1. 数据下载
2. 数据概述

根据作业要求，我们需要至少500份企业年报，考虑到中国A股市场的特点，市值较大企业的信息披露较为完善、财务造假风险较小，我们选择沪深300指数成分股作为股票域。

由于作业要求中涉及未来一年和未来两年的预测，因此我们需要预留2022年、2023年作为预测的年份，因此我们下载沪深300成分股2020和2021年的年报，其未来一年收益率分别对应2021年和2022年，未来两年收益率分别对应2022年和2023年。

1. 成分股数据

如上所述，我们从同花顺下载2024年8月4日沪深300指数的成分股数据，下载的数据如表1所示：

表 1 沪深300成分股数据

|  |  |  |
| --- | --- | --- |
| code\_full | name | code |
| 000001.SZ | 平安银行 | 000001 |
| 000002.SZ | 万科A | 000002 |
| 000063.SZ | 中兴通讯 | 000063 |
| 000100.SZ | TCL科技 | 000100 |
| 000166.SZ | 申万宏源 | 000166 |

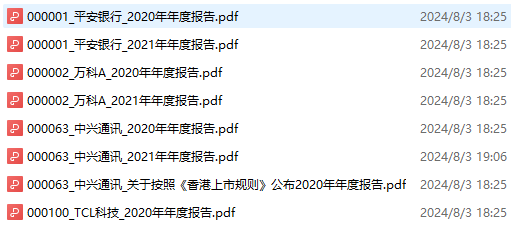
1. 年报数据

我们使用Python爬虫下载个股年报，使用巨潮资讯网，网址为：

<http://www.cninfo.com.cn/new/commonUrl/pageOfSearch?url=disclosure/list/search&lastPage=index>

通过爬虫，我们可以获取个股年报的下载链接，再使用Python的request命令下载对应链接下的文件即可。最终，下载好的文件如图1所示：

图 1 年报示例



1. 未来收益数据

如上所述，对于2020年的年报，我们下载2021年和2022年的收益率，对于2021年的年报，我们下载2022年和2023年的年报，最终结果如表2所示：

表 2 未来收益率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| code\_full | name | code | 2021 | 2022 | 2023 |
| 000001.SZ | 平安银行 | 000001 | -0.1327 | -0.1763 | -0.2416 |
| 000002.SZ | 万科A | 000002 | -0.2213 | -0.0216 | -0.2670 |
| 000063.SZ | 中兴通讯 | 000063 | 0.0014 | -0.2102 | 0.0369 |
| 000100.SZ | TCL科技 | 000100 | -0.1009 | -0.3265 | 0.2130 |
| 000166.SZ | 申万宏源 | 000166 | -0.0103 | -0.1860 | 0.1069 |

1. 数据检验

对于每一只股票，检验是否正确下载其年报文件。

如上所述，每一只股票应该对应2份年报，共600份，由于股票列表是2024年8月3日的数据，部分股票在2020、2021年可能还未上市，因此我们下载的文件数量为555份。

检查各文件名确认其为公司年报，发现如下几个异常情况：

1. 部分报告名为《×××持续督导年度报告》混入其中，故删除；
2. 部分公司的年报文件名以“年度报告全文”结尾，因此在之前爬虫的逻辑中并未考虑，由于年报数量已超过500份，故此部分年报未补充下载。
3. 部分公司年报包括境内（大陆）和境外（港股）的公示，我们认为这些报告均具有一定价值，因此保留了境外公示年报。

最终，我们一共有551份有效年报。

1. 数据处理

对于下载好的公司年报数据，文件格式为pdf，我们对其进行分词，并转换为txt文件，为后续建模和预测做准备。

对于每一份pdf文件，我们按顺序进行如下处理：

1. 使用Python的fitz库读取pdf文件中的文本，过滤表格、图片等内容。
2. 使用Python的re正则表达式过滤文本中的空字符。
3. 使用Python的jieba对文本进行分词。
4. 根据停用词表去除停用词。
5. 保存为txt文件。

最终，得到的数据如图2所示：

图 2 文本数据

