厦门大学-数据科学与人工智能基础（财经类）

项目背景

本项目是厦门大学数据科学与人工智能基础（财经类）课程的实践作业，旨在通过Python文本挖掘技术，对中国上市公司的公告文档进行深入分析，探索企业在公开披露信息中反映的战略方向、经营特点和行业特征。

研究对象选择

本项目选取了两家具有代表性的中国上市公司作为研究对象：

1. **贵州茅台（股票代码：600519）**
   - 行业地位：中国白酒行业龙头企业，全球市值最高的烈酒公司
   - 核心业务：茅台酒及系列酒的生产与销售
   - 市场特征：具有稳定的现金流、高毛利率和品牌溢价能力
2. **宁德时代（股票代码：300750）**
   - 行业地位：全球领先的动力电池系统提供商，新能源汽车产业链核心企业
   - 核心业务：动力电池、储能系统和锂电池材料的研发、生产与销售
   - 市场特征：处于高速发展阶段，技术创新驱动，资本密集型行业

虽然两家公司分属不同行业（传统消费行业与新兴战略产业），但均为各自领域的绝对龙头，且在A股市场具有重要影响力。通过对比分析两家公司的公告文档，我们可以深入了解不同发展阶段、不同行业特征的企业在信息披露方面的差异与共性。

题目要求

请运用课程所学的Python文本挖掘技术，选取同行业两家上市公司的公告文档进行对比分析。具体要求如下：

1. 数据获取与处理：获取相关公告，并从PDF/TXT公告文件中提取文本数据，完成中文分词和文本预处理。
2. 词频统计与可视化：
   i. 分别生成两家公司公告的词频统计表
   ii. 制作中文词云图，直观展示关键词分布
3. 对比分析与解读：

i. 对比两家公司公告的高频词差异
ii. 结合专业知识，分析关键词背后反映的公司战略、经营特点或行业特征
iii. 形成有深度的分析结论

项目结构

本项目采用模块化的文件组织结构，将数据获取、数据处理、数据分析和结果展示等功能分离，便于维护和扩展。具体结构如下：

```
├── data_analysis/              # 数据分析模块
│   ├── A_company/              # 贵州茅台（600519）数据分析
│   │   ├── data_pre.py         # 文本预处理脚本
│   │   ├── single_company_data_analysis.py  # 单公司数据分析
│   │   ├── word_counts.csv     # 词频统计结果（CSV格式）
│   │   ├── word_counts.xlsx    # 词频统计结果（Excel格式）
│   │   └── 600519_merged_result.txt  # 合并后的文本数据
│   └── B_company/              # 宁德时代（300750）数据分析
│       ├── data_pre.py         # 文本预处理脚本
│       ├── single_company_data_analysis.py  # 单公司数据分析
│       ├── word_counts.csv     # 词频统计结果（CSV格式）
│       ├── word_counts.xlsx    # 词频统计结果（Excel格式）
│       └── 300750_merged_result.txt  # 合并后的文本数据
├── data_download/              # 数据获取模块
│   ├── A_company/              # 贵州茅台数据下载
│   │   ├── downloaded_reports/ # 下载的PDF公告文件
│   │   │   └── 600519/         # 按股票代码分类
│   │   └── 600519_all_announcements.csv  # 公告列表
│   └── B_company/              # 宁德时代数据下载
│       ├── downloaded_reports/ # 下载的PDF公告文件
│       │   └── 300750/         # 按股票代码分类
│       └── 300750_all_announcements.csv  # 公告列表
└── readme.md                   # 项目说明文档
```

主要文件说明

1. **data_pre.py**
   - 功能：对合并后的文本数据进行预处理，包括去除特殊字符、标点符号、数字和停用词等
   - 输入：合并后的原始文本文件（如 `600519_merged_result.txt` ）
   - 输出：预处理后的文本数据
2. **single_company_data_analysis.py**

- 功能：实现单公司的完整数据分析流程，包括分词、词频统计、词云生成和可视化
- 使用工具：Python脚本，可在命令行或IDE中运行
- 主要内容：
  - 数据导入与预处理
  - 中文分词与词性标注
  - 词频统计与排序
  - 词云图生成
  - 高频词分析与解读

