted at Genetics and Molecular Biology but the most recent articles and more scraping friendly HTML were at the **NCBI** site.

Setup and Running

The project consists of several `R` files:

- `main.R`
- `issue_page_reader.R`
- `article_page_reader.R`

Required packages.

the following additional R packages are required:

- logging
- xml2
- httr
- rvest
- stringr

At startup, if these packages are not able to load or installed, a `STOP` error will occur. This is handled by the following R snippet.

```
# a precheck on required packages outside normal R core
needed_packages <- c("logging", "xml2", "httr", "rvest", "stringr")
packages_installed <- needed_packages %in% rownames(installed.packages())

if (!all(packages_installed))
  stop(paste('missing some needed packages check if all installed: ', paste(n
eeded_packages,packages_installed, collapse = ';')))
```

Running is as follows:

The function contained in `main.R`, `retrieve_all_content` takes a single required parameter and 1 optional parameter.

Parameters

- year (no default / required – must be between 1998 and 2019)
- file_path - a file name and full path, otherwise will write in current `getwd()`

```
full_df <- retrieve_all_content(2017, file_path = 'all_articles2017.csv')
```

Challenges and issues

Full text parsing

The full text and some other fields contained newline characters - `\n` - which for file persistence via the `write.csv` or `write.table` functions posed the greatest issue. For that a cleanup function was created that essentially strips the Text of these characters. In a more robust model, I would look to encode perhpas using HTML, URL, or something that would preserve the integrity of the original data but allow for easy persistence and reloading.

Using XML vs CSS DOM

Traditionally, navigating HTML documents via CSS DOM is far more proper vs. XML as HTML pages aren't required to be valid XML with current standards. However, the `xml2` library, along with `rvest` was able to handle the parsing and interpretation of CSS selectors (DOM) and create XML object. However, it was far more easier to convert as with with `as_list()` the XML object to native R list and use native R. Of course navigation through the nesting was a bit tiring, but once the pattern was know, it became reasonable easy to extract text, attributes, etc. from the HTML nodes.

`main.R` - Main Entry Point

Main just contains the primary logic to firstcall the **issue** related scraping, filter by the year, then using the **article** related scraping to retrieve, parse, then via a **Data.Frame** persist to a file.

`issue_page_reader.R` - Functions to read Issue metadata

Any issue related and main page issue discovery parsing logic

`article_page_reader.R` - Functions to read and parse Article data

Any article specific reading and parsing.

Output file format

The format is essentially a CSV file using the pipe symbol '|' for separation of fields, The first row are the column names:

- url
- title
- authors
- author affiliations
- correspondence author
- correspondence email
- publish date
- keywords
- abstract
- full text

