Foot scrapping

https://github.com/Eelai-Scheubel/scrapping

Eelai Scheubel, Laurentiu Istrati, Vasile Marcu

2025-01-14

Introduction

Ce projet utilise la librarie worldfootballR (https://cloud.r-project.org/web/packages/worldfootballR/readme/README.html) pour automatiser le processus d'extraction, filtrage et création d'une base de données.

Les données utilisées proviennent du site FBref (https://fbref.com/en/)

Objectifs

- La collecte d'une série de statistiques (tirs, passes, buts) pour la ligue de football d'un pays donné
- La création d'une base de données à partir de ces statistiques
- L'exportation au format Excel de la base de données

Extraire des données grâce à R

```r
# Fonction pour extraire les statistiques

filter_stat <- function(team_urls, stat_type, comp) {
  stats <- fb_team_match_log_stats(team_urls = team_urls, stat_type = stat_type) %>%
    filter(Comp == comp, ForAgainst == "For") %>%
    select(-all_of(if (stat_type == "shooting") {
      c("Team_Url", "ForAgainst", "Gls_Standard")
    } else {
      c(1:13)
    }))
  Sys.sleep(3)
  return(stats)
}
```

Extraire des données grâce à R

```r
# Fonction pour fusionner les statistiques d'une équipe

combine_team_stats <- function(team_url, stat_types, comp) {
  stats_list <- lapply(stat_types, function(stat_type) {
    filter_stat(team_url, stat_type, comp)
  })
  merged_df <- do.call(cbind, stats_list)
  return(merged_df)
}
```

Fonction principale

```
fetch_premier_league_stats <- function(url, stat_types, comp) {
  team_urls <- fb_teams_urls(url)
  final_data_frames <- lapply(team_urls, function(team_url) {
    combine_team_stats(team_url, stat_types, comp)
  })
  combined_df <- bind_rows(final_data_frames)
  return(combined_df)
}
```

Exemple d'utilisation

```
url <- "https://fbref.com/en/comps/9/2023-2024/2023-2024-Premier-League-Stats"
stat_types <- c("shooting", "passing", "keeper", "passing_types", "gca", "defense",
comp <- "Premier League"
name_xl <- "Database_PL.xlsx"
```

Temps d'exécution

```
t = Sys.time()
combined_df <- fetch_premier_league_stats(url, stat_types, comp)
Sys.time() - t
```

```
## Time difference of 27.35672 mins
```

Exemple d'application

```
head(combined_df[, 1:5]) %>%
  kbl(caption = "Application") %>%
  kable_styling(latex_options = c("striped", "hold_position"), full_width = FALSE)
```

Table 1: Application

Team	Date	Time	Comp	Round
Manchester City	2023-08-11	20:00	Premier League	Matchweek 1
Manchester City	2023-08-19	20:00	Premier League	Matchweek 2
Manchester City	2023-08-27	14:00	Premier League	Matchweek 3
Manchester City	2023-09-02	15:00	Premier League	Matchweek 4
Manchester City	2023-09-16	15:00	Premier League	Matchweek 5
Manchester City	2023-09-23	15:00	Premier League	Matchweek 6

Exportation

```
write_xlsx(combined_df, name_xl)
```