Improving The Usability of Software Systems Using Group Discussions: A Case Study on Galaxy

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Shamse Tasnim Cynthia

©Shamse Tasnim Cynthia, December 2022. All rights reserved.
Unless otherwise noted, copyright of the material in this thesis belongs to the author.

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Usability problems in software systems cause performance degradation, user dissatisfaction and loss in terms of cost. There is a growing need for the software systems to become more accessible, retrievable and usable for the users. The usability test of a software is conducted by getting the opinions directly from the users and its goal is to identify problems, uncover opportunities and learn about target users' preferences. But accessing real users is very difficult for certain software systems. However, there are many popular user forums such as Stack Overflow, Quora, Stack Exchange, Eclipse Community Forum etc. and people from different domains of knowledge use these forums to ask about their problems and post their concerns. So exploring these forums should provide significant knowledge for getting information about a system's usability issues. Previous studies show that investigating these group discussion forums discovered several usability issues that the system was unaware of such as topic categorization, automatic tag prediction, identifying reproducible codes etc. However, there are many Scientific Workflow Management Systems (SWfMSs) such as Galaxy, Taverna, Kepler, iPlant, VizSciFlow etc. and although these SWfMSs are emerging and important for data extensive research, no study has been done earlier to figure out the usability problems of these systems. Therefore, in this thesis, we take Galaxy, a well-known SWfMS, as our use case. We explore the user forum that Galaxy offers where users ask for help from experts and other Galaxy users. We search for the issues users are discussing in the forum and find out several usability problems in different categories. In our first study, we try to group the usability problems to easily identify them and galaxy community can be informed of the existing usability problems of the system. While exploring the posts, we find a significant percentage (up to 28%) of them lack tags. If tags are found, they do not reflect the context of the posts properly. This leads to one of the major usability problems for the discussion forums as users will be unable to identify suitable posts without proper tags. Moreover, users will face difficulties to explore the answers in those untagged questions. So in our second study, we try to suggest tags based on the context and proposed a method for automatically suggesting tags. Again in our extensive investigation, we find lots of usability issues but among them, the problem of finding and searching for the appropriate workflows emerges as a great usability problem of the system. Users, especially novice users, ask for workflow design recommendations from the experts but because of the domain-specific nature of SWfMSs, it gets difficult for them to design or implement a workflow according to their new requirements. Any software system's usability is called into question if users face trouble specifying or carrying out certain tasks and are not given the necessary resources. Therefore, to increase the usability of Galaxy, in our third study, we introduce a NLP-based workflow recommendation system where anyone can write their queries using natural language. Our system can recommend the users with the most relevant workflows in return. We develop a tool on the Galaxy platform based on the idea of the proposed method. Lastly, we believe our study findings can guide the Galaxy community to improve and extend the services according to the users' requirements. We are confident that our proposed methods can be applied to any software system to improve the usability of the system by exploring the user forums.

Acknowledgements

At first, I would like to praise the Almighty, who blessed me with the ability to carry out this work. Next, I would like to express my sincerest appreciation to my supervisor Dr. Banani Roy for her continuous support, guidance, motivation and remarkable endurance during this thesis work. Without her support, this work would have been unthinkable.

I would like to thank Dr. Julita Vassileva and Dr. Debajyoti Mondal for their willingness to take part in the evaluation and advisement of my thesis. In addition, I am grateful to them for their valuable feedback and suggestions.

I would like to thank anonymous reviewers of different conferences for their valuable comments and feedback which helped me to improve this thesis work.

I express my heartiest gratitude to my father, mother, brother, sister-in-law and my twin nieces. Their endless sacrifice, unconditional love and constant good wishes have made me reach this stage of my life. I am thankful to Shams Zerin Xian and Fariar Rahman for being there with me in my every happiness and sadness. They always stayed with me in ease and hardship and inspired me constantly. I would like to thank Al Muttakin for his constant support and ability to motivate me whenever I needed. I would also like to thank Nowshin Tabassum, Fahmida Promi, Tafannum Torsha and Shoshe Chowdhury for their consistent support in my postgrad life.

