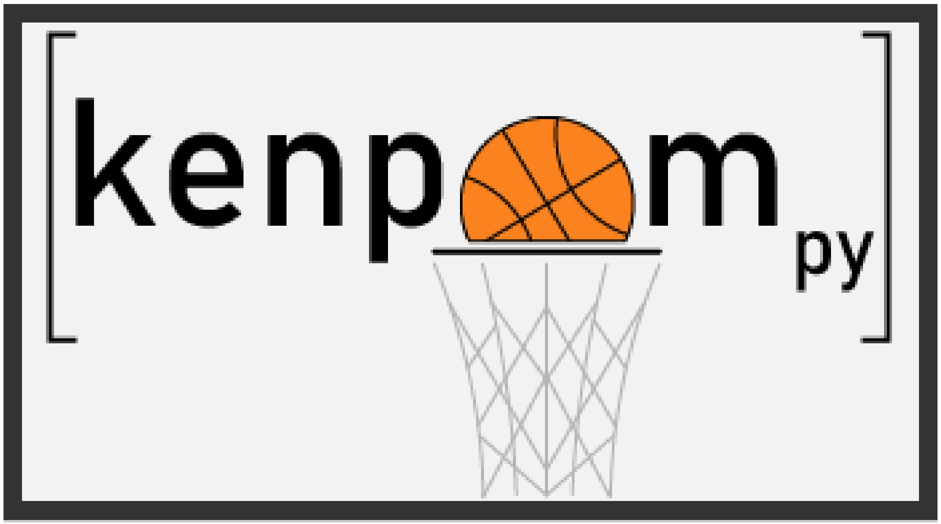
kenpompy

Release 0.1.0

Jared Andrews

Feb 17, 2020

CONTENTS:



This python package serves as a convenient web scraper for kenpom, which provides tons of great NCAA basketball statistics and metrics. It **requires a subscription to KenPomeroy's site for use**, otherwise only the home page will be accessible. It's a small fee for a year of access, and totally worth it in my opinion.

CHAPTER ONE

OBJECTIVE

Ultimately, this package is to allow both hobbyist and reknown sports analysts alike to get data from kenpom in a format more suitable for visualization, transformation, and additional analysis. It's meant to be simple, easy to use, and to yield information in a way that is immediately usable.

CHAPTER

TWO

RESPONSIBLE USE

As with many web scrapers, the responsibility to use this package in a reasonable manner falls upon the user. Don't be a jerk and constantly scrape the site a thousand times a minute or you run the risk of potentially getting barred from it, which you'd likely deserve. I am in no way responsible for how you use (or abuse) this package. Be sensible.

CHAPTER

THREE

BUT I USE R

Yeah, yeah, but have you heard of reticulate? It's an R interface to python that also supports passing objects (like dataframes!) between them.

CHAPTER

FOUR

INSTALLATION

kenpompy is easily installed via pip:

```
pip install kenpompy
```

CHAPTER

FIVE

WHAT IT CAN (AND CAN'T) DO

This a work in progress - it can currently scrape all of the summary, FanMatch, and miscellaneous tables, pretty much all of those under the Stats and Miscellany headings. `Team` and `Player` classes are planned, but they're more complicated and will take some time.

CHAPTER SIX

USAGE

kenpompy is simple to use. Generally, tables on each page are scraped into pandas dataframes with simple parameters to select different seasons or tables. As many tables have headers that don't parse well, some are manually altered to a small degree to make the resulting dataframe easier to interpret and manipulate.

First, you must login:

```python
from kenpompy.utils import login

# Returns an authenticated browser that can then be used to scrape pages that require authorization.
browser = login(your_email, your_password)
```

Then you can request specific pages that will be parsed into convenient dataframes:

```python
import kenpompy.summary as kp

# Returns a pandas dataframe containing the efficiency and tempo stats for the current season (https://kenpom.com/summary.php).
eff_stats = kp.get_efficiency(browser)
```

CHAPTER

SEVEN

FULL API REFERENCE

7.1 utils

The utils module provides utility functions, such as logging in.

`kenpompy.utils.`**`login`**`(`*email*, *password*`)`

Logs in to kenpom.com using user credentials.

Parameters

- **`email`** (*`str`*) – User e-mail for login to kenpom.com.
- **`password`** (*`str`*) – User password for login to kenpom.com.

Returns Authenticated browser with full access to kenpom.com.

Return type browser (mechanicalsoup StatefulBrowser)

7.2 misc

This module provides functions for scraping the miscellaneous stats kenpom.com pages into more usable pandas dataframes.

`kenpompy.misc.`**`get_arenas`**`(`*browser*, *season=None*`)`

Scrapes the arenas table (https://kenpom.com/arenas.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2010 is the earliest available season. Most recent season is the default.

Returns Pandas dataframe containing the arenas table from kenpom.com.