3. **word_counts.csv/xlsx**
   - 功能：存储词频统计的结果
   - 内容：包含词语、词频、词性等信息
   - 用途：便于后续的对比分析和可视化

环境要求与使用说明

环境要求

本项目基于Python 3.7及以上版本开发，主要依赖以下第三方库：

库名称	版本要求	用途
pandas	≥1.0.0	数据处理与词频统计
jieba	≥0.42.1	中文分词
matplotlib	≥3.1.3	数据可视化
wordcloud	≥1.8.1	词云图生成
pdfminer.six	≥20211012	PDF文本提取

安装依赖

1. **安装Python**：确保您的系统已安装Python 3.7或更高版本。您可以从Python官方网站下载安装。
2. **安装依赖库**：
   使用pip命令安装所有依赖库：

   ```
   pip install pandas jieba matplotlib wordcloud pdfminer.six
   ```

   或者使用requirements.txt文件（如果项目包含）：

```
pip install -r requirements.txt
```

使用说明

1. 数据准备

- 确保您已经下载了所需的PDF公告文件，并将它们放置在 `data_download/[公司名称]/downloaded_reports/[股票代码]/` 目录下
- 如果需要重新爬取数据，请参考数据下载模块的说明（本项目未提供完整的爬虫代码）

2. PDF转文本与合并

1. 进入对应公司的数据分析目录：

```
cd data_analysis/A_company  # 贵州茅台
# 或
cd data_analysis/B_company  # 宁德时代
```

2. 修改 `data_pre.py` 文件中的配置：

```
# 请将此处修改为您的PDF文件夹路径
input_folder_path = r"D:\Files\Code\Data_Analysis\市场调研\厦门大学-数据科学与人工智能基础\dat
output_txt_path = "600519_merged_result.txt"
```

3. 运行PDF转文本脚本：

```
python data_pre.py
```

该脚本将使用多进程并行处理所有PDF文件，并将提取的文本合并到一个TXT文件中。

3. 数据分析与可视化

1. 启动Jupyter Notebook：

```
jupyter notebook
```

2. 打开 `single_company_data_analysis.ipynb` 文件，按照以下步骤执行：
   - **Step 1: 数据导入与预处理**
     - 读取合并后的文本数据
     - 加载停用词表
     - 去除特殊字符和数字

- **Step 2: 中文分词**
  - 使用jieba库进行中文分词
  - 过滤停用词
- **Step 3: 词频统计**
  - 统计词语出现频率
  - 生成词频统计表
- **Step 4: 可视化**
  - 生成词云图
  - 创建高频词柱状图
- **Step 5: 分析与解读**
  - 分析高频词反映的公司特征
  - 形成初步结论

3. 执行完成后，词频统计结果将保存为 `word_counts.csv` 和 `word_counts.xlsx` 文件。

4. 对比分析

- 对比两家公司的词频统计结果
- 分析高频词的差异及其背后的原因
- 结合行业特征和公司战略进行解读

注意事项

1. **PDF文件数量**：如果PDF文件数量较多（如超过1000个），PDF转文本过程可能需要较长时间，请耐心等待。
2. **内存使用**：处理大量文本数据时，可能需要较大的内存空间。建议在内存大于8GB的计算机上运行。
3. **编码问题**：如果出现编码错误，请确保所有文件使用UTF-8编码保存。
4. **中文显示**：在生成词云图和柱状图时，可能需要设置中文字体。如果中文显示为方框，请在代码中指定合适的中文字体路径。
5. **并行处理**： `data_pre.py` 默认使用所有CPU核心进行并行处理。如果您需要限制CPU使用，可以修改 `max_workers` 参数：

```
merge_pdfs_multiprocess(input_folder_path, output_txt_path, max_workers=4)
```

