所有代码用jupyter notebook编写

pdf2text.ipynb :从pdf中解析出文本内容，并且放入txt文件夹

modify.ipynb ： 对解析出来的文本信息按句号断行，主要针对的是文件3.pdf、5.pdf，这两个pdf文本读取出来的txt文本仅有两行，所以需要分行处理

wordcut.ipynb :分词处理

LDA_model.ipynb ： model

stopwords.txt ： 停用词文本，里面手动加入了一些停用词，比如五个公司的名称等无关信息，避免对最终的关键词产生影响

文件夹:

PDF：里面包含了五份pdf，重命名为1，2，3，4，5

txt：里面包含了五份pdf提取出来的文本

output ： 里面包含五份分词之后的结果

WordListOutput ： 里面包含要求的10个主题词

FinalOutput ： 里面包含需要的段落，由于pdf文本提取之后的段落不是很干净，需要手动选出段落

附带了一个中文分类的项目例子，用的是Bert模型，也是一个多分类问题，数据集是hugging face的seamew/Weibo 微博数据集,model是bert-base-chinese