Names and ID's of all participants.

Dominik Chrząstowski-Wachtel, 262616

- Short description of the topic and the web page.

I wanted to gather data about judo athletes from judo database.

IJF - International Judo Federation

I chose to gather information about seniors in all weight catergories, both men and women, and to count their medals.

- If you omit Beautiful Soup scraper: justification, that the page is dynamic, and scraping can not be done with BS there.

-

- Short description of your scraper mechanics.

All three scrapers have the same approach -

1) I take values from country dropdown on https://www.ijf.org/judoka to check their country codes.

Since each country can bo chosen by swapping the url, I can get the list of countries by manipulating the url with the value from country dropdown

2) On page for each country, I gather data about each competitor in "Senior" category - I chose this approach to limit the amount of pages scraped. There were 4163 individual pages with competitors scraped.

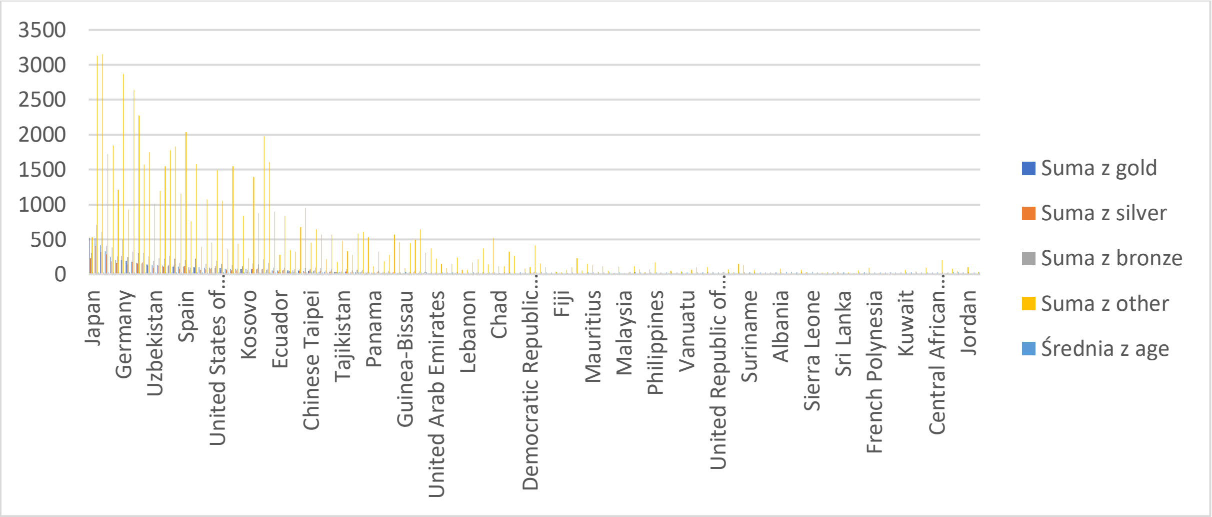
3) Having individual page for each competitor, I gather information about their name, age, weight category, country, and medals and other achievements

- Short technical description of the output you get.

- Extremely elementary data analysis - you need to prove, that collected data can be used for further analysis, but nothing more (hard limit of data analysis: one page).

I used pivot tables in Excel to see which country has the most medals among their current seniors

I also checked the average age of the competitors. There are many different interesting ways to interpret the data, and the scrapers can be extended easily, as the ijf website is quite consistent.



Etykiety wierszy	Sum gold	Sum silver	Sum bronze	Mean age
Japan	525	229	311	24,37037037
Russian Federation	517	410	704	25,72072072
France	420	320	609	25,52173913
Brazil	335	285	409	25,88043478
Netherlands	247	186	387	24,13043478
Georgia	202	162	257	24,14084507
Germany	202	264	498	24,63809524
Republic of Korea	191	139	250	24,59701493
Italy	183	183	322	24,55555556
Hungary	166	155	312	23,81538462
Mongolia	156	172	311	24,45454545
Kazakhstan	143	125	252	23,61589404
Uzbekistan	133	98	194	23,4047619
Canada	130	129	234	23,32142857
Azerbaijan	128	112	223	24,13924051
Great Britain	119	137	265	25,25396825
Ukraine	113	124	226	24,725
Slovenia	112	83	167	24,61290323
Spain	109	119	202	25,72289157
Australia	104	68	95	25,68888889
Israel	101	95	222	23,65517241
Tunisia	100	64	67	24,84210526
Turkey	94	74	146	23,44230769
Algeria	92	80	87	25,55263158
United States of America	90	118	198	24,62666667
Belgium	90	61	148	25,4137931
Dominican Republic	83	61	72	23,7173913
Romania	82	58	134	24,43902439

- Detailed description which participant wrote which part of the project.

BeautifulSoup:

judo1.py : Dominik Chrząstowski-Wachtel

Scrapy:

judo2.py : Dominik Chrząstowski-Wachtel

judo3.py : Dominik Chrząstowski-Wachtel

judo4.py : Dominik Chrząstowski-Wachtel

Selenium:

judo5.py : Dominik Chrząstowski-Wachtel

judo6.py : Dominik Chrząstowski-Wachtel

Other:

README.md: Dominik Chrząstowski-Wachtel

description.pdf: Dominik Chrząstowski-Wachtel

======

4155.311539888382 seconds for BeautifulSoup

In order to get the correct csv values order for Scrapy:

settings.py needs to have this part added:

```
  FEED_EXPORT_FIELDS = [
   'name',
   'age',
   'country',
   'wCat',
   'gold',
   'silver',
   'bronze',
   'other'
]
```

```
scrapy crawl country_list -O countries.csv
scrapy crawl judoka_list -O athletes.csv
scrapy crawl judoka -O results.csv
```

```
'start_time': datetime.datetime(2021, 5, 9, 12, 47, 7, 849328)}
'finish_time': datetime.datetime(2021, 5, 9, 12, 50, 1, 778329),
datetime.timedelta(seconds=173, microseconds=929001)
```