KDD Lab 作业 2

4/14 14:00 ~ 4/28 22:00 通过 blackboard 上传作业文件，作业不接受任何理由的补交或缓交，请同学们注意尽早完成并提交。截止时间为第十一周周四晚上十点。Blackboard 上有 3 次提交机会，希望同学们提交前能好好检查完再上传，如果真的有需求重新上传，请联系助教赵奕丞处理。

逾期不收！逾期不收！逾期不收！

第二次作业的任务是要同学们完成一套完整的数据收集→数据清洗→模型预测→结果展示的过程。本次作业的任务是：通过爬虫手段获取腾讯新闻：疫情实时追踪里的数据（提示：在浏览器中查看网络记录可以找到对应的 API），尤其是其中的“全国现有确诊”“全国新增确诊”“全国新增疑似”“全国累计确诊”“全国累计治愈”“全国累计死亡”等数据。在爬虫获得原始文件后提取出其中的关键信息，并对该关键信息进行一个基本的线性回归未来预测：腾讯 API 中有提供过去两个月的疫情相关数据，同学们需要针对其中的最近 30 天（新冠病毒在国内传播时间段）的 6 种统计数据（全国“现有确诊”，“新增确诊”，“新增疑似”，“累计确诊”，“累计治愈”，“累计死亡”等）分别进行线性回归拟合，并预测出下一天的对应数值。

