Remote Linux Cluster Parallelization in R

Daniel Huang

12 November, 2021

Here is the code from my R script (results copy pasted below):

```r
library(doFuture)

nCores <- 4
plan(multisession, workers = nCores)
registerDoFuture()

do_analysis <- function(filename, loadPackages = TRUE) {
  if(loadPackages){
    library(readr)
    library(dplyr)
    library(stringr)
  }

  data <- read_delim(filename, delim = " ", col_types = "ccccdd", quote = "",
                     col_names = c("Date", "Time", "Language", "Webpage", "Num_Hits", "Page_Size"))
  data2 <- filter(data, str_detect(Webpage, "Barack_Obama"))
  return(data2)
}

#Find files in /tmp folder
files <- list.files("/tmp", full.names = TRUE, pattern = "part-")

result <- foreach(i = seq_len(length(files))) %do% {
  cat('Starting ', i, 'th job.\n', sep = '')
  output <- do_analysis(files[i], F)
  cat('Finishing ', i, 'th job.\n', sep = '')
  output # this will become part of the out object
}

final <- bind_rows(result)

write.csv(as.data.frame(final), "Obama.csv")
```

Before running the script, I unzipped all of the .gz files in the folder to the /tmp/ folder using these bash commands:

```bash
#Code found here:
#https://superuser.com/questions/139419/how-do-i-gunzip-to-a-different-destination-directory

#After copying ../dated folder into /tmp/, from the /dated folder. This makes all of the part-XXXXX
#files in the /tmp folder to be read in by the R script in that folder
for f in *.gz; do
```

```
  STEM=$(basename "${f}" .gz)
  gunzip -c "${f}" > /tmp/"${STEM}"
done
```

After running the script, I used scp to get the Obama.csv back onto my local machine (as Obama2.csv, since I was using Obama.csv to test), which I then loaded in here for the analysis.

```
obama <- read_csv("Obama2.csv", col_types = "dccccdd")
```

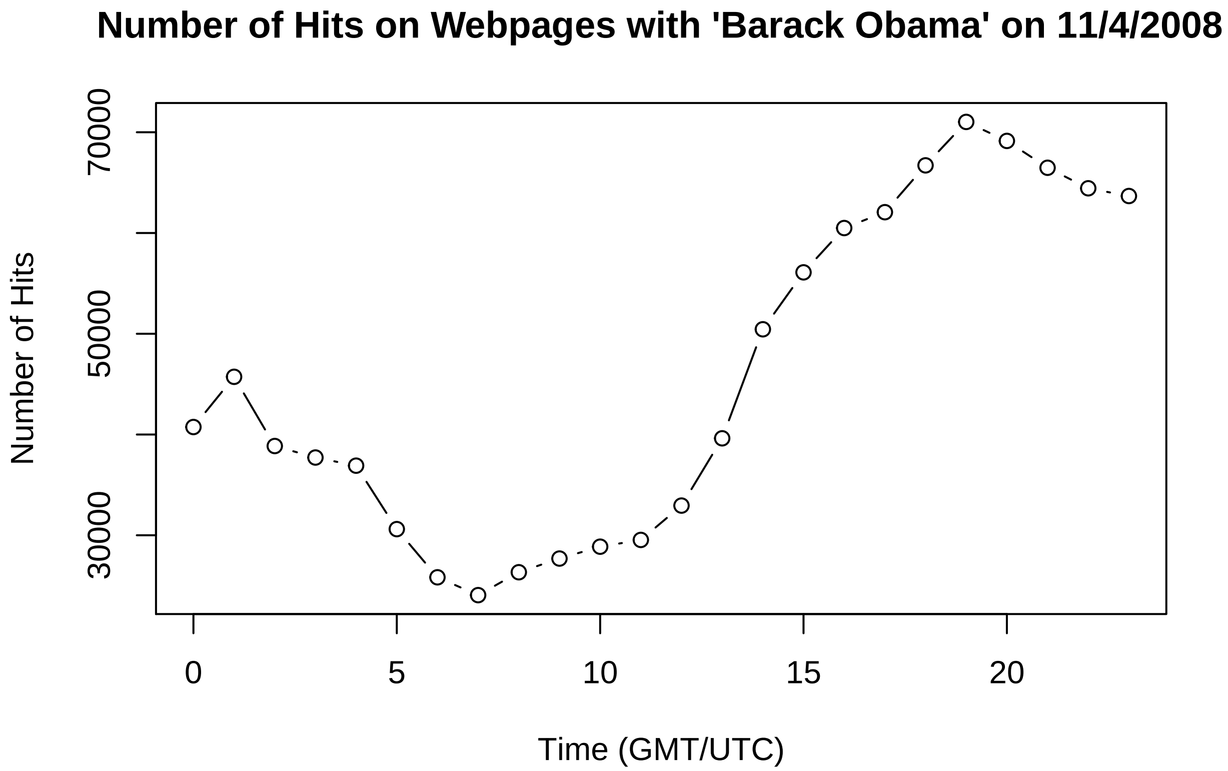
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
obama_mod <- mutate(obama, time2 = as.numeric(str_sub(Time, 1, str_length(Time) - 4)))

#Take a look at our results
head(obama_mod)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 6 x 8
##      X1 Date     Time   Language  Webpage               Num_Hits Page_Size time2
##   <dbl> <chr>    <chr>  <chr>     <chr>                    <dbl>     <dbl> <dbl>
## 1     1 20081104 110000 eu        Barack_Obama                 2     20799    11
## 2     2 20081104 080000 pl        Specjalna:Eksport/Ba~        1     79214     8
## 3     3 20081104 080000 pl        Specjalna:Export/Bar~        1       951     8
## 4     4 20081104 070000 commons.m Barack_Obama                20    552411     7
## 5     5 20081104 070000 en        President_Barack_Oba~        1     38176     7
## 6     6 20081104 060001 en        Family_of_Barack_Oba~      918  39132991     6
```

```
#Using sqldf package to do sql query
summary <- sqldf("select time2, sum(Num_Hits) from obama_mod group by time2 order by time2 desc")

plot(summary$time2, summary$`sum(Num_Hits)`, xlab = "Time (GMT/UTC)", ylab = "Number of Hits", type='b'
```



b.

I wrote a wikisubR script and used the existing wikipedia_analysis.R script. I scp'ed both to my home directory on the cluster and then ran sbatch wikisubR. Here is the wikisubR script:

```
#!/bin/bash


####################
# SBATCH OPTIONS
####################
#SBATCH --job-name=DanielWikiSub
#SBATCH --partition=low
#SBATCH --error=wiki.err
#SBATCH --output=wiki_analysis.out
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4

# Copy files over

#Remove any files in tmp beforehand
rm -rf /tmp/*
echo "Before copying:"
ls /tmp
```

```
cd /scratch/users/paciorek/wikistats/dated_2017_small/dated

for f in *.gz; do
    STEM=$(basename "${f}" .gz)
    gunzip -c "${f}" > /tmp/"${STEM}"
done

echo "After copying:"
ls /tmp
echo "Done unzipping and copying"

cd
R CMD BATCH --no-save wikipedia_analysis.R
```

We get a wikipedia_analysis.Rout file as well as the desired Obama.csv, which we can then scp back to our laptop in order to do analysis.