海洋舆情项目工作周计划（2 月 21 日——2 月 28 日）

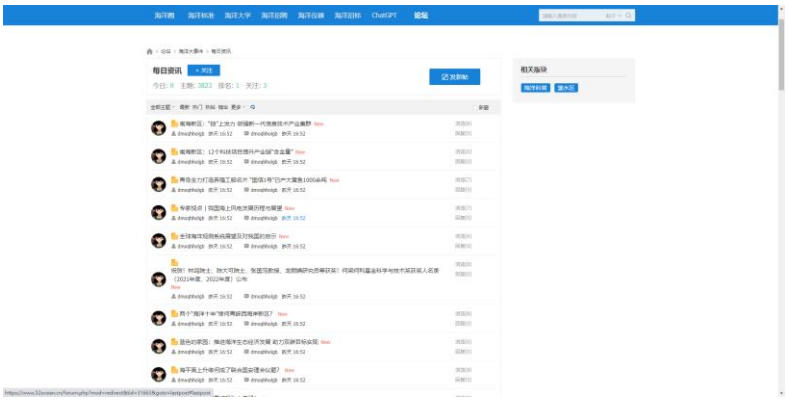
任务一：海洋舆情、政策可爬取网站搜集

任务内容：除甲方提供的几个网站外，搜集了一些含有国内外海洋政策标准的网站（舆情收集平台可能可以从微博等社交平台上获取，由于工作量可能较大暂不考虑），网站由 ChatGPT 查找有待验证：

1. 海洋专业知识服务系统:https://ocean.ckcest.cn/web/index_new.view（除主页面外无服务）
2. 中国海洋信息网:https://www.nmdis.org.cn/
3. 国家海洋环境监测中心:https://www.nmemc.org.cn/
4. 吾爱海洋论坛:https://www.52ocean.cn/
5. 中华人民共和国自然资源部：http://www.mnr.gov.cn/sjzx/hygb/
6. 国家海洋局：http://www.soa.gov.cn/
7. 中国知网：https://www.cnki.net/
8. 中国法律文献库：http://www.pkulaw.cn/more/marine.html
9. 海洋与渔业标准信息服务平台：http://standard.soa.gov.cn/
10. 联合国海洋法公约：https://www.un.org/Depts/los/index.htm
11. 国际海事组织：https://www.imo.org/
12. 国际海洋标准化组织：https://www.iso.org/committee/6545330.html
13. 欧盟海洋政策：https://ec.europa.eu/maritimeaffairs/policy_en
14. 美国国家海洋和大气管理局：https://www.noaa.gov/

任务二：爬虫初步实施

任务内容：一个舆情网站的信息爬取，如吾爱海洋论坛的海洋咨询模块。



任务三：政策文件 pdf 转文本

任务内容：收集一月二月政策文件，这部分任务可采用两种方法使用 OCR 模型进行文本识别和使用 python 库进行 pdf 文本提取。

二者各有优劣，python 的 pdf 文本提取库可以简单实现文本提取，但有时 pdf 内文字以图片形式存在，此方式无法提取；OCR 可以识别出文本信息但可能存在识别错误和格式错误，而且如百度 OCR 库识别有图片的 pdf 同样会报错。

暂考虑以简单的 pdf 文本提前库实现。

任务四：情感分析模块和语义分析模块

任务内容：使用 jieba+ Snownlp 进行文本分词、语义分析、情感评价、关键词提取。Snownlp 可以支持这些功能的简单实现。可参考：https://cloud.tencent.com/developer/article/1699688

任务五：简单的展示可视化

任务内容：简单的舆情数据存储展示、政策文件文本的存储展示和简单搜索、一些舆情分析结果的可视化，可先实现简单的前后端。

任务分工：

王申宇：flask 后台搭建支持接入各项功能。

胡炅炫：负责任务一收集相关可用网站，以及任务三的下载一月二月相关政策、标准文件（直接在网页内公布的要给出网址），舆情信息爬取及数据处理。

徐翰文：负责实现 pdf 文本提取模块、情感分析、关键词提取、语义分析模块。

陈玉仪：前端可视化界面，包含舆情信息的展示、搜索，政策标准的展示搜索，舆情情感热

度等可视化展示（图表）。