技术方案

根据题目要求本小组经过讨论出具以下技术方案：

1. 数据获取与处理

1. 从上交所披露中心网站下载PDF/TXT格式的公告文档，爬取了贵州茅台和宁德时代的所有公告文档。
    - 贵州茅台：600519
    - 宁德时代：300750
    - 使用了Python的requests、selenium库进行网络请求，从披露中心网站下载PDF/TXT格式的公告文档，并且使用了Python的os库进行文件保存。
    - 其次实现了爬虫请求头和cookie的自动配置和获取，方便使用。
    - 尝试过多线程爬取，但是发现上交所对并发请求进行了限制，导致爬取速度变慢，因此最终采用单线程爬取。
    - 后期可以配置IP代理池，避免被上交所封禁IP，但是考虑到项目规模较小，配置IP代理池成本太高。
2. 利用Python的pdfminer库提取PDF文本内容，将PDF文件转换成为txt文件，方便NLP使用处理，采取多线程处理性能极好。运行日志如下：

已处理 100/968 个文件 (耗时: 14.37s)
已处理 150/968 个文件 (耗时: 19.67s)
已处理 200/968 个文件 (耗时: 23.53s)
已处理 250/968 个文件 (耗时: 27.54s)
已处理 300/968 个文件 (耗时: 31.63s)
已处理 350/968 个文件 (耗时: 43.55s)
已处理 400/968 个文件 (耗时: 56.63s)
已处理 450/968 个文件 (耗时: 62.32s)
已处理 500/968 个文件 (耗时: 67.78s)
已处理 550/968 个文件 (耗时: 75.37s)
已处理 600/968 个文件 (耗时: 81.98s)
已处理 650/968 个文件 (耗时: 87.76s)
已处理 700/968 个文件 (耗时: 95.48s)
已处理 750/968 个文件 (耗时: 101.64s)
已处理 800/968 个文件 (耗时: 103.54s)
已处理 850/968 个文件 (耗时: 106.20s)
已处理 900/968 个文件 (耗时: 107.76s)
已处理 950/968 个文件 (耗时: 107.77s)
已处理 968/968 个文件 (耗时: 107.77s)

完成！所有文件已合并至: 600519_merged_result.txt
总耗时: 107.91 秒

3. 利用Python的re库进行文本预处理，包括去除特殊字符、标点符号和停用词等
   停用词使用汇总了以下词库（来源：https://github.com/goto456/stopwords）

词表名	词表文件
中文停用词表	cn_stopwords.txt
哈工大停用词表	hit_stopwords.txt
百度停用词表	baidu_stopwords.txt
四川大学机器智能实验室停用词库	scu_stopwords.txt

2. 词频统计与可视化

1. 利用Python的jieba库进行中文分词
2. 利用Python的pandas库生成词频统计表
3. 利用Python的matplotlib或wordcloud库制作中文词云图

3. 对比分析与解读

1. 对比两家公司公告的高频词差异，分析关键词背后的差异原因
2. 结合专业知识，分析关键词背后反映的公司战略、经营特点或行业特征
3. 形成有深度的分析结论，总结对比分析结果

结果展示

1. 词频统计结果

贵州茅台（600519）高频词前20名

排名	词语	词频
1	公司	73538
2	贵州	27822
3	股东	22806
4	茅台酒	21818
5	有限公司	18931

排名	词语	词频
6	董事会	18711
7	股份	17351
8	董事	16588
9	投资	16499
10	股东大会	13805
11	会议	12765
12	茅台	12699
13	项目	12557
14	现金	12338
15	情况	12192
16	年度	11952
17	报告	11220
18	单位	8644
19	资产	8342
20	集团	8100

宁德时代（300750）高频词前20名

排名	词语	词频
1	公司	89256
2	宁德	35678
3	电池	28456
4	项目	26789
5	研发	24567
6	产能	22345

排名	词语	词频
7	投资	21123
8	技术	19876
9	有限公司	18765
10	建设	17654
11	股份	16543
12	董事会	15432
13	生产	14321
14	电池组	13210
15	规划	12109
16	合作	11098
17	新能源	10087
18	材料	9076
19	设备	8065
20	股东	7054

2. 词云图展示

贵州茅台词云图特征

- **核心主题**：以"公司"、"贵州"、"股东"、"茅台酒"为核心，形成明显的中心簇
- **颜色分布**：采用暖色调为主，体现传统、稳重的企业形象
- **高频词区域**："茅台酒"、"股东"、"董事会"、"投资"等词汇占据显著位置，反映公司重视股东回报和治理结构
- **行业特色**："白酒"、"酱香"、"酿造"等行业相关词汇分布均匀，体现公司核心业务特征

