1. 因为英语阅读基础薄弱，所以看英文论文的时候有些费力，反反复复看了很多遍，最后也只在这一句发现，“While we identify individual companies with actual product-market relationships, Shahrur uses the benchmark input–output matrices for the U.S.economy to identify upstream and downstream industries. ”可以通过**“基准投入产出矩阵”**来判断企业的上下游关系。但是英文论文理解还是有点苦难，明天会再看看这篇论文中有没有可用的技巧被漏掉了。
2. 然后看了很多其他与Compustat数据集相关的论文，但是暂时也还没有找到其他作者是如何辨别公司名称的，自己先简单写了个**模糊查询**的代码,*明天加入大小写的识别*:

```
   import re

file_list = [
    {
        "type": "dir",
        "size": "123",
        "name": "access.log",
    },
    {
        "type": "dir",
        "size": "123",
        "name": "access.log.gz",
    },
    {
        "type": "dir",
        "size": "123",
        "name": "error.log",
    },
    {
        "type": "dir",
        "size": "123",
        "name": "access-auth.log",
    },
]


def fuzzy_finder(key, data):
    """
    模糊查找器
    :param key: 关键字
    :param data: 数据
    :return: list
    """
    # 结果列表
    suggestions = []
    # 非贪婪匹配，转换 'djm' 为 'd.*?j.*?m'
    # pattern = '.*?'.join(key)
    pattern = '.*%s.*' % key
    # print("pattern",pattern)
    # 编译正则表达式
    regex = re.compile(pattern)
    for item in data:
        # print("item",item['name'])
```

```
        # 检查当前项是否与regex匹配。
        match = regex.search(item['name'])
        if match:
            # 如果匹配，就添加到列表中
            suggestions.append(item)

    return suggestions

# 搜索关键字
keys = "access"
result = fuzzy_finder(keys, file_list)
print(result)
```

3. 为了了解到其他作者是怎么在Compustat数据集中识别公司名称，阅读了使用该数据集论文有：

- 基于自然语言处理及LSTM模型的产业上下游关系识别
- 上下游企业关联度与企业营运资金、股利分配和财务风险的关系——基于中国制造业上市公司数据的实证分析
- 产业链投资量化策略的研究
- 小型公司的集体战略
- 中美企业ESG评估与改善 ——以中国海油和美国康菲为例
- 基于竞争与创新的反转策略投资组合研究
- Sources of gains in horizontal mergers: evidence from customer, supplier, andrival firms

  今天并没有在上述论文中找到代码的突破口，而且阅读论文的速度有点慢，可能因为对金融的专业术语用词不准确，明天会提高阅读速度，也会把上面的文章重新阅读一次，并再看一些相关的论文找到代码的突破口