

Documentación de Usuario

Pipeline de Extracción y Procesamiento de Datos del Fútbol Femenino

Introducción

Esta documentación describe un sistema de extracción y procesamiento de datos estadísticos de las principales ligas femeninas de fútbol europeo. El sistema está compuesto por dos archivos principales:

1. `000_raw_to_gold.ipynb`: Notebook principal que ejecuta la secuencia de procesamiento
2. `functions.py`: Biblioteca de funciones que contiene toda la lógica de extracción y procesamiento

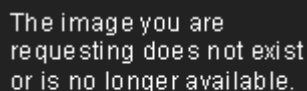
El sistema implementa un pipeline ETL (Extract, Transform, Load) que obtiene datos de varias fuentes web, los procesa y los almacena en una estructura organizada, siguiendo un enfoque de procesamiento por capas (raw → bronze → silver → gold).

Ligas incluidas

El sistema recopila datos de las siguientes ligas femeninas:

- Liga F (España)
- Women's Super League (Inglaterra)
- Frauen-Bundesliga (Alemania)
- Première Ligue (Francia)
- Serie A (Italia)

Estructura del Pipeline



The image you are requesting does not exist or is no longer available.

imgur.com

El proceso completo está organizado en las siguientes fases:

1. Configuración inicial

Importación de bibliotecas y configuración del entorno para el proceso de extracción.

2. Extracción de datos (Raw)

- `fbref_extract_all_stats()`: Extrae estadísticas detalladas de jugadoras y equipos desde fbref.com
- `extract_team_logos()`: Obtiene los escudos y nombres oficiales de los equipos desde los sitios oficiales de las ligas
- `concat_teams_info_dfs()`: Consolida la información de equipos en archivos unificados

- `fbref_player_ids()`: Recopila información detallada sobre jugadoras (URLs, fotos, altura, peso, etc.)

3. Procesamiento Bronze

Limpieza inicial y estructuración de los datos extraídos.

4. Procesamiento Silver

- `data_clearing_silver()`: Selecciona campos relevantes, estandariza formatos y realiza transformaciones intermedias

5. Procesamiento Gold

- `missing_data()`: Complementa datos faltantes en registros incompletos
- `data_clearing_gold()`: Realiza la transformación final de los datos, limpieza avanzada y consolidación

Archivos generados

El sistema genera varios conjuntos de archivos organizados por capas de procesamiento:

Capa Raw

Datos en bruto extraídos de las fuentes web:

- `df_{id_competición}_{tipo_estadística}_stats.csv`

Capa Bronze

Datos estructurados con limpieza inicial:

- `df_{id_competición}_{tipo_estadística}_bronze.csv`

Capa Silver

Datos procesados con selección de campos relevantes:

- `df_{id_competición}_players_silver.csv`
- `df_{id_competición}_keepers_silver.csv`

Capa Gold

Datos finales completamente procesados:

- `df_players_gold_1.csv`
- `df_keepers_gold_1.csv`

Información adicional

- Datos de equipos: `df_{id_competición}_teams_info.csv`

- Datos de jugadoras: `df_{id_competición}_players_info.csv` y `df_{id_competición}_keepers_info.csv`
- Archivos consolidados globales en carpetas `/big/`

Uso del sistema

Requisitos previos

- Python 3.x
- Bibliotecas: pandas, numpy, beautifulsoup4, selenium, requests, pygit2, tqdm
- WebDriver compatible con Chrome

Ejecución del pipeline completo

Para ejecutar el pipeline completo, simplemente abra y ejecute el notebook `000_raw_to_gold.ipynb`. El proceso seguirá automáticamente todas las fases descritas anteriormente.

```
# Ejemplo de ejecución de fases individuales desde Python
from functions import fbref_extract_all_stats, data_clearing_silver,
data_clearing_gold

# Extraer datos
fbref_extract_all_stats()

# Procesamiento Silver
data_clearing_silver()

# Procesamiento Gold
data_clearing_gold()
```

Tipos de estadísticas recopiladas

El sistema recopila múltiples categorías de estadísticas:

1. **standard**: Estadísticas básicas (goles, asistencias, minutos jugados)
2. **shooting**: Estadísticas de tiro (disparos, precisión, xG)
3. **passing**: Estadísticas de pase (pases completados, distancia, xA)
4. **passing_types**: Tipos de pases (centros, pases en profundidad)
5. **gca**: Acciones que generan goles
6. **defense**: Estadísticas defensivas (intercepciones, despejes)
7. **possession**: Estadísticas de posesión (toques, conducción)
8. **playing_time**: Tiempo de juego (titularidades, minutos)
9. **misc**: Estadísticas misceláneas (tarjetas, faltas)
10. **keeper**: Estadísticas de porteras
11. **keeper_adv**: Estadísticas avanzadas de porteras

Funciones principales

Extracción de datos

fbref_extract_all_stats()

Extrae todas las estadísticas de jugadoras y porteras desde fbref.com para cada liga configurada.

