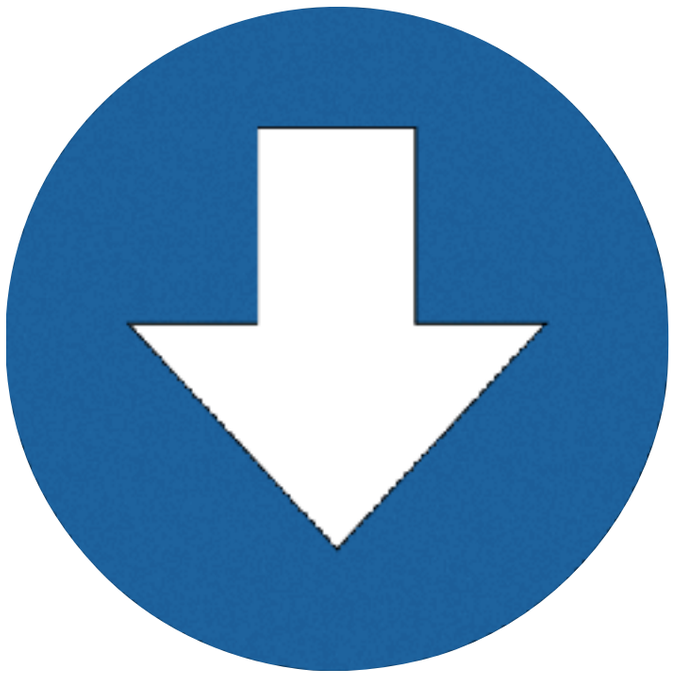
Tìm hiểu hệ thống mã nguồn mở Crawler

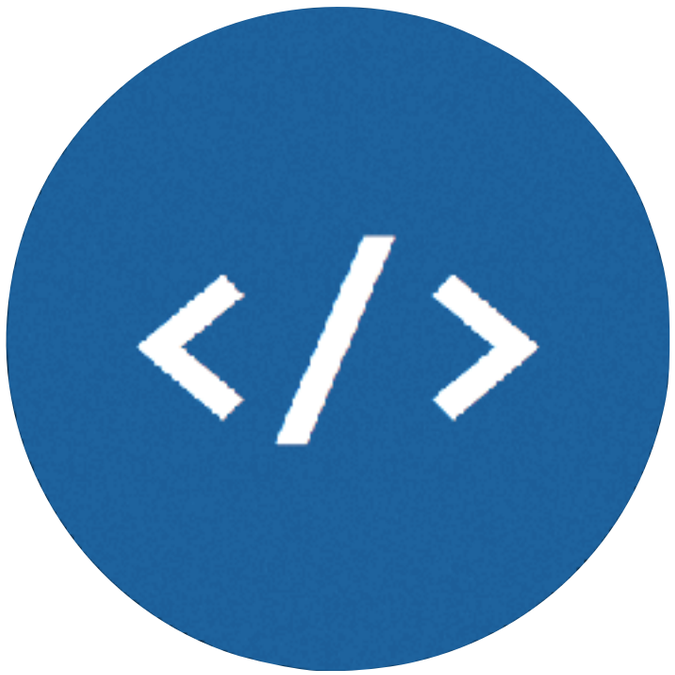
Group 3

Members:

- Trần Minh Tuấn
- Trần Thị Thanh Huyền
- Hoàng Văn Hải
- Nguyễn Thị Hồng Hải



Fetch



Extract

Store

Fetch one page

```
import requests

page = requests.get('http://dictionary.cambridge.org')
```

Fetch a set of results

```
import requests

base_url = ‘http://dictionary.cambridge.org/
dictionary/english-vietnamese/%s'

for word in list_words:
    url = base_url % word
    page= requests.get(url)
```

Fetch a set of results

```
import requests

page='requests.get(url)'

with open(page,’"wb") as html:
    # extract structure webpage
```

Parsing data

- Regular Expressions
- CSS Selectors
- XPath
- Object Searching

Parsing data

```
<html>
  <head><title>Green Eggs and Ham</title></head>
  <body>
    <ol>
      <li>Green Eggs</li>
      <li>Ham</li>
    </ol>
  </body>
</html>
```

