Kendall Goodland
EPPS 6302.001
Dr. Karl Ho

Assignment #7: Government data and parallel processing

1. Run govdata01.R. and parallel01.R.

```r
# Function to download pdfs
download_govfiles_pdf <- function(url, id) {
  tryCatch({
    destfile <- paste0(save_dir, "govfiles_", id, ".pdf")
    download.file(url, destfile = destfile, mode = "wb") # Binary files
    Sys.sleep(runif(1, 1, 3))  # Important: random sleep between 1 and 3 seconds to avoid suspicion of "hacking" the server
    return(paste("Successfully downloaded:", url))
  },
  error = function(e) {
    return(paste("Failed to download:", url))
  })
}

# Download files, potentially in parallel for speed
# Simple timer, can use package like tictoc
start.time <- Sys.time()
message("Starting downloads")
results <- 1:length(pdf_govfiles_url) %>%
  purrr::map_chr(~ download_govfiles_pdf(pdf_govfiles_url[.], pdf_govfiles_id[.]))
message("Finished downloads")
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken

# Print results
print(results)
```

```r
# Get the number of cores available on your machine
num_cores <- detectCores()

# Initialize a cluster with the number of available cores
cl <- makeCluster(num_cores)

# Load libraries and functions in each cluster
clusterEvalQ(cl, library(pdftools))

# Define the function to be parallelized after making the clusters
read_pdf_to_text <- function(uri) {
  text <- pdftools::pdf_text(uri)
  return(text)
}

# Export any libraries or objects that will be used within the parallel code

# Perform the parallel computation
tic()
pdf_texts <- parLapply(cfa_ch, read_pdf_to_text, mc.cores = num_cores)
toc()

# Don't forget to stop the cluster
stopCluster(cl)
```

2. Storage and computational resources
    a. Note the space and time taken

        i. This process takes up quite a bit of time and space, and we can analyze these things by setting up a timer with packages like “tictoc”. This allows us to see the results of our download time. To note, instead of exporting the csv, I exported the file to Excel and loaded the government data that way.
    b. Plan data management
        i. Data can be stored in data frames using the strategies explained in Assignment 6. By storing data in data frames, we can easily manipulate and visualize the data using packages like “ggplot2”.
3. Organize data
    a. Data is organized in data frames using R. Once the data has been organized into data frames, we can easily organize data based on text from the PDFs.
4. I tried storing and organizing data using arrow and parquet in R. Parquet is generally most favorable when working with very big and complex datasets and focuses on aspects of storage like compression. Arrow specializes more in the movement of data across processes. I favored parquet due to its capabilities in processing large data sets.