Here is an example header and first record.

```
url"|"title"|"authors"|"author_affiliations"|"correspondence"|"correspondence
_email"|"publish_date"|"keywords"|"abstract"|"full_text"
"https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5901503/?report=classic"|"Homoz
ygous sequence variants in the WNT10B gene underlie split hand/foot malformat
ion"|"Asmat Ullah;Ajab Gul;Muhammad Umair; Irfanullah;Farooq Ahmad;Abdul Aziz
;Abdul Wali;Wasim Ahmad"|"  1Department of Biochemistry, Faculty of Biologica
l Sciences, Quaid-i-Azam University, Islamabad, Pakistan.  2Department of Bio
technology and Informatics, BUITEMS, Quetta, Pakistan.  3Department of Comput
er Sciences and Bioinformatics, Khushal Khan Khattak University, Karak, Pakis
tan.Send correspondence to Wasim Ahmad, Department of Biochemistry, Faculty o
f Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan. E-mail:
kp.ude.uaq@damhawContributed by *These authors contributed equally to this st
udy."|"Send correspondence to Wasim Ahmad, Department of Biochemistry, Facult
y of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan. E-mai
l: kp.ude.uaq@damhaw"|"kp.ude.uaq@damhaw"|"Published online 2018 Jan 22. "|"S
plit-Hand-Foot Malformation 6, WNT10B gene, sequence variants"|"Split-hand/sp
lit-foot malformation (SHFM), also known as ectrodactyly is a rare genetic di
sorder. It is a clinically and genetically heterogeneous group of limb malfor
mations characterized by absence/hypoplasia and/or median cleft of hands and/
or feet. To date, seven genes underlying SHFM have been identified. This stud
y described four consanguineous families (A-D) segregating SHFM in an autosom
al recessive manner. Linkage in the families was established to chromosome 12
p11.1–q13.13 harboring WNT10B gene. Sequence analysis identified a novel homo
zygous nonsense variant (p.Gln154*) in exon 4 of the WNT10B gene in two famil
ies (A and B). In the other two families (C and D), a previously reported var
iant (c.300_306dupAGGGCGG; p.Leu103Argfs*53) was detected. This study further
expands the spectrum of the sequence variants reported in the WNT10B gene, wh
ich result in the split hand/foot malformation."|"Peripheral blood samples we
re obtained from 19 individuals in EDTA containing vacutainer sets (BD, Frank
lin Lakes, NJ, USA). Genomic DNA extraction was performed using a standard ph
```

```
enol-chloroform procedure. DNA was quantified using a Nanodrop-1000 spectroph
otometer (Thermal Scientific, Wilmington, MA).;Linkage in the families was se
arched by genotyping microsatellite markers mapped in the flanking regions of
autosomal dominant and autosomal recessive forms of SHFM. This included SHFM1
(D7S2537, D7S2481, D7S630, D7S492, D7S627, D7S1813, D7S657, D7S527, D7S479) a
t chromosome 7q21, SHFM3 (D10S520, D10S91, D10S1736, D10S1726, D10S603, D10S1
710, D10S383, D10S1264) at chromosome 10q24, SHFM4 (D3S3570, D3S3600, D3S3596
, D3S1661, D3S2747, D3S1662, D3S2311, D3S1305) at chromosome 3q27, SHFM5 (D2S
124, D2S2345, D2S294, D2S2302, D2S1274, D2S2257, D2S2173, D2S2978) at chromos
ome 2q31, SHFM6 (D12S1034, D12S823, D12S1042, D12S1337, D12S1698, D12S87, D12
S1584, D12S1621, D12S291, D12S1301, D12S1713, D12S1701, D12S339, D12S1590, D1
2S1620, D12S1635, D12S347, D12S297, D12S368, D12S398, D12S1604, D12S325) at c
hromosome 12q11-q13, and another SHFM locus mapped on chromosome 8q21.11–q22.
3 (D8S526, D8S2321, D8S1119, D8S1818, D8S1129, D8S1714, D8S556) (Gurnett et a
l., 2006). PCR amplification of the microsatellite markers was performed as p
reviously described (Ullah et al., 2015). The amplified PCR products were res
olved on 8% non-denaturing polyacrylamide gels, stained with ethidium bromide
, and genotypes were assigned by visual inspection. DNA ladders of 5, 10 and
20 bp (MBI Fermentas®, Life Sciences, York, UK) were used to determine allele
size for respective microsatellite markers. Markers used in the genotyping we
re arranged according to Rutgers combined linkage-physical map (Build 36.2) o
f the human genome (Matise et al., 2007). Haplotypes were analyzed by SIMWALK
2 (Sobel and Lange, 1996).;Primers used for PCR amplification, sequencing and
coding of intron-exon junctions of the WNT10B gene were the same as described
earlier (Khan et al., 2012). The PCR-amplified products were purified with a
commercially available kit (Axygen MD, USA) and sequenced using ABI BigDye Te
rminator Sequencing Kit v.3.1 (Applied Biosystems, Foster City, CA, USA). Seq
uence variants were identified via the BIOEDIT sequence alignment editor, ver
sion 6.0.7 (Ibis Biosciences, CA, USA).;The pathogenicity index of the sequen
ce variants identified here was calculated using the following softwares: Mut
ation Taster (http://www.mutationtaster.org/), Polymorphism Phenotyping V2 (P
olyPhen-2) (http://genetics.bwh.harvard.edu/pph2/) and Sorting Intolerant Fro
m Tolerant (SIFT) (http://sift.bii.a-star.edu.sg/). The frequency of the vari
ants in the general population was determined using the Exome Variant Server
(EVS) (http://evs.gs.washington.edu/EVS/), and 1000 genomes.; Associate Edito
r: Maria Rita Passos-Bueno "
"
```