Regular Expressions

```
import re

item_re'='re.compile("<li[^>]*>([^<]+?)</li>")'

item_re.findall(html)
```

CSS Selectors

```
from bs4 import BeautifulSoup

soup= BeautifulSoup(html)

[s.text for s in soup.select("li")]
```

XPath

```
from lxml import etree

 s.text for s in tree.xpath('/ol/li')
```

Write JSON

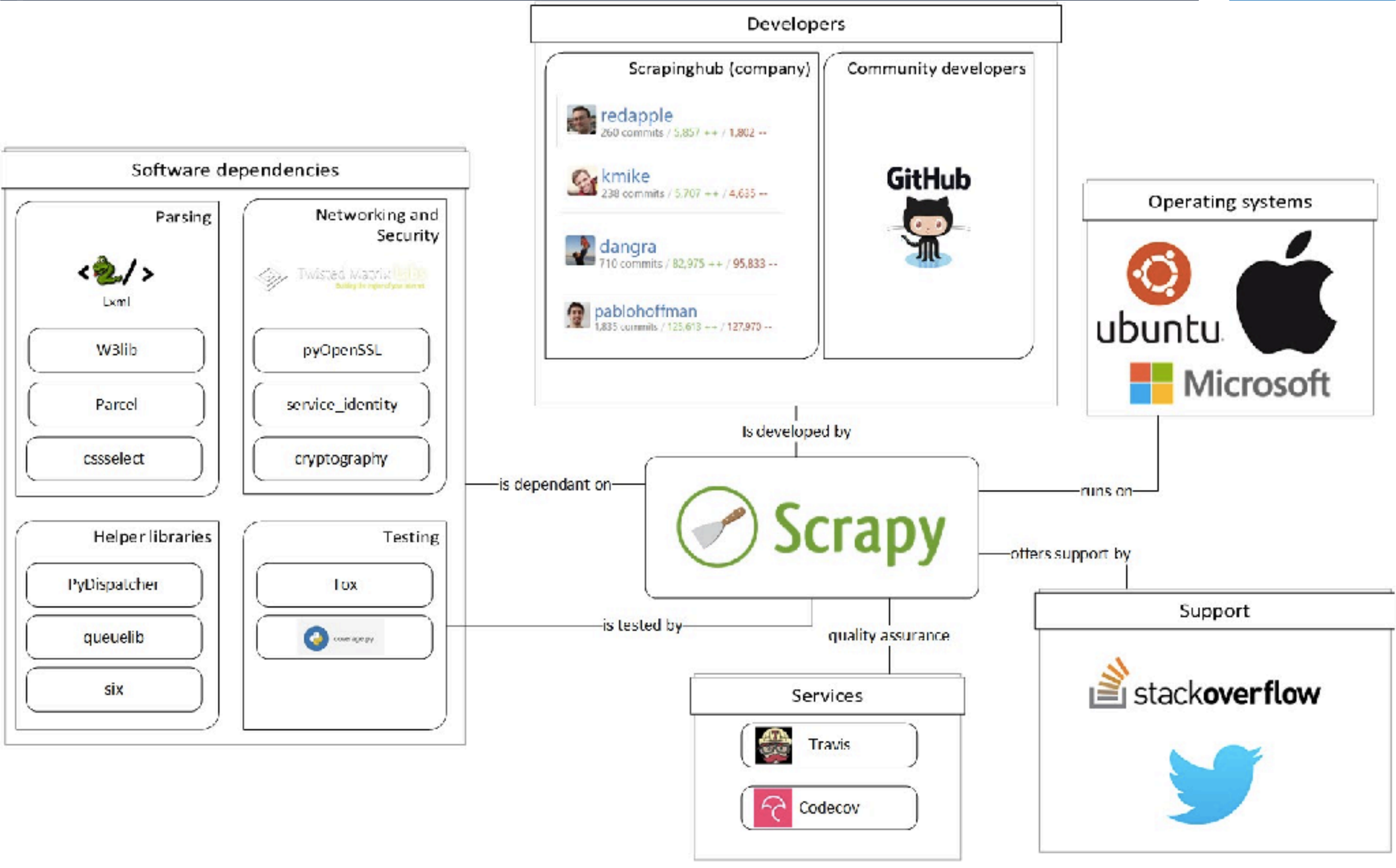
```
import json

 with open(‘data.json','wb') as outfile:
        json.dump(my_json, outfile)
```