I would like to thank the Interactive Software Engineering Lab (iSE Lab) and Software Research Lab (SR Lab) members for the good time we have together. Specially, I would miss the lively discussions during coffee breaks and lunch time. In particular, I would like to thank Dr. Chanchal Roy, Dr. Farouq Al-omari, Mohammad Mainul Hossain, Saikat Mondal, Amit Kumar Mondal, Md. Abdul Awal, Shamimur Rahman, Debashish Chakroborti, Sristy Sumana Nath, Md Nadim, Khairul Islam, Subroto Nag Pinku, Saumo Roy, Palash Ranjan Roy, Jarin Tasnim, Parnian, Justin Schneider and Rayhan Islam Shuvo.

Special thanks goes to C M Khaled Saifullah, Afsana Sultana Ruma, Avijit Bhattacharjee, Naz Zarreen Oishie, Shahrima Jannat Oishwee, Mobassir Hossain Irteza, Shubhashish Karmakar and Rabiul Awal for their love and mental support during my stay in Saskatoon. I am very much thankful to C M Khaled Saifullah, Afsana Sultana Ruma and Avijit Bhattacharjee for being the guardian I needed in Saskatoon.

I would like to thank the Computer Science department of the University of Saskatchewan for their financial assistance through scholarships, awards, bursaries which helped me to focus on this thesis work. Moreover, I would like to thank all the staff of the Department for their constant support. In particular, I would like to thank Sophie Findlay, Heather Webb, Maurine Powell, Shakiba Jalal, Greg Oster, Jeff Long and Tamarra Calver.

Lastly, I would like to express my gratitude to Saikat Mondal for the help and support he provided me in my thesis work. Along with him, I am also thankful to Subroto Nag Pinku and Saumo Roy for our hangouts together.

I dedicate this thesis to my father (Md Shahadat Hossain), my mother (Khairun Nahar), my brother (Imtiaz Hossain Nisat), my sister-in-law (Banani Hoque Raka) and my twin nieces (Inaya Hossain and Nayyara Hossain), who always believed in me and inspired me to become the best version of myself.

Contents

List of Tables

List of Figures

List of Abbreviations

SWfMSs	Scientific Workflow Management Systems
NLP	Natural Language Processing
NLU	Natural Language Understanding
NASA-TLX	NASA Task Load Index
UI	User Interface

1 Introduction

In this chapter, we describe the motivation and the problem statement of the thesis. In Section 1.1, we present the motivation of the thesis. We then state the problem statement and our contributions in Section 1.2 and Section 1.3. We show the list of the prepared and submitted papers from this thesis in Section 1.4. Finally, in Section 1.5, we provide an outline of the remaining chapters of the thesis.

1.1 Motivation

Usability refers to the definition of measuring a system's effectiveness, efficiency and satisfaction [9]. The absence of those criteria is considered as a usability problem. The usability evaluations are applied to evaluate the quality of a user interaction design and establish a basis for improving it. Usability also implies that a software system can deliver its content to the users in an efficient and ordered way. When an interactive software system is built, ensuring the usability of the software is one of the key factors that developers should keep in mind [42]. But software systems suffering from usability problems are well known in the research field [28, 36, 83]. The presence of poor usability of health information technology products has already shown profound negative impacts. Research showed that 22 types of medication error risks were generated only by the usability issues with computerized provided single cases [56]. It has also been found that for software development, agile practitioners hardly emphasize developing software that is usable for the users [12, 21, 31]. This leads to raising dissatisfaction, frustration and the tendency to lose interest in using that particular software system among the users. For any software systems developments, it should be kept in mind that user experience is threatened without ensuring the proper usability of the system. When a system's usability is tested, then the system developers can better understand the user's needs and can build plans or strategies to meet the users' requirements. To our knowledge, usability testing includes several steps to accomplish its goal. From planning the session to recruiting participants and getting their opinions all of them play significant roles in getting the job done. But what happens when the users are not available? Or the users are not very accessible from the outside? All these questions led us to find the options on how we can get users' opinions, their concerns or their questions about the system. We investigated and explored the user forums or group discussion platforms such as Stack Overflow [1], Stack Exchange [2], Eclipse Community

