# 规划

## 专家意见整理

### 主要问题

* 没有考虑对GitHub已有的Topic生成的研究。
* 没有清晰地证明研究能够在实际上对程序员起作用。
* 为什么要考虑多源数据，从GitHub还有Stackoverflow社区中获取的数据为什么能够弥补不足呢？

### 专家1

* 忽略了现有对GitHub Topics生成的研究：
  + [文章1](https://dl.acm.org/doi/pdf/10.1145/3084226.3084287)
  + [文章2](https://dl.acm.org/doi/pdf/10.1145/3382494.3410690)
  + [文章3](https://arxiv.org/abs/1910.07115)
  + [文章4](https://ieeexplore.ieee.org/document/8816780)
* 到底从GitHub还有Stackoverflow中提取了哪些数据，和DockerHub中提取的短文本描述和长文本描述有什么区别。
* 技术描写的细节过多，例如为什么选用Vertical Tag Extension，为什么要用那些技术，应该举更多的例子。
* 实验部分：
  + 数据是如何具体划分的？
  + 为什么使用具体的阈值呢？
  + recall并不重要，与其提供错误的标签还不如不提供标签。

### 专家2

* 质疑意义：提供标签对于开发者并不重要，用户并不会通过搜索这种广义的方式来寻找镜像，往往是通过某个具体的需求来搜索的。
* 方法没有什么创新性。
* 实验部分几乎没有方法论的体现。

### 专家3

* 在Motivation部分更应该体现出对于开发者来说有用的标签是什么，从这个层次进行对比。
* 为什么还要添加更多的数据源，DockerHub本身的不够吗？
* 采用不同的NLP方法的原因不明确。
* 实验不明确。

## 修改规划

1. 删除GitHub和Stackoverflow的部分，感觉增加了反而造成了不好的影响。
2. 修改方法和实验的部分，着重体现方法论的体现，内容不好改。

## 期刊投向

[Information Retrieval Journal](https://www.springer.com/journal/10791/)