Write JSON

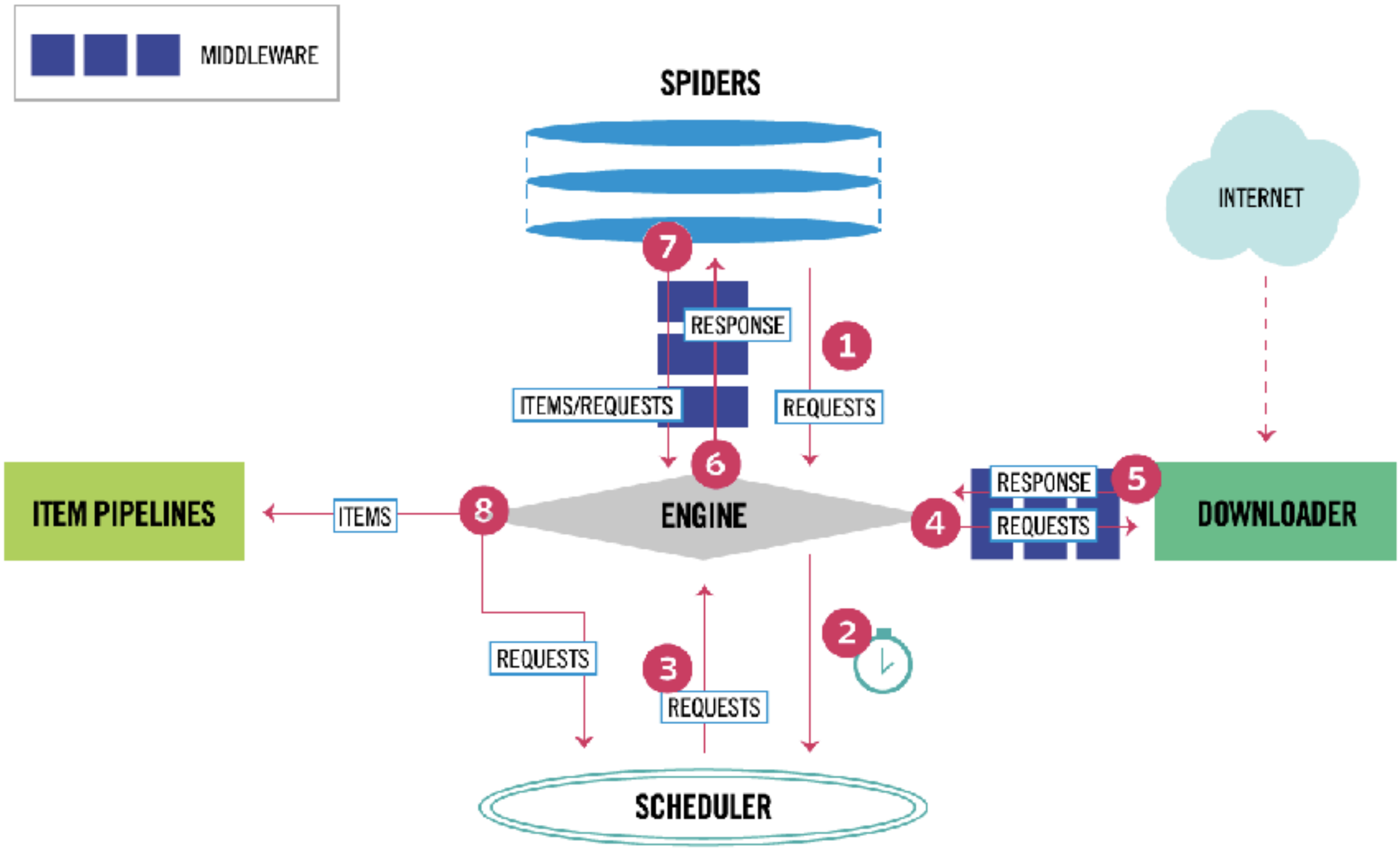
```
import json

 with open(‘data.json','wb') as outfile:
        json.dump(my_json, outfile)
```