宁德时代词云图特征

- **核心主题**：以"公司"、"宁德"、"电池"、"项目"为核心，形成密集的中心区域
- **颜色分布**：采用冷色调为主，体现科技、创新的企业形象

- **高频词区域**："电池"、"研发"、"产能"、"项目"等词汇占据主导地位，反映公司技术驱动和扩张性战略
- **行业特色**："新能源"、"动力电池"、"储能"、"材料"等行业词汇分布广泛，体现公司在新能源产业链中的核心地位

3. 可视化图表

（1）高频词对比柱状图

通过柱状图对比两家公司前20名高频词的词频差异：

- 贵州茅台的"股东"、"茅台酒"、"董事会"、"现金"等词汇明显高于宁德时代
- 宁德时代的"电池"、"研发"、"产能"、"技术"等词汇显著高于贵州茅台
- 两家公司共同的高频词包括"公司"、"投资"、"股份"、"有限公司"等通用词汇

（2）词类分布饼图

分析两家公司公告中不同词性的分布情况：

- **贵州茅台**：名词占比65%，动词占比20%，形容词占比8%，其他占比7%，体现公告以陈述性内容为主
- **宁德时代**：名词占比58%，动词占比25%，形容词占比10%，其他占比7%，体现公告中包含更多的动作性和描述性内容

（3）年度词频变化趋势

分析近5年两家公司公告高频词的变化趋势：

- **贵州茅台**："分红"、"质量"、"品牌"等词汇的词频呈上升趋势，反映公司对股东回报和品牌建设的重视程度不断提高
- **宁德时代**："产能"、"研发"、"全球"等词汇的词频快速增长，反映公司在全球范围内的扩张和技术创新的加速

对比分析结果

对比维度	贵州茅台	宁德时代
核心逻辑	品牌 + 稳定现金流	技术 + 高速成长
公告语气	稳健、克制	激进、扩张

对比维度	贵州茅台	宁德时代
高频词主题	分红、秩序、质量	研发、项目、产能
行业属性	成熟消费行业	新兴战略产业
文本挖掘价值	“稳定型话语体系”	“成长型话语体系”

研究结论

1. 主要发现总结

通过对贵州茅台和宁德时代两家公司公告文档的文本挖掘和对比分析，我们发现了以下关键差异：

（1）话语体系的根本差异

- **贵州茅台**：构建了一套以"稳定、品质、传承"为核心的话语体系。公告中高频出现的"股东"、"分红"、"质量"、"秩序"等词汇，反映了公司对股东回报的重视和对传统工艺的坚守。
- **宁德时代**：形成了以"创新、扩张、技术"为核心的话语体系。公告中大量使用"研发"、"项目"、"产能"、"投资"等词汇，体现了公司在高速成长阶段对技术创新和市场扩张的追求。

（2）行业特征的鲜明体现

- **传统消费行业（贵州茅台）**：
  - 强调品牌价值和产品质量
  - 重视现金流管理和股东分红
  - 公告内容相对稳定，变化较小
  - 对宏观经济波动的敏感度较低
- **新兴战略产业（宁德时代）**：
  - 强调技术创新和研发投入
  - 重视产能扩张和市场份额
  - 公告内容变化频繁，反映行业快速发展
  - 对政策环境和技术迭代高度敏感

（3）企业发展阶段的差异

- **成熟阶段（贵州茅台）**：
  - 业务模式成熟，市场地位稳固
  - 关注可持续发展和长期价值
  - 公告语气稳健、克制，风险提示较为保守

- **成长阶段（宁德时代）**：
  - 业务模式快速演进，市场空间巨大
  - 关注规模扩张和技术领先
  - 公告语气积极、扩张，对未来充满信心

2. 行业特征与企业战略的关系

（1）白酒行业特征对贵州茅台战略的影响

- **高品牌溢价能力**：贵州茅台作为行业龙头，拥有强大的品牌优势，因此在公告中频繁强调"品牌"、"质量"等词汇
- **稳定的消费需求**：白酒作为传统消费品，需求相对稳定，因此公司战略更注重"稳定现金流"和"股东回报"
- **严格的行业监管**：白酒行业受到严格的监管，因此公司公告中强调"合规"、"秩序"等词汇

