archivist: Managing Data Analysis Results

https://github.com/pbiecek/archivist

Marcin Kosiński[1,2],
Przemysław Biecek[2]

[1]IT Research and Development
Grupa Wirtualna Polska

[2]Faculty of Mathematics, Informatics
and Mechanics, University of Warsaw

14 October, 2015

About Grupa Wirtualna Polska

The Leader Of The Polish Internet

SPORT Piłka nożna Siatkówka Żużel Tenis Sportowy Bar

Piękna Caroline szaleje w bikini i chwali się zdjęciami

Rywale Polaków w el. MŚ: nie chcieliśmy was. Macie dużo lepszych piłkarzy

Real Madryt dostał ofertę za Cristiano Ronaldo. 120 milionów euro

Skazywali ich na spadek, na razie grają rewelacyjnie. "Zwycięstwa dodają radości"

- Real podpisał umowę z prawdziwym gigantem
- ATP Hamburg: Jerzy Janowicz w II rundzie
- Polski siatkarz zagra w irańskiej potędze
- Świetny debiut gwiazdy rajdów w Hiszpanii
- Legia zmiażdżyła rywala. Berg zabrał głos
- Transfery w PlusLidze: nowe twarze w PGE Skrze
- Ranking ATP: Jerzy Janowicz poza "50"

REKLAMA Sprawdź jak schudnąć błyskawicznie

Kibice ich uwielbiali. Czarnoksiężnicy futbolu

Duży awans Urszuli Radwańskiej w rankingu WTA

REKLAMA

Monstrualna potencja każdego dnia. Jeden prosty sposób.

REKLAMA

Schudnij do 20kg na prostej diecie

Siłą odciągnął kolegę, który chciał "zabić" Polaka. Dostał za to nagrodę fair play

Zbigniew Boniek o losowaniu MŚ: "Absurd, głupota, nonsens"

- Nowy supersnajper Legii: to moja strategia
- 0:5 z Legią. "Pierwsza taka porażka w karierze"
- Baja Aragon: cenne punkty polskiej załogi
- Napastnik Man City blisko transferu do Serie A
- 8 goli w pierwszym meczu Pirlo w MLS
- Wybrano MVP finałów WGP
- El. MŚ: "Lewy i Krychowiak postraszą Rumunów"

REKLAMA

Tak wygląda brzuch 50 letniej kobiety. Odkryj jej sekret!

Taką talię uzyskasz nawet w 30 dni. Sprawdź ich metodę

Taką talię uzyskały w 30 dni? Sprawdź ich metodę!

FINANSE WIG -0,30% WIG20 -0,09% EUR 4,1259 -0,32%

Interesują Cię korzystniejsze kursy walut? Walutomat Oszczędzaj nawet 8%

Podatek PiS-u może przynieść 5 mld zł rocznie

Amerykański gigant odzieżowy wchodzi do Polski. Będzie rewolucja?

Sieć klubów fitness po lupą. Zwrócą pieniądze klientom?

- Takich inwestycji na kolei jeszcze nie było
- Nowe prawo odmieni życie dużej grupy Polaków?
- Ateny siadają do negocjacji z wierzycielami
- Kontrole skutkują. Pracodawcy boją się nie płacić
- Tak teraz kradną Polacy
- Będzie nowy unijny podatek?
- Oto najwspanialsze jachty świata
- Pieniądze za działalność opozycyjną w PRL

WP Wiesz, kto produkuje te ciastka? Niewielu Polaków już się w tym orientuje

Deklaracje PiS o zniesieniu podatku pomogły wielkiej spółce

Sebastian Ogórek

Kultowa polska marka wraca na rynek. I znów zrobi się głośno

REKLAMA

Tornister od 27.07 59,- ZOBACZ >

Nawałnice nad Polską. Straty są ogromne, ludzie mają żal

WP Trzęsienie ziemi na giełdzie w Chinach. Spadki największe od ośmiu lat

Polska daje "olbrzymie zachęty" i wygrywa ze Słowacją. Fabryka Jaguara prawdopodobnie w Krakowie

- WP Największe linie lotnicze w tarapatach
- WP Fatalne dane z Rosji
- I ty masz kredyt, o którym nie wiesz?
- Powstały trzy nowe banki. Zniszczą obecny ład?
- Wkurzyła cię reklama? Zaskarż ją!
- WP Zobacz, kto pije najwięcej piwa
- W Polsce spada produkcja alkoholu

REKLAMA

Wielka wyprzedaż do -70%. Dostawa i zwrot gratis – sprawdź!

Jest tak dobrze, że kiedy inni zwalniają, oni przyjmują do pracy

Pięknie i niedrogo. Oto idealne miejsca na emeryturę

Pomysł z Francji uratuje zyski McDonald's?

Figure : wp.pl

IT Research and Development

- large-scale online learning
- personalized news article recommendation
- e-mail targeting
- text mining
- web user behavior identification

archivist: What does it do?

