Scraping from Wikipedia

In [2]:
```
import requests
import pandas as pd
import bs4
res = requests.get('https://en.wikipedia.org/wiki/Machine_learning')
sp = bs4.BeautifulSoup(res.text , 'lxml')
type(sp)
```

Out[2]:
```
bs4.BeautifulSoup
```

In [3]:
```
sp.select('.mw-headline')
```

Out[3]:
```
[<span class="mw-headline" id="Overview">Overview</span>,
 <span class="mw-headline" id="Machine_learning_tasks">Machine learning tasks</
span>,
 <span class="mw-headline" id="Machine_learning_applications">Machine learning
applications</span>,
 <span class="mw-headline" id="History_and_relationships_to_other_fields">Histo
ry and relationships to other fields</span>,
 <span class="mw-headline" id="Relation_to_statistics">Relation to statistics</
span>,
 <span class="mw-headline" id="Theory"><span id="Generalization"></span> Theory
</span>,
 <span class="mw-headline" id="Approaches">Approaches</span>,
 <span class="mw-headline" id="Decision_tree_learning">Decision tree learning</
span>,
 <span class="mw-headline" id="Association_rule_learning">Association rule lear
ning</span>,
 <span class="mw-headline" id="Artificial_neural_networks">Artificial neural ne
tworks</span>,
 <span class="mw-headline" id="Deep_learning">Deep learning</span>,
 <span class="mw-headline" id="Inductive_logic_programming">Inductive logic pro
gramming</span>,
 <span class="mw-headline" id="Support_vector_machines">Support vector machines
</span>,
 <span class="mw-headline" id="Clustering">Clustering</span>,
 <span class="mw-headline" id="Bayesian_networks">Bayesian networks</span>,
 <span class="mw-headline" id="Reinforcement_learning">Reinforcement learning</
span>,
 <span class="mw-headline" id="Representation_learning">Representation learning
</span>,
 <span class="mw-headline" id="Similarity_and_metric_learning">Similarity and m
etric learning</span>,
 <span class="mw-headline" id="Sparse_dictionary_learning">Sparse dictionary le
arning</span>,
 <span class="mw-headline" id="Genetic_algorithms">Genetic algorithms</span>,
 <span class="mw-headline" id="Rule-based_machine_learning">Rule-based machine
learning</span>,
 <span class="mw-headline" id="Learning_classifier_systems">Learning classifier
systems</span>,
 <span class="mw-headline" id="Applications">Applications</span>,
 <span class="mw-headline" id="Model_assessments">Model assessments</span>,
 <span class="mw-headline" id="Ethics">Ethics</span>,
 <span class="mw-headline" id="Software">Software</span>,
 <span class="mw-headline" id="Free_and_open-source_software">Free and open-sou
rce software</span>,
 <span class="mw-headline" id="Proprietary_software_with_free_and_open-source_e
ditions">Proprietary software with free and open-source editions</span>,
 <span class="mw-headline" id="Proprietary_software">Proprietary software</span
>,
 <span class="mw-headline" id="Journals">Journals</span>,
 <span class="mw-headline" id="Conferences">Conferences</span>,
 <span class="mw-headline" id="See_also">See also</span>,
 <span class="mw-headline" id="References">References</span>,
 <span class="mw-headline" id="Further_reading">Further reading</span>,
 <span class="mw-headline" id="External_links">External links</span>]
```

In [4]:
```
lst = []
for i in sp.select('.mw-headline'):
    lst += [i.text]
print(lst)
```

```
['Overview', 'Machine learning tasks', 'Machine learning applications', 'Histor
y and relationships to other fields', 'Relation to statistics', ' Theory', 'App
roaches', 'Decision tree learning', 'Association rule learning', 'Artificial ne
ural networks', 'Deep learning', 'Inductive logic programming', 'Support vector
machines', 'Clustering', 'Bayesian networks', 'Reinforcement learning', 'Repres
entation learning', 'Similarity and metric learning', 'Sparse dictionary learni
ng', 'Genetic algorithms', 'Rule-based machine learning', 'Learning classifier
systems', 'Applications', 'Model assessments', 'Ethics', 'Software', 'Free and
open-source software', 'Proprietary software with free and open-source edition
s', 'Proprietary software', 'Journals', 'Conferences', 'See also', 'Reference
s', 'Further reading', 'External links']
```

In [5]:
```
ml = pd.DataFrame(lst , columns = ['Content'])
ml
```

Out[5]:

	Content
0	Overview
1	Machine learning tasks
2	Machine learning applications
3	History and relationships to other fields
4	Relation to statistics
5	Theory
6	Approaches
7	Decision tree learning
8	Association rule learning
9	Artificial neural networks
10	Deep learning
11	Inductive logic programming
12	Support vector machines
13	Clustering
14	Bayesian networks
15	Reinforcement learning
16	Representation learning
17	Similarity and metric learning
18	Sparse dictionary learning
19	Genetic algorithms
20	Rule-based machine learning
21	Learning classifier systems
22	Applications
23	Model assessments
24	Ethics
25	Software
26	Free and open-source software
27	Proprietary software with free and open-source...
28	Proprietary software
29	Journals
30	Conferences
31	See also
32	References

	Content
33	Further reading
34	External links

In [9]:
```
ml.to_csv('web_scrape.csv' , index = False)
"""Successfully saved the scraped data as a csv file"""
```