Return type arenas_df (pandas dataframe)

Raises **`ValueError`** – If *season* is less than 2010.

`kenpompy.misc.`**`get_gameattribs`**`(`*browser*, *season=None*, *metric='Excitement'*`)`

Scrapes the Game Attributes tables (https://kenpom.com/game_attrs.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.

- **season** (*str, optional*) – Used to define different seasons. 2010 is the earliest available season. Most recent season is the default.
- **metric** (*str, optional*) – Used to get highest ranking games for different metrics. Available values are: 'Excitement', 'Tension', 'Dominance', 'ComeBack', 'FanMatch', 'Upsets', and 'Busts'. Default is 'Excitement'. 'FanMatch', 'Upsets', and 'Busts' are only valid for seasons after 2010.

Returns Pandas dataframe containing the Game Attributes table from kenpom.com for a given metric.

Return type ga_df (pandas dataframe)

Raises

- **ValueError** – If *season* is less than 2010.
- **KeyError** – If *metric* is invalid.

kenpompy.misc.**get_hca**(*browser*)

Scrapes the home court advantage table (https://kenpom.com/hca.php) into a dataframe.

Parameters

- **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **season** (*str, optional*) – Used to define different seasons. 2010 is the earliest available season.

Returns Pandas dataframe containing the home court advantage table from kenpom.com.

Return type hca_df (pandas dataframe)

kenpompy.misc.**get_program_ratings**(*browser*)

Scrapes the program ratings table (https://kenpom.com/programs.php) into a dataframe.

Parameters **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.

Returns Pandas dataframe containing the program ratings table from kenpom.com.

Return type programs_df (pandas dataframe)

kenpompy.misc.**get_refs**(*browser*, *season=None*)

Scrapes the officials rankings table (https://kenpom.com/officials.php) into a dataframe.

Parameters

- **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **season** (*str, optional*) – Used to define different seasons. 2016 is the earliest available season. Most recent season is the default.

Returns Pandas dataframe containing the officials rankings table from kenpom.com.

Return type refs_df (pandas dataframe)

Raises **ValueError** – If *season* is less than 2016.

kenpompy.misc.**get_trends**(*browser*)

Scrapes the statistical trends table (https://kenpom.com/trends.php) into a dataframe.

Parameters **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.

Returns Pandas dataframe containing the statistical trends table from kenpom.com.

Return type trends_df (pandas dataframe)

7.3 summary

This module provides functions for scraping the summary stats kenpom.com pages into more usable pandas dataframes.

`kenpompy.summary.`**`get_efficiency`**`(`*browser*, *season=None*`)`

Scrapes the Efficiency stats table (https://kenpom.com/summary.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2002 is the earliest available season but possession length data wasn't available until 2010. Most recent season is the default.

Returns Pandas dataframe containing the summary efficiency/tempo table from kenpom.com.

Return type eff_df (pandas dataframe)

Raises **`ValueError`** – If *season* is less than 2002.

`kenpompy.summary.`**`get_fourfactors`**`(`*browser*, *season=None*`)`

Scrapes the Four Factors table (https://kenpom.com/stats.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2002 is the earliest available season. Most recent season is the default.

Returns Pandas dataframe containing the summary Four Factors table from kenpom.com.

Return type ff_df (pandas dataframe)

Raises **`ValueError`** – If *season* is less than 2002.

`kenpompy.summary.`**`get_height`**`(`*browser*, *season=None*`)`

Scrapes the Height/Experience table (https://kenpom.com/height.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2007 is the earliest available season but continuity data wasn't available until 2008. Most recent season is the default.

Returns Pandas dataframe containing the Height/Experience table from kenpom.com.

Return type h_df (pandas dataframe)

Raises **`ValueError`** – If *season* is less than 2007.

`kenpompy.summary.`**`get_kpoy`**`(`*browser*, *season=None*`)`

Scrapes the kenpom Player of the Year tables (https://kenpom.com/kpoy.php) into dataframes.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2011 is the earliest available season. Most recent season is the default.

Returns

List of dandas dataframes containing the kenpom Player of the Year and Game MVP leaders tables from kenpom.com. Game MVP table only available from 2013 season onwards.

Return type kpoy_dfs (list of pandas dataframe)

Raises **`ValueError`** – If *season* is less than 2011.

