一、编程题：

1. 无
2. 见代码
3. 见 WordListOutput 文件
4.

文本段落	一级主题
北辰鹿鸣院获得成都房地产绿色发展示范奖北辰鹿鸣院项目在成都房协首届成都房地产绿色发展示范奖评选中成功获得“绿色发展示范奖”。该项目通过在节约土地资源、优化建筑布局、节约建材、采用节能设备、非传统水源利用、采用节水设备、景观配置、BIM 技术、装配式技术等 9 大方面专项设计的应用，实现可再生资源利用率、减少污染、降低能源消耗，达到北辰鹿鸣院的样板间绿色可持续发展目标。	环境
北辰实业自成立至今，始终秉承公益传统，积极践行社会责任。目前，本公司在参与社会公益方面的主要形式包括实施精准扶贫和公益慈善捐款。精准扶贫助力产业发展 2020 年是脱贫攻坚收官之年，本公司按照市委、市政府、市国资委党委关于精准帮扶工作的具体要求，通过党建共建、消费帮扶、就业帮扶等多个方面，实施帮扶措施，为受援地区实现“造血”发展，加快脱贫致富步伐。	社会
其次，坚守对安全与健深度参与全球竞争，全面向创新型科技企和发展的核心，汽车行业的跨时空巨变一直没有停止，实现中国汽车工业高质量发展	社会
吉利有效利用信息技术、产品处理产生的建筑垃圾、装饰废料等进行合理分类，同时，引入柔性生产技术，通过智能控并妥善处置未靠近自然保护区等生吉利汽车宁波杭州湾第二制造基地通过态敏感区域，不会影响当地生物多样性和全生命周期	环境
通过加强船序管理、优化生产业务	治理

流程、作业人员换班管理、人员备份、“陆转铁”等措施，保障港口作业有序进行，积极降低疫情对生产影响	
社会责任愿景，公司将充分发挥在大连东北亚国际航运中心建设及辽宁沿海经济带发展中的核心与旗舰作用，利用优越的自然条件，发扬自身经营管理优势	社会
对当地社区有实际或潜在重大负面影响的运营供应商社会评估使用社会标准筛选的新供应商供应链对社会的负面影响以及采取的行动对产品和服务类别的健康与安全影响的评估涉及产品和服务的健康与安全影响的违规事件	社会
神马股份不存在任何使用童工、强迫劳动、歧视和骚扰等行为，同时，公司高度重视员工的职业发展规划，倡导职业教育与通用教育相结合、内部教育与外部教育相结合的培训模式，鼓励员工依照自身需求，有针对性地开展精益生产、营销知识、办公技能等多层次多方面的培训内容，以提升员工的综合素质	治理
履行央企社会责任与担当，在陕西省西乡县消费帮扶项目中采购 200 多万元农特产品，持续推进能源管理体系和环境管理体系建设，加大节能环保技术升级	环境

5. 无

二、简答题

1. 已经有这三个一级主题，可以设置成 0，1，2 标签
2. 之后采用 bert-base-chinese 进行 Tokenizer，然后导入 bert-base-chinese Model，定义下游任务，最基础的可以是简单的一个（768-3）的全连接层，然后对每个向量的第一维度 CLS 进行分类即可，或者不采用 CLS 而是 mean method，这些需要根据具体的数据进行调整。
3. 难点在于需要获取已有标签的段落去训练模型进行微调，并且 pdf 的文本提取并不是很干净并且段落划分存在问题，个人感觉跟 pdf 里面的一些格式有关，找了一些处理 pdf 的教程以及库，但对于个别 pdf 会存在效果一般的现象，之后如果段落处理干净并且打上标签，就是一个很简单的 bert 文本分类任务，个人做过一些这方面的项目并且有调参经验，加在了整个文件夹里（文本分类.ipynb），用的是 hugging face 上的微博数据集，也是文本分类，主要用 transformers 的一些库。