以下作业均在收到作业的【24小时】 交付

B站爬虫：

以下仅以某内容为例，最后输出的代码要求适用于搜索任何关键词（输入变量也可为URL），推荐代码部分以Def、方程或封装等形输出

例如，B站搜索 Fate，出现以下界面（内容仅为截图，包括且不少于）
URL：https://search.bilibili.com/all?keyword=fate

题目如下：

一、输出基本搜寻结果

综合 视频 99+ 番剧 18 影视 9 直播 11 专栏 99+ 话题 5 用户 99+

- 相关内容主要分为哪几大类（例如：视频、番剧、影视、直播等），每类中有多少作品
    - 例如，Fate搜索结果出现的番剧，有19个；视频作品有至少 4*5*50（页）个
- 每日新增视屏相关信息
    - 视频数
    - 视屏时长
    - 视频相关作者基本信息（该作者是否为过去一周相关主题Fate主题下活跃作者:是否过去一周上传相关视频不限于一个）
    - 视屏播放量
    - 视屏弹幕数量
    - 视频评论数

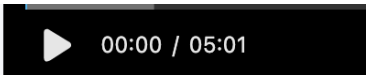
注:如果量过大，可以酌情缩短检索时间窗口

二、针对单个视频（跟上述一中的部分内容可重合处理）

以该链接为例 (需要将大量弹幕和评论等极端【重量级】案例的应对措施考虑进爬虫需求，不仅针对少数量级案例）

https://www.bilibili.com/video/BV1us411h7BQ?from=search&seid=9220732790560164695&spm_id_from=333.337.0.0

输出以下结果：（最后输出结果以xlsx或csv为主，灵活运用行、列、dataframe效果展示，便于逻辑性运算和二次加工）

- 播放时长 

- 点赞数、投币数、收藏数、转发数

- 标签

- 视频标题、视频介绍
- 爬取弹幕（包含详细弹幕内容、时间戳timestamp、弹幕发出的用户账号信息）
    - 弹幕会有品类之分，请考虑重点突出的弹幕如何归类
- 评论内容（内容、发布时间、回复内容）
    - 有多少LV4+以上的账号回复（回复中的账号等级）
- 输出视频品质信息（比如解析度等）

重点问题：

- 如果需要real time的增量内容的爬取，如何处理，请输出可落实方案及代码；
- 增量处理会出现需要数据存储（本地数据库）等情况，请输出应对方案；
- 如果每天弹幕有爬取数量上限MAX，如何做增量处理该上限问题，超过MAX的部分如何提取，并保证不会爬取出重复内容，请输出应对方案及代码及Plan B；
- 保证历史数据的爬取；

三、针对UP主（依然可和 一、二中需求结合）

URL举例：怕上火暴王老菊
https://space.bilibili.com/423895?from=search&seid=5117393271046525453&spm_id_from=333.337.0.0

输出结果：（最后输出结果以xlsx或csv为主，灵活运用行、列、dataframe效果展示，便于逻辑性运算和二次加工）

- 基本信息:UP主注册信息、动态、投稿、视频数量、关注数、粉丝数是否有其他链接等

- 所有视频的发布时间戳、年限、月份、天数
    - 输出案例，以excel列为主，仅供参考
      #个视频发布于2001 |#个视频发布于2002 |...| #个视频发布于2021 | 总计视频数 | 总计评论数量 | …)
- 所有视频的基本信息（包含且不限于 时长、评论数、弹幕数量）
- 关注人数中 不同账号等级各有多少人（是否存在头部KOL 或者 高LV账号比例）

四、代理IP

背景:考虑到爬取数据和文本会遇到IP切换等情况，避免过度使用某一IP造成局域瘫痪

- 代理IP的处理方式，请提供可落地方案及代码

五、NLP自然语言处理情感语义分析

选择不少于10W+以上的弹幕 或 评论区 爬取的内容，输出NLP语义分析模型结果

- NLP模型可针对日文、英文、中文等做出相应的归类
- 正向、中立、负向关键词中的TOP N表现
- 模糊语境的中立关键词进行更精准的语义分析

输出要求：

- 提供源代码，以Python、Java为主，数据以Excel、CSV、dataframe等形式交付
- 验证可操作性：可采取远程桌面测试、提供代码测试，并能够逐行解释代码的原理及运算目的
- 数据文本爬取结果可以CSV, EXCEL等形式输出，格式统一，易于使用
- 通过自选的案例（要求符合以上标准）输出1-2页的NLP语义结果分析归纳，以PPT或者Word为主
- 请勿直接使用网络上源代码