```
# web scraping
!pip install gazpacho

```

```
from gazpacho import Soup
import requests
import pandas as pd
import numpy as np
```

```
from gazpacho import Soup
import requests
import pandas as pd
import numpy as np
url = "https://www.animenewsnetwork.com/encyclopedia/ratings-anime.php?top50=popular&n=100"
html = requests.get(url)
top_100 = Soup(html.text)
```

```
# Find anime name
titles = top_100.find("a", {"href" : "/encyclopedia/anime.php"})
titles[0:5]
```

```
[<a href="/encyclopedia/anime.php">Anime</a>,
 <a href="/encyclopedia/anime.php?id=2960">Fullmetal Alchemist (TV)</a>,
 <a href="/encyclopedia/anime.php?id=6592">Death Note (TV)</a>,
 <a href="/encyclopedia/anime.php?id=13">Cowboy Bebop (TV)</a>,
 <a href="/encyclopedia/anime.php?id=377">Spirited Away (movie)</a>]
```

```
# Clean titles
# Delete non anime name
clean_titles =[title.strip() for title in titles ]
clean_titles.pop(0)
clean_titles[0:5]
```

```
['Fullmetal Alchemist (TV)',
 'Death Note (TV)',
 'Cowboy Bebop (TV)',
 'Spirited Away (movie)',
 'Princess Mononoke (movie)']
```

```
# Find rating
ratings = top_100.find("td", {"class" : "r"})
ratings[0:6]
```

```
[<td class="r">rating</td>,
 <td class="r">nb. votes</td>,
 <td class="r">8.72</td>,
 <td class="r">13523</td>,
 <td class="r">8.86</td>,
 <td class="r">13107</td>]
```

```
# Clean rating
clean_ratings = [rating.strip() for rating in ratings]
clean_ratings.pop(0)
clean_ratings.pop(0)
clean_ratings[0:10]
```

```
['8.72',
 '13523',
 '8.86',
 '13107',
 '8.93',
 '12444',
 '8.96',
 '10706',
 '8.93',
 '9974']
```

```
# In the list, it contain rating and number vote.
# So, we have to extract and create new list for that.
# Extract only rating
float_numbers = [float(number) for number in clean_ratings]
filtered_ratings = [rating for rating in float_numbers if rating <= 10]

```

```
# Extract only number vote
int_numbers = [int(float(number)) for number in clean_ratings]
filtered_number_vote = [number for number in int_numbers if number >10]


```

```
df = pd.DataFrame(data = {
    "titles" : clean_titles,
    "rating" : filtered_ratings,
    "number_vote" : filtered_number_vote
})

df.head()
```

	titles	rating	number_vote
0	Fullmetal Alchemist (TV)	8.72	13523
1	Death Note (TV)	8.86	13107
2	Cowboy Bebop (TV)	8.93	12444
3	Spirited Away (movie)	8.96	10706
4	Princess Mononoke (movie)	8.93	9974

```
# Further more, we want to extract type of anime in parentheses
# Extract type into the new column
df["Type"] = df["titles"].str.extract(r'\((.*)\)', expand =False)

# Remove () in tiltes
df['titles'] = df['titles'].str.replace(r'\(.*?\)', '').str.strip()

df.head(10)
```

```
<ipython-input-17-8598cdf67f3e>:6: FutureWarning: The default value of regex will change from True to F
  df['titles'] = df['titles'].str.replace(r'\(.*?\)', '').str.strip()
```

	titles	rating	number_vote	Type
0	Fullmetal Alchemist	8.72	13523	TV
1	Death Note	8.86	13107	TV
2	Cowboy Bebop	8.93	12444	TV
3	Spirited Away	8.96	10706	movie
4	Princess Mononoke	8.93	9974	movie
5	Melancholy of Haruhi Suzumiya	8.56	10225	The) Melancholy of Haruhi Suzumiya (TV
6	Elfen Lied	8.29	10530	TV
7	Neon Genesis Evangelion	8.32	10372	TV
8	Code Geass: Lelouch of the Rebellion	8.85	9308	TV
9	Bleach	7.94	9242	TV

```
# There are still wrong values in index = 5
# Because, the original name is (The) Melancholy of Haruhi Suzumiya (TV)
# It was extract by regex
# So, just replace it.
df.iloc[5,3] = "TV"
df.iloc[5,0] = "(The) Melancholy of Haruhi Suzumiya"

df.head(10)
```

	titles	rating	number_vote	Type
0	Fullmetal Alchemist	8.72	13523	TV
1	Death Note	8.86	13107	TV
2	Cowboy Bebop	8.93	12444	TV
3	Spirited Away	8.96	10706	movie
4	Princess Mononoke	8.93	9974	movie
5	(The) Melancholy of Haruhi Suzumiya	8.56	10225	TV
6	Elfen Lied	8.29	10530	TV
7	Neon Genesis Evangelion	8.32	10372	TV
8	Code Geass: Lelouch of the Rebellion	8.85	9308	TV
9	Bleach	7.94	9242	TV

✓ 0s completed at 10:10 PM