1. 疑问：
2. 希望明确一下展示全部数据需要的硬件条件具体是怎样的。（以CA数据为例）

每种实例的数量在不同的数据中出现情况肯定不同，因此上限数量也无法固定。

以CA数据为例，从机器配置的角度，应该需要空余300MB内存就可以打开CA。

建议预留1g内存左右，保证最低体验。

查看剩余内存可以用任务管理器

1. EXCEL统计表中，每类问题的样例可以全部列出吗？

可以列出全部错误类型，可以使用参数配置每种错误的例子

无论是何种方式，只要不是建立完整的大数据平台系统，就必然用本地文件的方式进行存储，因此就必然受到本地机器配置的影响，解决方法是合理划分文件即可。强求单个大文件的完整性是不必要的，如果对单个大文件的完整性有强烈需求，则必然考虑提升机器配置。

1. 个人机环境下，既有的可视化工具能支持多大量级的分析？联动实例的上限数量是多少？

从机器配置的角度，预留内存不到1g的机器还是建议联动10以内就可以了；内存比较充裕、不紧张的话，可以尝试更多。

1. 目前的参考文献识别，是否支持将整个参考文献以“连续的块”识别出来，只有以连续的块识别出来才能方便打标签？

目前使用的正则表达式可以由开发者嵌入任何系统，已经可以简单识别出参考文献内容，可以不纠结“连续“、”块“的概念。打标签是后续功能，按需进行评估和对接即可。

1. 如果是多进程并行，每种问题的统计分析应该是只针对输入文件中的数据，而非全量数据，无法满足最初的需求，该问题如何解决？

此话题涉及的概念很多很杂，细细的梳理一下。

多进程。属于加速方法。也是为了达到目标做的优化。

输入文件。无论是什么软件，都是只对软件的输入进行处理，一定不会对不输入的文件进行处理。

最初的需求和最初的解决方案。都是双方的理想化后的最高的期待。这个最高期待往往是需要补充后续的新的信息的。尤其在实践工作中，一定需要在掌握了新的信息之后进行分析利用，才能做到符合常理。

跑全量数据，要进行合理的调研分析计划。首先，要清楚这是下一个阶段，不仅仅是一个学术层面、或者说算法架构层面、或者说工具层面的阶段了。

是需要按步骤进行分析计划的，至少包括如下问题：到底为什么而跑、跑了之后仍然是大数据规模的结果又要如何利用、到底有多少数据、数据分别是什么样子、有哪些特殊数据、需要多长时间、需要多少台服务器。

接下来说怎么跑。

目前来看，全量数据规模在千万以上，参照2020年专利局其他千万以上数据规模的项目的处理经验、时间、预算、服务器配置。

以Ca数据3400篇进行推算：

数据量：每个问题至少得到了3000个以上的实例，因此按1千万篇来算，会得到1千万个以上的实例。

时间：CA使用常见PC机，要跑206s，8个进程，因此按1千万来算，普通PC机的话时间是不可接受的（400天）。因此全量数据一定需要使用服务器级别的机器。建议一般的服务器都有64核，如果有10+台服务器，还是可以完成的。

1. 补充需求：
2. 建议增加配置选项，联动实例的个数可以根据各自的环境配置情况进行选择。10个、20个、30个、50个或者100个

已添加，通过设置config.ini中的MAX\_ERROR\_TIMES\_PERTAG\_PERTYPE的大小，可对实例个数进行配置。

# 在可视化中输出的每个类型的错误数

MAX\_ERROR\_TIMES\_PERTAG\_PERTYPE = 10

# 在csv中输出的每个类型的错误数

MAX\_ERROR\_TIMES\_PERTAG\_PERTYPE\_CSV = 100

# 标签内容的最大长度，10240就是说每个标签下最多有一万字

MAX\_LENGTH\_PERTAG = 10240

1. 目前的可视化分析工具支持json格式的文件，建议增加对excel、txt等格式文件的支持。

可视化需求是在检测工具开发之后补充的，是为了对检测工具的中间结果进行可视化。中间文件JSON是复杂的数据结构，表格和纯文本不容易承载这样的功能。

另外，如果需要，可以交付可视化分析工具的源码，方便以后其他相关项目的二次开发。

1. 可视化分析结果除了饼图，建议增加柱状图、折线图等呈现方式。

可以利用现有饼图的数据，再画出柱状图、折线图。