# 湖南大学实训报告

课程名称：计算与人工智能概论　实验类型：团队实训

实验项目名称：世界 500 强公司信息爬取

学生姓名：苏晨阳　班级：2402　学号：202410040223

同组姓名：　班级：　学号：

## 一、实训目的

1．巩固计算与人工智能概念课程所学基础知识

2．拓展训练学生计算和 AI 思维能力

3．加强团队分工与合作，培养学生团队协作能力

## 二、实训内容和团队分工

1．实训内容：获取注册资金 500 强公司的名字和注册资金，并通过 matplotlib 绘制出注册资金最多的公司 top20

2.团队分工：苏晨阳负责全部

## 三、实训环境

库：

requests

re

函数

Findall ( )

## 四、实训方法和步骤

步骤 1：导入 requested 库

```
#### 代码窗口
#首先我们需要导入 requests 库
import requests
```

步骤 2：找到请求 url，确认请求模式

```
# 请求的url
url = "https://top.chinaz.com/gongsitop/index_500top.html"
# 设置请求头信息
headers = {
    "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/95.0.4638.69 Safari/537.36"
}
# 使用reqeusts模快发起 GET 请求
response = requests.get(url, headers=headers)
# 获取请求的返回结果
html = response.text
```

步骤 3：获取结果并解析，考虑使用正则表达式解析工具，因此也需要导入 python 中内置的 re 模块来实现

```
# 导入 re 模快
import re
```

```
# 存储内容
message = []
# 总共16个页面的数据
for page in range(16):
    # 组装url
    if page == 0:
        url = "https://top.chinaz.com/gongsitop/index_500top.html"
    else:
        url = "https://top.chinaz.com/gongsitop/index_500top_{}.html".format(page + 1)
    # 使用reqeusts模快发起 GET 请求
    response = requests.get(url, headers=headers)
    html = response.text
    # 使用 findall 函数来获取数据
    # 公司名
    company = re.findall('<a.*?target="_blank">(.+?)</a></h3>', html)
    #注册资本
    registered=re.findall('注册资本：</span>(.*?)</p>',html)
    pageOne = list(zip(company,registered))
    # 合并列表
    message.extend(pageOne)
```

步骤 4：保存内容

```
# 导入python中的内置模块csv
import csv
with open("content.csv", "w") as f:
    w = csv.writer(f)
    w.writerows(message)
```

```
In [7]: !cat content.csv
```

```
中国石油化工股份有限公司,1210.71亿元
中国石油天然气股份有限公司,1830.21亿元
中国建筑股份有限公司,300亿元
中国工商银行股份有限公司,3493.212346亿人民币
上海汽车集团股份有限公司,116.83亿元
中国移动通信有限公司,164184.83万元人民币
中国中铁股份有限公司,228.44亿元
中国平安保险(集团)股份有限公司,182.8亿元
中国建设银行股份有限公司,25001097.748600万人民币
中国铁建股份有限公司,135.8亿元
中国农业银行股份有限公司,3247.94亿元
中国人寿保险股份有限公司,282.65亿元
中国银行股份有限公司,2943.88亿元
中国人民保险集团股份有限公司,4242399.0583万人民币
中国交通建设股份有限公司,161.75亿元
中国中信集团有限公司,18419815.685903万人民币
中国联合网络通信股份有限公司,211.97亿元
中国太平洋保险(集团)股份有限公司,90.62亿元
中国中车股份有限公司,286.99亿元
```

步骤 5：数据可视化

```
import pandas as pd

# 读取数据
df = pd.read_csv("content.csv", names=["company", "registered"])
df.head()
```

| | company | registered |
|---|---|---|
| 0 | 中国石油化工股份有限公司 | 1210.71亿元 |
| 1 | 中国石油天然气股份有限公司 | 1830.21亿元 |
| 2 | 中国建筑股份有限公司 | 300亿元 |
| 3 | 中国工商银行股份有限公司 | 3493.212346亿人民币 |
| 4 | 上海汽车集团股份有限公司 | 116.83亿元 |

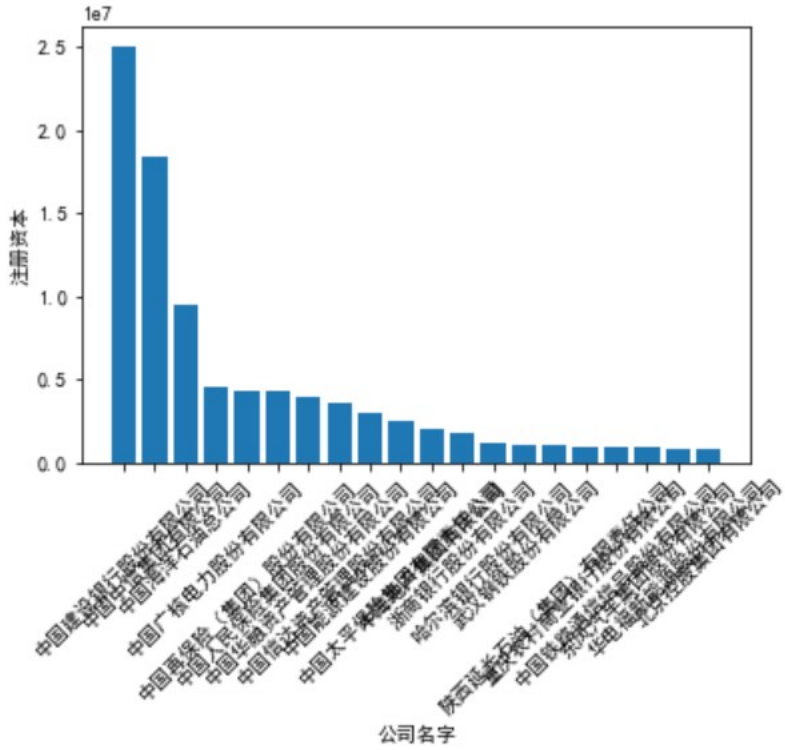
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 474 entries, 0 to 473
Data columns (total 2 columns):
company       474 non-null object
registered    405 non-null object
dtypes: object(2)
memory usage: 7.5+ KB
```

```
[0]:
# 在jupyter中直接展示图像
%matplotlib inline
import matplotlib.pyplot as plt

# 用黑体显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']

#
df['registered']=df['registered'].astype(str)
df['registered'] = df['registered'].str.extract('(\d+(?:\.\d+)?)',expand=False).astype(float)

#降序
df_sorted=df.sort_values(by='registered',ascending=False)

#提取前20位
top_20=df_sorted.head(20)

#柱状图
plt.bar(top_20['company'],top_20['registered'])
plt.xticks(rotation=45)
plt.xlabel('公司名字')
plt.ylabel('注册资本')

#
plt.show
```

# 五、实训结果和分析

<function matplotlib.pyplot.show(close=None, block=None)>




分析：

在进行爬虫工作时仍然需要我们获得很多的前置信息，在掌握了这些信息之后我们便能很好的爬取信息，requests 等工具很好的帮助我们建立起了爬虫的前置环境与条件，只需善于利用它便可，此外，能够清晰有条理的呈现数据也是很重要的一件事，利用饼图，柱状图等图表工具是个不错的选择

## 六、讨论与心得

1.合理利用工具可以帮助我们高效的搜集信息，磨刀不误砍柴工
2.数据具有多样性，需要充分考虑同类数据的不同呈现方式以便获得完整且真实的信息
3.需重视隐私规则，避免因为爬取而违规泄密