Scrapy

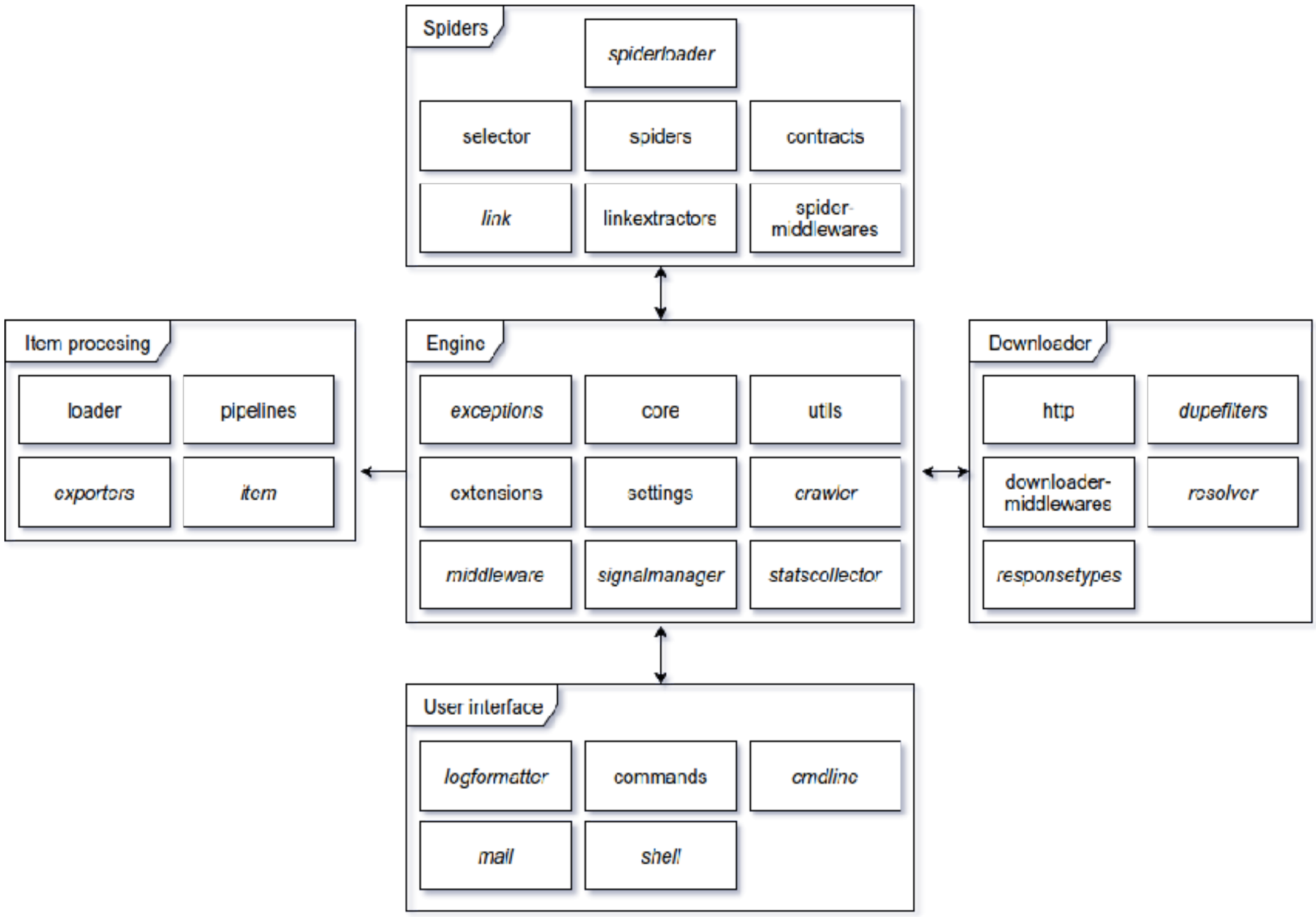
Data flow Scrapy



Data flow Scrapy



Data flow Scrapy



Data flow Scrapy