[1] https://stackoverflow.com/

[2] https://stackexchange.com/

Forum [3], Jupyter Discourse Forum [4] etc. All these user forums are either serving as a question-answer site for any kind of software system or for some dedicated software systems. And analyzing these forums led to various research studies conducted over the years and very significant findings have been discovered from the analysis, for example, topic categorization studies [65, 102], predicting tags or suggesting tags [95, 109], finding reproducibility issues [91] and so on. Therefore, we realized that user forums could be a good source to find out the usability problems of the software systems too. Thus, we move our research direction to find the usability issues in software systems by using the group discussion forums. However, among the various types of software systems, we want to find the usability problems in scientific workflow management systems (SWfMSs). As an emerging type of software system, SWfMSs have gained much popularity in helping scientists to conduct scientific computations with less effort and time. Many of the SWfMSs are used by scientists around the world and usability issues have been found on them over the years [96, 35, 49].

A scientific workflow management system aims to automate its process's life cycle, which is composition, deployment, execution and result analysis. The increased amount of data production accelerated the users' urge to have a system that can manage and execute scientific experiments without much hassle. Data analysis and visualization also played a significant role in developing any software system only dedicated to scientific experiments. Several SWfMSs have been developed over the years such as Galaxy [20], Taverna [100], Kepler [5], Pegasus [26], VizSciFlow [44] etc, but no usability issue investigation over these SWfMSs has been conducted ever. Now to investigate more on the usability issues of the SWfMSs, one needs to contact the users and test the usability issues of the system. But a few of the systems provide any platforms for the users to share their experiences, problems or concerns. So it is quite difficult to reach out to the users and ask their opinions about the usability issues of the SWfMSs. With the ever-growing development of new SWfMSs, the number of users is also increasing. Without considering the users' requirements and expectations from these systems, user satisfaction cannot be provided. Many of the SWfMSs have become obsolete just because they did not emphasize the users' needs or focus on the issues users are discussing when faced with problems. As the importance of analyzing and conducting high computing scientific experiments is increasing day by day, it is not an ideal solution that the existing SWfMSs would go obsolete, reducing the number of available SWfMSs for the users.

Hence, a usability study of the existing SWfMSs can open up a new research area on meeting users' requirements. Creating an environment where large computing analysis is performed while keeping users' satisfaction in mind should be an ideal platform for accelerating scientific experiments. Several studies present the motivations and possibilities that SWfMSs ensuring users' needs can improve the data analysis on a large scale [64, 112, 61]. But reaching out to the users of SWfMSs happens to be a difficult task. As no other previous studies have been performed including the users and getting their feedback, so exploring the users' opinions became a cumbersome process to perform. For conducting our research, we tried to reach out to

[3] https://www.eclipse.org/forums/

[4] https://discourse.jupyter.org/c/jupyterhub/10

the users of SWfMSs, but unfortunately, we were not successful in getting the reach of them. We tried to contact the developers of the SWfMSs and some other users who provided their contact information on those systems, but we have yet to hear back from them. Thus, to investigate the user's opinions and their point of view, we took the help of the user forums provided by SWfMSs. But the main challenge here was to find out a reliable and efficient user forum that is still available for the users. After some research, we find Galaxy [20] has the most convenient and helpful user forum among all other SWfMSs. There are some more user forums or group discussion forums available on the internet but most of them became non-usable [5] and users stopped to write on those forums long ago [6]. Galaxy is one of the leading SWfMSs, having more than 125K users who conduct thousands of scientific experiments on this platform on a regular basis. It is also a system where users' accessibility, reproducibility and flexibility of use are ensured. Galaxy also provides a platform [7] for the users to post their concerns and questions and can also ask for suggestions. Each day, a large number of questions are posted in this forum. In this thesis, we focus on the Galaxy help forum to find out the usability issues users are discussing and get insights into improving those issues so that other software systems can follow the same approach.

1.2 Problem Statement

1. **Users Posts Are Uncategorized and Non-traceable.** As mentioned above, we find that the number of posts in the Galaxy user forum is large in number. But there is no effective categorization of the posts and users need to go through the keywords searching option to find out any solutions. Moreover, the existing usability issues are discussed here. So exploring and categorizing the posts to find out the usability issues of this system can make it more usable and guide the users to find their answers more effectively.