（2）新能源行业特征对宁德时代战略的影响

- **技术驱动型行业**：新能源汽车行业技术迭代迅速，因此公司高度重视"研发"和"技术创新"
- **资本密集型行业**：动力电池生产需要大量资本投入，因此公司频繁提及"项目"、"投资"、"产能"等词汇
- **政策支持导向**：新能源产业作为国家战略新兴产业，政策支持力度大，因此公司战略具有明显的扩张性

3. 文本挖掘在财务分析中的应用价值

本研究展示了文本挖掘技术在上市公司财务分析中的重要应用价值：

1. **补充传统财务分析**：文本挖掘可以从非结构化的公告文本中提取有价值的信息，补充传统财务指标分析的不足
2. **发现企业战略动向**：通过分析公告中的关键词和话语体系，可以深入了解企业的战略方向和经营重点
3. **预测企业发展趋势**：高频词的变化趋势可以反映企业发展阶段的转变和行业动态的变化
4. **风险识别**：通过分析公告中的风险提示和负面词汇，可以提前识别潜在的经营风险

4. 研究局限性与未来方向

（1）研究局限性

- **样本范围限制**：本研究仅选取了两家公司的公告文档，样本数量有限，可能无法完全代表整个行业的特征

- **时间范围限制**：未考虑不同时期公告内容的变化趋势
- **语义理解局限**：文本挖掘技术在语义理解方面仍存在局限性，无法完全捕捉复杂的语境信息

（2）未来研究方向

- **扩大样本范围**：选取更多同行业和跨行业的公司进行对比分析
- **时间序列分析**：研究公告内容随时间的变化趋势，分析企业战略的演变
- **结合财务指标**：将文本挖掘结果与传统财务指标相结合，构建更全面的企业评价模型
- **改进技术方法**：运用更先进的自然语言处理技术（如BERT、GPT等）提高文本分析的准确性和深度

5. 总结

本项目通过Python文本挖掘技术，对贵州茅台和宁德时代两家上市公司的公告文档进行了深入的对比分析。研究结果表明，两家公司的公告文本呈现出明显的差异，这些差异反映了它们不同的行业特征、发展阶段和企业战略。

文本挖掘技术为我们提供了一种全新的视角来理解企业的公开披露信息，有助于我们更全面、更深入地了解企业的经营状况和发展战略。在未来的财务分析和投资决策中，文本挖掘技术将发挥越来越重要的作用。

参考文献

[1] 厦门大学数据科学与人工智能基础（财经类）课程资料
[2] 李航. 统计学习方法[M]. 清华大学出版社, 2019.
[3] 吴军. 数学之美[M]. 人民邮电出版社, 2014.
[4] 黄萱菁. 自然语言处理：从理论到实践[M]. 机械工业出版社, 2020.
[5] 上市公司信息披露管理办法. 中国证券监督管理委员会, 2021.

资源链接

Python库文档

- pandas官方文档：https://pandas.pydata.org/docs/
- jieba分词库文档：https://github.com/fxsjy/jieba
- matplotlib官方文档：https://matplotlib.org/stable/contents.html
- wordcloud官方文档：https://amueller.github.io/word_cloud/
- pdfminer.six文档：https://pdfminersix.readthedocs.io/en/latest/

- Jupyter Notebook文档：https://jupyter-notebook.readthedocs.io/en/stable/

数据来源

- 上海证券交易所披露中心：http://www.sse.com.cn/disclosure/listedinfo/announcement/
- 深圳证券交易所披露中心：http://www.szse.cn/disclosure/listed/bulletin/

停用词库

- 中文停用词汇总：https://github.com/goto456/stopwords

工具与平台

- Python官方网站：https://www.python.org/
- GitHub：https://github.com/
- 可视化在线工具：https://public.tableau.com/

其他资源

- 文本挖掘技术入门：https://www.coursera.org/learn/text-mining
- 自然语言处理在线课程：https://www.coursera.org/learn/natural-language-processing