**阶段一、数据爬取**

该阶段主要是使用分布式爬虫获取所需的文本集，难点是如何实现PathFinder算法。现在很多网站都采取了反爬措施，如何绕过关卡爬取数据也必须重点关注。

阶段一的准备时间可能较长，但只要考虑周到，具体实施时应该不会有很大困难。主要由周余承担。

**阶段二、数据处理**

为了便于后续的分词、知识抽取等一系列操作，要对阶段一获取的数据进行清洗、过滤等。由于之前没有接触过相关内容，所以这一阶段要对数据进行哪些处理，才能使得生成的数据更有利于下一步操作，不好把握。

预计一周之内完成本阶段工作，主要由文华承担。

**阶段三、 特征提取与文本分类**

采用VSM描述文本，先利用word2vec将本文中的专业术语提取出来，表示成向量；特征提取采用χ2统计，在Spark上并行处理。将提取到的特征作为参数，利用有监督学习对文本进行学习，完成分类的范围划定。初步分类使用k-means算法，在Spark上并行处理。文本分类使用Bayes，在Spark上并行处理。

本阶段应当是本项目疑难点最多的地方，很多细节都需要推敲，可能有些原定的步骤用不到或者该用的没想到。

预计花两个星期完成阶段三的工作，由刘宏鑫、周余和文华一同承担。

**阶段四、模型优化与数学解释**

对已经分类的文本和训练的模型进一步优化，争取在数学上解释算法和模型。

预计花五天完成本阶段工作，主要由周余承担。

**阶段五、术语关键词联想**

基于Bayes完成知识图谱关键词联想，预计花五天完成本阶段工作，由刘宏鑫和文华一同承担。

**阶段六、可视化UI设计**

对软件UI进行设计，预计花三天完成本阶段工作，主要由刘宏鑫承担。

**阶段七、构建可视化映射函数**

设计一整套完整的知识图谱的可视化函数，并提供开发文档，预计花一周完成本阶段工作，主要由周余和文华承担。

**阶段八、构建D3.JS平台与知识图谱可视化**

将抽取的知识可视化，并提供开发文档，预计花一周完成，主要由刘宏鑫、周余和文华完成。

表 1 项目日程计划

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 阶段 | 预计耗时/天 | 前序工作 | 承担人 |
| 1 | 数据爬取 | 10 | 无 | 周 |
| 2 | 数据处理 | 7 | 1 | 文 |
| 3 | 特征提取与文本分类 | 14 | 2 | 周、文 |
| 4 | 模型优化与数学解释 | 5 | 3 | 周、文 |
| 5 | 术语关键词联想 | 5 | 4 | 周、刘 |
| 6 | 可视化UI设计 | 3 | 无 | 刘 |
| 7 | 构建可视化映射函数 | 7 | 5 | 周、文 |
| 8 | 构建D3.JS平台与 知识图谱可视化 | 7 | 6 | 周、文 |
|  | 合计 | 51 |  |  |

麻烦老师給我们看看，怎么计划比较稳妥、更有可行性。我觉得队友都还是那种肯干事的人。