网上系统项目申报

* **项目研究的目的和主要内容**

经过两年多计算机专业的学习与实践，我们拥有了一定的思维能力，掌握了一定的专业技能。为了学有所用并进一步训练和提升自己的创新能力，我们希望通过创新项目来锻炼自己。项目的主要内容包括以下：

① 代码数据的爬取、代码数据库的搭建、代码分析模型的搭建。首先我们需要获取尽可能多的代码建立数据库，并将收集到数据作为训练集，训练好分析模型。代码数据可能不只包括代码本身，还会附带很多附带信息，如git中提供的日志信息等。当代码仓库数据输入时，能够显示其中的基本信息，并给出相应分析模型得到的结果。

② 实证分析角度的选取。基于Git开源代码仓库日渐流行。对于贡献者人数众多的开源项目。贡献者的贡献数据对于软件的健康演化十分重要。为此，我们可以提出很多问题，比如贡献者的贡献信息在时间上的分布有无关联？贡献者的贡献度是否与语言种类等项目属性有关联？等等。这些问题的答案对于软件的开发与演化都可能为软件开发提供指导性的建议，或者可以较好地展示当前软件开发的趋势。

③ 对分析结果进行可视化。可视化的特性就是直观、使观者快速、准确获取信息。可视化实现的方法是利用Wicket框架并结合JavaScript可视化方法实现效果良好的用户界面。

④ web插件开发。可以将上述功能封装到web应用中，作为代码仓库平台的插件，更好地服务使用者。

* **国内外研究现状和发展动态**
  + 开源代码仓库日渐盛行，因此对代码仓库中数据的分析也愈发重要。反观当下流行的几种代码仓库，大部分都只提供了一些基本信息的可视化，而缺少一种有效的模型对仓库数据进行刻画，软件开发人员和软件管理人员缺少有指导意义的指标来了解软件的演化，因此这是一个比较新的领域。
  + 实证软件工程作为软件工程的一部分，近年来发展迅速。前些日(10月11、12日)在芬兰召开了第十二届ESEM研讨会，讨论了诸多关于实证软件工程的话题，比如类比其他领域的经验实证工作以启发实证软件工程的发展，以及机器学习技术在实证软件工程中的应用等。
  + 近年来，数据挖掘与分析技术十分流行，受众越来越广，普及度越来越高。使用Python、JavaScript等语言实现数据可视化的技术现已较为成。同时，人工智能和机器学习技术可以让数据分析与可视化更智慧地服务于人们。
* **项目特色与创新之处**

项目的目标是完成一篇时政研究以及数据分析插件的开发。将应用Wicket开发框架等完成前后端开发，使用机器学习技术对数据进行建模分析，对分析结果进行可视化并提高其用户友好度。

实证分析角度的选取是本项目的核心。好的角度选取对软件开发过程将有着重要的指导意义。因此本项目的创新之处在于实证分析角度的选取以及相应的数据建模方法。

* **技术路线、拟解决的问题及预期结果**

技术路线：

① 爬取数据，对得到的代码仓库数据分析其基本信息，并可视化。爬取数据的来源可以选择Github，有海量且质量很高的开源项目；分析其基本信息即使用现在常见的统计模型进行统计，比如贡献次数、活跃时间区间等。这部分我们将通过Java语言，结合Wicket框架，使用JavaScript可视化方法实现。本阶段的目的在于掌握基本的数据爬取与可视化的基本方法，为后期进一步工作打下基础。

② 通过数据挖掘、机器学习的方法，从选取的研究角度分析开发者行为特征进行研究分析。项目的主旨在于分析开发者的行为特征，为了实现这一点，需要广泛调查了解当下开源软件工程中开发人员的工作状态和贡献行为，选取可行性高、有研究价值的方向，建立相应的分析模型，并将现有代码仓库作为训练集，得到相应的规律，并总结出结论，给出有指导性的建议。这部分机器学习可以使用Python语言结合TensorFlow等机器学习库，并使用R语言作为统计工具实现。

预期成果：

形成一篇实证研究文章，针对某一特定的角度对开源仓库中贡献者的行为进行分析实证，以获取代码贡献者的行为特性及其贡献度与软件演化的关系，为软件开发提出指导性建议；形成可以发布的web应用，用户可以从多个角度了解代码仓库中各贡献者的基本信息以及贡献行为特性的分析结果，丰富代码信息统计的内容，为代码仓库管理者管理仓库、规划软件开发提供方便。