`kenpompy.summary.`**`get_playerstats`**(*browser*, *season=None*, *metric='EFG'*, *conf=None*, *conf_only=False*)

Scrapes the Player Leaders tables (https://kenpom.com/playerstats.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **`season`** (*`str, optional`*) – Used to define different seasons. 2004 is the earliest available season. Most recent season is the default.
- **`metric`** (*`str, optional`*) – Used to get leaders for different metrics. Available values are: 'ORtg', 'Min', 'eFG', 'Poss', 'Shots', 'OR', 'DR', 'TO', 'ARate', 'Blk', 'FTRate', 'Stl', 'TS', 'FC40', 'FD40', '2P', '3P', 'FT'. Default is 'eFG'. 'ORtg' returns a list of four dataframes, as there are four tables: players that used >28% of possessions, >24% of possessions, >20% of possessions, and with no possession restriction.
- **`conf`** (*`str, optional`*) – Used to limit to players in a specific conference. Allowed values are: 'A10', 'ACC', 'AE', 'AMER', 'ASUN', 'B10', 'B12', 'BE', 'BSKY', 'BSTH', 'BW', 'CAA', 'CUSA', 'HORZ', 'IND', IVY', 'MAAC', 'MAC', 'MEAC', 'MVC', 'MWC', 'NEC', 'OVC', 'P12', 'PAT', 'SB', 'SC', 'SEC', 'SLND', 'SUM', 'SWAC', 'WAC', 'WCC'. If you try to use a conference that doesn't exist for a given season, like 'IND' and '2018', you'll get an empty table, as kenpom.com doesn't serve 404 pages for invalid table queries like that. No filter applied by default.
- **`conf_only`** (*`bool, optional`*) – Used to define whether stats should reflect conference games only. Only available if specific conference is defined. Only available for seasons after 2013. False by default.

Returns Pandas dataframe containing the Player Leaders table from kenpom.com.

Return type ps_df (pandas dataframe)

Raises

- **`ValueError`** – If *season* is less than 2004 or *conf_only* is used with an invalid *season*.
- **`KeyError`** – If *metric* is invalid.

`kenpompy.summary.`**`get_pointdist`**(*browser*, *season=None*)

Scrapes the Team Points Distribution table (https://kenpom.com/pointdist.php) into a dataframe.

Parameters

- **`browser`** (*`mechanicalsoup StatefulBrowser`*) – Authenticated browser with full access to kenpom.com generated by the *login* function.

- **season** (*str, optional*) – Used to define different seasons. 2002 is the earliest available season. Most recent season is the default.

Returns Pandas dataframe containing the Team Points Distribution table from kenpom.com.

Return type dist_df (pandas dataframe)

Raises **ValueError** – If *season* is less than 2002.

`kenpompy.summary.`**`get_teamstats`**(*browser*, *defense=False*, *season=None*)

Scrapes the Miscellaneous Team Stats table (https://kenpom.com/teamstats.php) into a dataframe.

Parameters

- **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **defense** (*bool, optional*) – Used to flag whether the defensive teamstats table is wanted or not. False by default.
- **season** (*str, optional*) – Used to define different seasons. 2002 is the earliest available season. Most recent season is the default.

Returns Pandas dataframe containing the Miscellaneous Team Stats table from kenpom.com.

Return type ts_df (pandas dataframe)

Raises **ValueError** – If *season* is less than 2002.

7.4 FanMatch

This module contains the FanMatch class for scraping the FanMatch pages into more usable objects.

class `kenpompy.FanMatch.`**`FanMatch`**(*browser*, *date=None*)

Object to hold FanMatch page scraping results.

This class scrapes the kenpom FanMatch page when a new instance is created.

Parameters

- **browser** (*mechanicalsoup StatefulBrowser*) – Authenticated browser with full access to kenpom.com generated by the *login* function.
- **date** (*str*) – Date to scrape, in format "YYYY-MM-DD", such as "2020-01-29".

url

Full url for the page to be scraped.

Type str

date

Date to be scraped.

Type str

lines_o_night

List containing lines of the night if games have taken place.

Type list

ppg

Average points per game for the day.

Type float

avg_eff

Average efficiency for the day.

Type float

pos_40

Possessions per 40 minutes for the day.

Type float

mean_abs_err_pred_total_score

The mean absolute error of predicted total score for the day.

Type float

bias_pred_total_score

The bias of predicted total score for the day.

Type float

mean_abs_err_pred_mov

The mean absolute error of the predicted margin of victory for the day.

Type float

record_favs

Record of favorites for the day.

Type str

expected_record_favs

Expected record of favorites for the day.

Type str

exact_mov

Number of games where margin of victory was accurately predicted out of total played.

Type str

fm_df

Pandas dataframe containing parsed FanMatch table.

Type pandas dataframe

CHAPTER

EIGHT

CONTRIBUTING

You can contribute by creating issues to highlight bugs and make suggestions for additional features. Pull requests are also welcome.

CHAPTER

NINE

LICENSE

kenpompy is released on the GNU GPL v3.0 license. You are free to use, modify, or redistribute it in almost any way, provided you state changes to the code, disclose the source, and use the same license. It is released with zero warranty for any purpose and I retain no liability for its use. Read the full license for additional details.

CHAPTER

TEN

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

INDEX