Main features

- allows to **store** and **archive** objects in repositories (stored on a local disk or via GitHub/Dropbox)
- provides handy tools facilitating objects' **search** and **recovery**
- ideally performs as **cache**
- supports the philosophy of **reproducible research**

archivist: Why is it useful?

Solves reproducible research problems

- sometimes raw data are large or with limited access
- computations take a lot of time or require specialized hardware
- reproducibility requires specific versions of packages

archivist: Cache Use Case

```r
getMaxDistribution <- function(D = rnorm, N = 10,
                               R = 1000000) {
  res <- replicate(R, max(D(N)))
  summary(res)
}

# or get directly from GitHub
library(archivist)
loadFromGithubRepo(md5hash="c", user="MarcinKosinski",
                   repo="Museum")
```

archivist: Cache Use Case

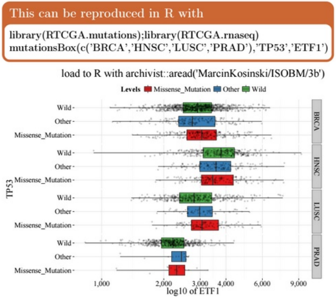
Using the archivist one can prepare a repository which stores calls and results of the cache() function to avoid their re-call in the future.

```
cacheRepo <- tempdir()
createEmptyRepo(cacheRepo)

system.time(cache(cacheRepo, getMaxDistribution,
                             rnorm, 10) )
   user  system elapsed
  5.230   0.090   5.315
system.time(cache(cacheRepo, getMaxDistribution,
                             rnorm, 10) )
   user  system elapsed
  0.009   0.000   0.008
```

archivist: Retrieving an Object Use Case

Fragment of a poster



This can be reproduced in R with

```
library(RTCGA.mutations);library(RTCGA.rnaseq)
mutationsBox(c(’BRCA’,’HNSC’,’LUSC’,’PRAD’),’TP53’,’ETF1’)
```

load to R with archivist::aread('MarcinKosinski/ISOBM/3b')

Figure : ggplot object

Using archivist to retrieve an object

```
aread(’MarcinKosinski/ISOBM/3b’)
      -> mutationsPlot
plot(mutationsPlot)
```

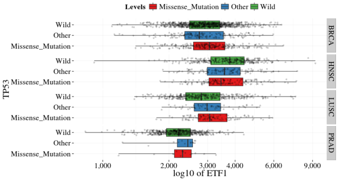


Figure : downloaded object

archivist: Object’s Pedigree Use Case

```
createEmptyRepo("FORUM_BI", default = TRUE)
invisible(aoptions("silent", TRUE))

data(iris)
iris %a%
  dplyr::filter(Sepal.Length < 16) %a%
  lm(Petal.Length~Species, data=.) %a%
  summary() -> obj

ahistory(obj)

   iris                                       [ff575c261c949d073b2895b05d1097c3]
-> dplyr::filter(Sepal.Length < 16)           [9f7045b7322cdf3a9071377c6fe9c175]
-> lm(Petal.Length ~ Species, data = .)       [0a82efeb8250a47718cea9d7f64e5ae7]
-> summary()                                  [671a0b89fccdf02087acb002374a0fcd]
```

archivist: Objects' Exploration Within a Repository

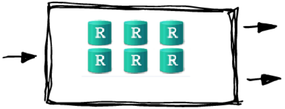
```
models <- asearch("pbiecek/graphGallery",
 patterns = c("class:lm", "coefname:Sepal.Length"))
lapply(models, coef)
```

```
[[1]]
(Intercept) Sepal.Length
  -7.101443     1.858433

[[2]]
      (Intercept)      Sepal.Length Speciesversicolor  Speciesvirginica
       -1.7023422         0.6321099         2.2101378         3.0900021
```

archivist: How does it work?

createEmptyRepo(repoDir)
summaryLocalRepo(repoDir)
summaryGithubRepo(repo, user)
deleteRepo(repoDir)



saveToRepo(artifact, repoDir)

loadFromLocalRepo(md5hash, repoDir)
loadFromGithubRepo(md5hash, repo, user)

searchInLocalRepo(pattern, repoDir)
searchInGithubRepo(pattern, repo, user)

Each repository contains a database with objects metadata.
Objects are stored as binary files.
Each object has a unique key - md5 hash.
Metadata, like object class, name, creation date, relations with other objects are useful when searching for an object in a repository.

```
library("archivist")
```

archivist: Plans & Prototypes

Automated repository creation on github, commit, push and a return of a hook to an object.

```
archive(iris, "MarcinKosinski",
 "archivist-Museum-RforeveR_last3",
        USER_EMAIL, USER_PASSWORD, app_key, app_secret,
            github_token) -> aread_input
# function returns a hook
archivist::aread("MarcinKosinski/archivist-Museum-
   RforeveR_last3/ff575c261c949d073b2895b05d1097c3")
archivist::aread(aread_input) -> x
digest::digest(x)

[1] "ff575c261c949d073b2895b05d1097c3"

identical(iris,x)

[1] TRUE
```

archivist: Learn More

http://pbiecek.github.io/archivist/