

# 上市公司财报数据采集

爬取中国国贸(SH600007) 资产负债表。

```
In [1]: from selenium import webdriver
        from selenium.webdriver.common.by import By
        from selenium.webdriver.chrome.options import Options
        from selenium.webdriver.chrome.service import Service
        import pandas as pd
        import time
```

```
In [2]: options = Options()
        options.add_argument('--headless=new')
        service = Service(r'D:\mypython\myprojects\env\chromedriver-win64\chromedriver.exe')
        browser = webdriver.Chrome(service=service, options=options)
```

```
In [3]: browser.get('https://emweb.securities.eastmoney.com/NewFinanceAnalysis/Index?type=w
        time.sleep(3)
        element = browser.find_element(By.CSS_SELECTOR, '#report_zcfzb')
        td_content = element.find_elements(By.TAG_NAME, 'td')
```

```
In [4]: lst = []
        for td in td_content:
            lst.append(td.text)
        lst[:12]
```

```
Out[4]: ['流动资产',
        '',
        '',
        '',
        '',
        '',
        '货币资金',
        '39.04亿',
        '34.85亿',
        '34.69亿',
        '30.69亿',
        '44.42亿']
```

```
In [5]: col = 6
        lst2 = [lst[i:i + col] for i in range(0, len(lst), col)]
        lst2[:5]
```

```
Out[5]: [['流动资产', '', '', '', '', ''],
        ['货币资金', '39.04亿', '34.85亿', '34.69亿', '30.69亿', '44.42亿'],
        ['', '', '', '', '', ''],
        ['', '', '', '', '', ''],
        ['', '', '', '', '', '']]
```

```
In [6]: lst3 = [row for row in lst2 if row != ['', '', '', '', '', '']]
        lst3[:5]
```

```
Out[6]: [['流动资产', '', '', '', '', ''],
         ['货币资金', '39.04亿', '34.85亿', '34.69亿', '30.69亿', '44.42亿'],
         ['应收票据及应收账款', '2.368亿', '2.104亿', '2.238亿', '2.276亿', '2.127亿'],
         ['其中:应收账款', '2.368亿', '2.104亿', '2.238亿', '2.276亿', '2.127亿'],
         ['预付款项', '1934万', '3486万', '1521万', '1838万', '1891万']]
```

```
In [7]: df_table = pd.DataFrame(lst3)
df_table.head()
```

```
Out[7]:
```

	0	1	2	3	4	5
0	流动资产					
1	货币资金	39.04亿	34.85亿	34.69亿	30.69亿	44.42亿
2	应收票据及应收账款	2.368亿	2.104亿	2.238亿	2.276亿	2.127亿
3	其中:应收账款	2.368亿	2.104亿	2.238亿	2.276亿	2.127亿
4	预付款项	1934万	3486万	1521万	1838万	1891万

```
In [8]: df_table = df_table.iloc[1:, [0, 1, 2, 3, 4, 5]]
df_table.columns = ['项目', '月份1', '月份2', '月份3', '月份4', '月份5']
df_table.index.name = '序号'
df_table = df_table.replace('--', 0)
df_table.head(10)
```

```
Out[8]:
```

	项目	月份1	月份2	月份3	月份4	月份5
序号						
1	货币资金	39.04亿	34.85亿	34.69亿	30.69亿	44.42亿
2	应收票据及应收账款	2.368亿	2.104亿	2.238亿	2.276亿	2.127亿
3	其中:应收账款	2.368亿	2.104亿	2.238亿	2.276亿	2.127亿
4	预付款项	1934万	3486万	1521万	1838万	1891万
5	其他应收款合计	609.4万	557.9万	754.4万	629.6万	587.7万
6	存货	2685万	2777万	2873万	2920万	2937万
7	其他流动资产	17.17万	17.17万	17.17万	17.17万	18.85万
8	流动资产合计	41.94亿	37.64亿	37.45亿	33.51亿	47.09亿
9	非流动资产					
10	长期股权投资	3261万	3118万	3185万	3025万	2921万

```
In [9]: browser.quit()
```