Parámetros:

- **comps**: Lista de IDs de competiciones a procesar (por defecto, todas las configuradas)
- **attributes**: Lista de tipos de estadísticas a extraer

Ejemplo de uso:

```
# Extraer solo datos de la Liga F (230) y WSL (189)
fbref_extract_all_stats(comps=['230', '189'])

# Extraer solo estadísticas estándar y de tiro
fbref_extract_all_stats(attributes=['standard', 'shooting'])
```

extract_team_logos()

Extrae los logos y nombres oficiales de los equipos de cada liga desde sus sitios web oficiales.

Ejemplo de uso:

```
# Extraer logos de todas las ligas configuradas
extract_team_logos()

# Extraer logos solo de la Liga F
extract_team_logos(comps=['230'])
```

fbref_player_ids()

Extrae información detallada de los perfiles de las jugadoras, incluyendo IDs, fotos, datos físicos y más.

Ejemplo de uso:

```
# Extraer perfiles de todas las jugadoras de las ligas configuradas
fbref_player_ids()
```

Procesamiento de datos

data_clearing_silver()

Transforma los datos raw/bronze a la capa silver, seleccionando campos relevantes y realizando transformaciones.

Ejemplo de uso:

```
# Procesar todas las ligas
data_clearing_silver()
```

data_clearing_gold()

Realiza el procesamiento final de los datos, generando los archivos consolidados de la capa gold.

Ejemplo de uso:

```
# Procesar todas las ligas
data_clearing_gold()
```

Notas importantes

1. El sistema implementa mecanismos de espera aleatoria entre solicitudes para evitar bloqueos por parte de los servidores web.
2. Algunas jugadoras pueden tener información incompleta que se complementa con el archivo `missing_data.csv`.
3. El sistema realiza una normalización de nombres de equipos para asegurar la consistencia entre diferentes fuentes.
4. El archivo `functions.py` contiene funciones auxiliares como `setup_selenium()`, `random_sleep_time()` y `retry_request()` que manejan aspectos técnicos de la extracción.
5. El código incluye barras de progreso (usando `tqdm`) para visualizar el avance de procesos largos como la extracción de información de perfiles de jugadoras.

Estructura de directorios

```
Proyecto_DL/
├── data/
│   ├── raw_data/           # Datos en bruto extraídos
│   ├── data_bronze/        # Datos tras limpieza inicial
│   ├── data_silver/        # Datos procesados con selección de campos
│   ├── data_gold/          # Datos finales procesados
│   ├── players_info/       # Información de jugadoras
│   │   ├── big/            # Archivos consolidados globales
│   └── teams_info/         # Información de equipos
│       ├── big/            # Archivos consolidados globales
```

Resolución de problemas comunes

Error al extraer información de jugadoras

Si encuentra errores como `'NoneType' object has no attribute 'find_all'` durante la extracción de información de jugadoras, asegúrese de:

1. Verificar la conexión a internet
2. Comprobar que las URLs configuradas siguen siendo válidas
3. Aumentar los tiempos de espera entre solicitudes modificando la función `random_sleep_time()`
4. Ejecutar nuevamente solo la función `fbref_player_ids()` para completar la extracción

Información faltante

El sistema está diseñado para manejar información incompleta. Si observa datos faltantes en los archivos finales:

1. Añada manualmente los datos en el archivo `missing_data.csv`
2. Ejecute nuevamente la función `data_clearing_gold()`

Mantenimiento y actualizaciones

Para mantener el sistema actualizado:

1. Verifique regularmente las URLs de extracción en caso de cambios en las fuentes
2. Actualice los mapeos de nombres de equipos en `data_clearing_gold()` si se producen cambios en los equipos
3. Revise periódicamente el archivo `missing_data.csv` para añadir información de nuevas jugadoras

Esta documentación proporciona una visión general del sistema. Para detalles específicos sobre cada función, consulte los comentarios en el código fuente de `functions.py`.