**企业年报数据抽取与分析**

一、数据分析

本文在企业年报数据抽取与分析实验，主要选取了交通运输、仓库和邮政业领域中11家代表性企业，例如韵达控股股份有限公司、中铁特货物流股份有限公司、南京盛航海运股份有限公司、招商局公路网络科技控股股份有限公司等等企业。

数据分析要求具体如下：

A）pdf文件读取：选择合适的pdf文件处理库，读取企业年报，并从年报中提取相关信息：企业名、年报时间、当年营业收入、当年经营成本、当年管理费用、当年研发成本，以及其他您关心的数据；

B）用Pandas将相关数据保存到excel表格；

二、数据分析流程：

（1）导入必要库和函数定义：

本文实验部分使用了Python中pdfplumber、pandas、os、re基本库进行企业年报的数据抽取与分析。其中pdfplumber用于处理PDF文件，re进行正则表达式匹配，pandas进行数据处理，os用于文件操作。

在实验的核心代码中，主要定义了clean\_filename、extract\_value\_after\_keyword，这二个主要函数。其中，定义clean\_filename 函数用于规范文件名；定义extract\_value\_after\_keyword函数用于在文本中查找关键字后的第一个数字。

（2）提取公司信息主函数extract\_company\_info：

在extract\_company\_info函数中，通过文件名正则匹配提取年度报告时间。其次初始化公司信息的变量：company\_name、revenue、cost、management\_expense、r\_d\_cost。使用pdfplumber打开PDF文件。在每一页PDF中搜索关键字，提取公司中文名称和年报时间。而针对营业总收入、营业总成本、管理费用、研发费用等关键信息，在每一页中通过调用extract\_value\_after\_keyword函数提取对应数字。将提取的信息构建为一个Pandas DataFrame。

（3）文件夹分析主函数analyze\_folder：

在analyze\_folder函数中，初始化结果DataFrame result\_df用于存储所有公司信息。然后，使用os.walk遍历指定文件夹下的所有PDF文件。对每个PDF文件，调用extract\_company\_info 函数提取公司信息，并将结果追加到result\_df中。最后，生成输出路径，保存结果DataFrame到Excel文件。

（4）主函数main：

在main函数中，指定实际的输入文件夹路径input\_folder和输出文件夹路径output\_folder。然后，调用analyze\_folder 函数执行文件夹中所有PDF文件的信息提取和保存。

（5）执行main函数与数据输出：

在执行main函数中，实际执行流程，遍历输入文件夹中的PDF文件，提取公司信息，保存结果到指定输出文件夹。最后，以Excel格式保存提取的公司信息，生成文件名为"Corporate\_Finance\_All.xlsx"。