简要设计文档

本项目使用了 scrapy 作框架，实现了棉花批次数据的获取。在核心程序外层，提供日期范围选择的参数选项(如果不选择则默认为是今天)。把日期范围按天拆分，然后进行 for 循环，取当前循环日期的所有仓库的所有批次，再获取具体批次的 json 数据，以批次号为文件名存入文件。

下面简要介绍获取数据的核心过程，大概是三步。

首先，以 http://www.cottoneasy.com/storage/maStorageProgressListInit 作为入口 URL，分析它的 html 页面结构，发现每个仓库某天的详细信息的请求地址是有规律的，请求 URL 由一个仓库编号和一个日期组成的，比如 http://www.cottoneasy.com/storage/maStoragePlanDetailInit?depotCode=19914&apptime=2017-01-16 这个网页是编号为 19914 的仓库（新疆伊犁州陆德棉麻有限责任公司）在 2017 年 1 月 16 日的详细信息。因此，第一步的任务是通过 xpath 解析页面，得到所有仓库的编号，然后就可以加上相应日期，拼接成我们所需要的 URL。这一部分对应了 cspider.py 里的 parse()函数。

其次，由上一步拼接到的 URL，进入到某个仓库某日的详细页面内，每个详细页面内都显示有若干批次号，如图所示：

批号	加工企业	品名	加工方式
65566161117	沙湾县元康棉业有限公司	细绒棉	锯齿
65069161167	乌苏市哈图布呼农牧发展有限责任公司	细绒棉	锯齿
65678161069	奎屯恒锦棉业有限公司	细绒棉	锯齿

同样通过 xpath 对 HTML 页面进行解析，获取所有批次号。这样就可以将 http://www.ccqsc.gov.cn/query/compareBatchInfoData.action?batchCodeInput=和批次号进行拼接，获得某个批次 json 信息的 URL。这一部分对应了 cspider.py 里的 parse2()函数。

最后，也是最容易的一部，依次访问上一步拼接的 URL，即可获得返回的对应批号的 json 数据。这一部分对应了 cspider.py 里的 parse3()函数。