****本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测****

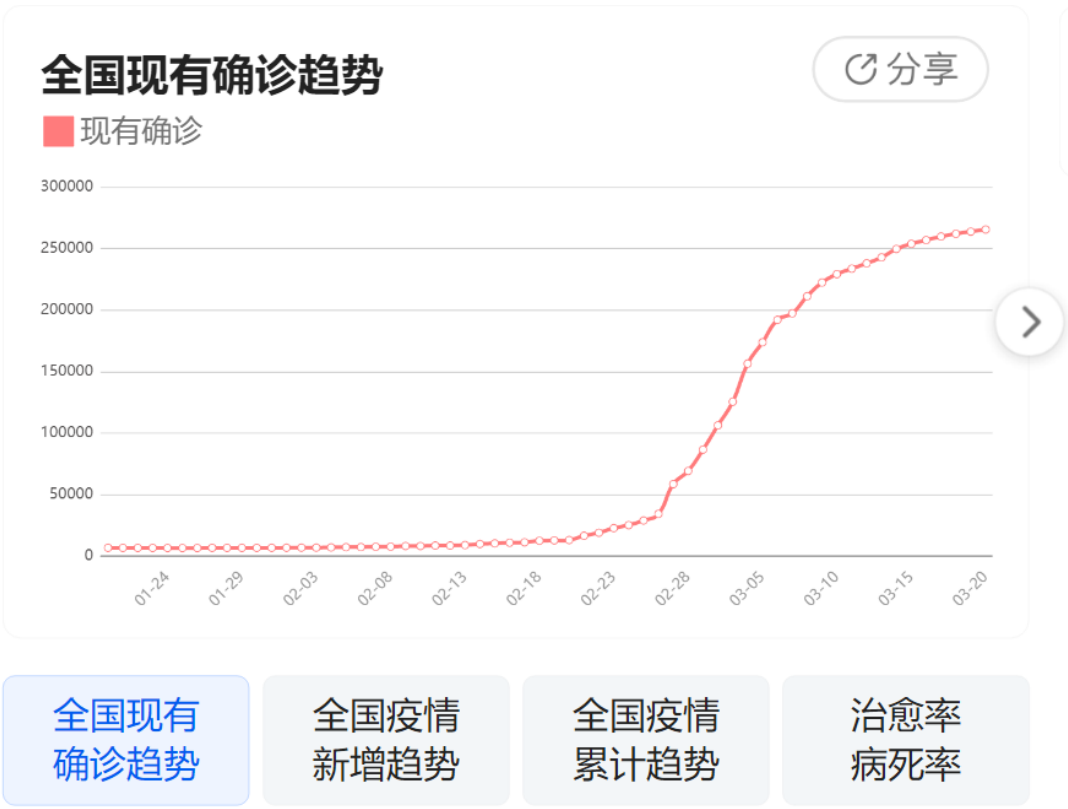
腾讯新闻：疫情实时追踪网址：

https://news.qq.com/zt2020/page/feiyan.htm#/

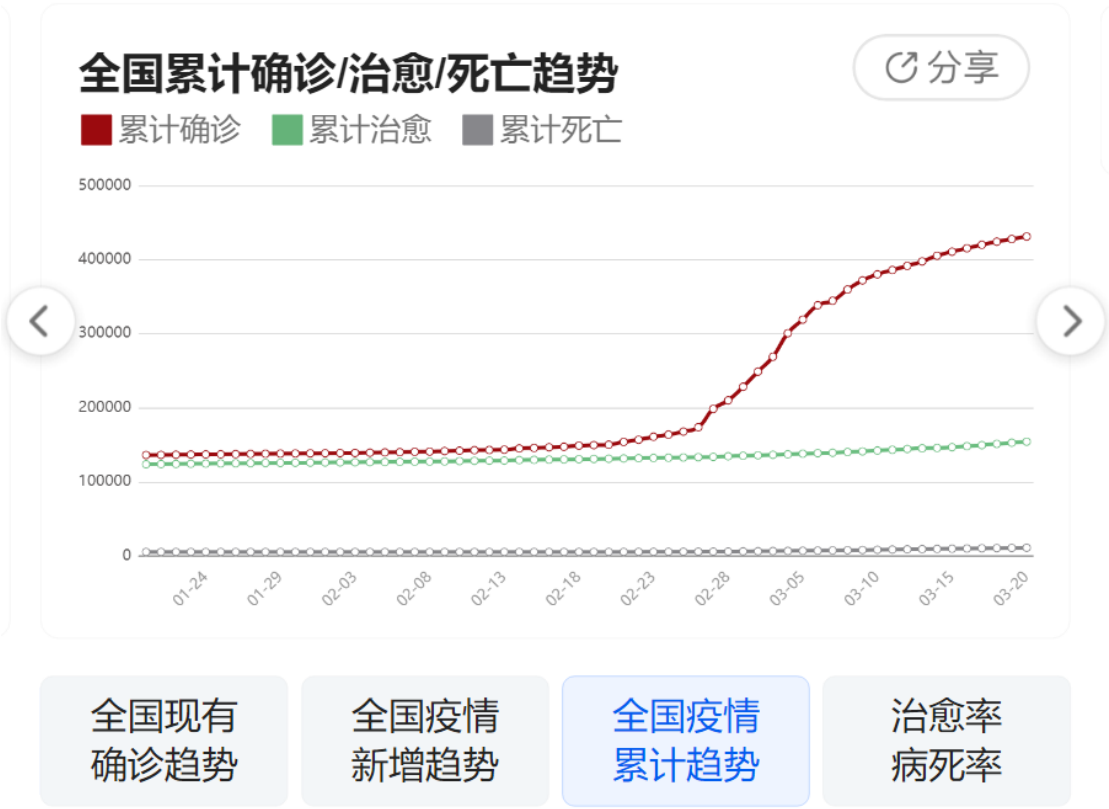
同学们需要找到可以爬取/申请访问的 API 接口的地址（提示：数据并不在腾讯新闻的页面上；通过观察腾讯新闻页面的网络读取部分可以找出对应的数据接口 API，数据接口 API 会长得像 https://api.inews.qq.com/.....类似的页面中，是 json 结构的数据）

所需数据如图所示：

****本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测****



本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测



在成功提取过去两个月的全部数据后，使用其中的最近 30 天的数据，分别针对全国“现有确诊”，“新增确诊”，“新增疑似”，“累计确诊”，“累计治愈”，“累计死亡”等 6 种数据建立 6 个简单线性回归模型，用于疫情数据预测。每个模型的输入是一个日期，输出是预测值。在建模过程中，先用 24 天的历史数据作为训练集拟合模型（train 的过程）；再用 6 天的历史数据来评估模型（validate 的过程），评估方式可以使用 RMSE；最后，输入第 31 天的日期到模型中得到对应的预测值（test 的过程）。

（举例说明：假设需要预测 2022 年 4 月 14 号的“现有确诊”值，

****本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测****

那么需要用到 3 月 15 号到 4 月 13 号（最近 30 天）的历史数据，用其中 24 天数据进行模型拟合，6 天的数据进行模型评估，最后预测 4 月 14 号（第 31 天）的“现有确诊”值。）

请将 6 个模型的**拟合得到的方程表达式、模型评估结果、第 31 天的预测值**等信息分别进行打印。

作业要求：如上述介绍，本次作业需要提交一套完整的数据收集，清洗，模型训练与未来预测流程。最后需要打印出对应的内容与结果。请同学们只使用 python 原生的库和 lab 课上使用过的库来完成此次作业的 baseline。请明确的在程序代码中备注原因和过程，最后按照要求命名你的 ipynb 文件上传作业

命名要求：请按照“HW2_你的学号”的方式命名你的文件，假如学号为 11956789，那么作业文件名为 HW2_11956789.ipynb

备注：

（1）为方便作业批改，必须提交规定格式的 ipynb 文件，且提交前先清除所有输出，减少 ipynb 文件的大小

（2）数据收集，数据清洗，模型训练，和未来预测都需要在程序中保留并提交，本次作业只收取 ipynb 文件，请同学们不要读入其他数据文件，数据需要通过网络爬虫实时获取，实时清洗训练以及预

****本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测****

测

（3）基本的线性回归模型仅仅作为最基础的 baseline 用于疫情数据的拟合和预测，本作业不要求使用其他高阶模型进行疫情数据预测。因此，**除非是额外完成了非 baseline 的预测，否则请同学们不要超纲，使用 lab 课上还没提及的第三方库，以免作业批改时产生不必要的报错**

作业计分标准（满分 10 分，扣完为止）

内容得分点：

数据获取：2 分

数据清洗与整理：2 分

每种基于线性回归模型预测的实现（共 6 种）：1（共 6）分

其他扣分点：

迟交：-10 分

作业文件不按照命名规则提交：-1 分

程序无法执行/程序出现无法继续运行的报错：-3 分

并没有提交 ipynb 文件，而是提交了 py 文件：第一次-5 分，第二次-8 分，第三次-10 分

****本次作业数据涉及新冠病毒疫情数据，但本作业仅用于让同学们练习使用线性回归模型，真实案例下要对疫情未来趋势进行预测需要很多其他的额外知识与数据，本作业的结果不能被当做是对疫情走势的权威预测****