2. **Posts Having No Tags or Nonrepresentable Tags.** While exploring the posts of the forum, we see that most of the questions do not contain any tags (up to 28%). Even if the tags are present there, they often fail to represent the context of the posts. Not only that, an abundance of unnecessary tags are identified on those posts too. This leads to disappointment and frustration for the users when they are not provided with posts containing appropriate tags. Users finding others' posts based on their requirements gets hampered because of the misuse or unnecessary use of the tags. Also, posts containing no tags will not be visible if someone is looking for an answer by searching for a specific tag. Thus, to ensure the system's usability, proper and representable tags should be included in the posts and users should be aware that without tags, their posts might not benefit others.

3. **Lack of Efficient Ways to Design Workflows Without Prior Domain Knowledge.** Moreover,

[5] https://tnrs.biendata.orgquestions/scope:all/sort:activity-desc/tags:CoGe/

[6] https://myexperiment.org/home

[7] https://help.galaxyproject.org/

we see a lot of posts [8],[9] from users asking for recommendations for implementing a workflow according to their criteria. We investigate the problem of why users are not finding the appropriate workflows or why they are not able to design as per their own needs. We find that being the SWfMSs very domain specific and for the shortness of proper description for the tools, and annotations for the shared workflows, the users often find it difficult to understand the working procedure of certain workflows. Thus, they cannot get an actual and workable solution from the repository to meet their requirements. Also the search option only works with keywords. So users need to look for the keywords one by one for designing the workflows and they also need to have some prior knowledge to find out which particular workflow that would serve their purposes. Hence, if the system could introduce a tool where users can write their queries or requirements in plain natural language and the tool could recommend the relative workflows, then the usability of the system would increase profoundly.

1.3 Our Contribution

We focus on the above-mentioned research problems in SWfMSs and in this thesis, we make three major contributions through our studies. In this section, we briefly present our contribution to the study.

1.3.1 Study 1: Understanding the users' posts and investigating the topics

In this study, we investigate the user forum and try to find what type of help users are asking for. We examine the questions and answers and investigate the forum. We attempt to understand the context of the posts by exploring the posts and the solutions provided to them. The outcome of our study is: we find that there are six types of issues users are asking questions on. According to our study, users ask for help on tools, Galaxy server or installation setup problems, job running issues, upload or input/output issues and other miscellaneous problems. Our study also found that among the problems, the usage of tools is the most frequently seen usability challenge. We thus further investigate the tools category to find more specific insights.

1.3.2 Study 2: Providing automatic tag suggestion technique

This study proposes a method for automatically suggesting tags to the users. We investigate the users' posts in a more detailed process and found the existing posts having no tags or non-representable tags. We thus propose a method which can suggest tags to the users according to the context of the post. We conduct a user study to evaluate the relevance level of the suggested tags and posted tags. We find that our suggested tags are more relevant than the posted tags. Moreover, for the posts where tags are not provided, our method can suggest appropriate tags too. We present 10 sample questions to the participants to rate the relevance

[8] https://help.galaxyproject.org/t/create-reference-genome-from-my-wgs-data-custom-reference/1051

[9] https://help.galaxyproject.org/t/metagenomics-analysis-on-pacbio-sequenced-data/8808

level of the suggested tags. We find that users found the suggested tags more relevant than the posted tags for nine of the questions.

1.3.3 Study 3: Providing a tool for the users to get workflow recommendations using natural language

In this study, we offer, for the first time for the users to have a tool that can convert queries written into the natural language to get some recommendations of the related workflows. In our investigation of the user forum, we find the problem users are facing in implementing their ideas into workflow designs. Often, the users need concrete suggestions for creating effective workflows or they need help finding the appropriate workflow from the repository. Moreover, the users need to have domain-specific knowledge to search for the workflows they need to get ideas. We discovered the problem in the search option where the search only works in the case of keywords. So, according to our investigation, we get the motivation to build a tool for the sole purpose of recommending related and workable workflows to the users. We build NL2WF tool on the Galaxy platform as in this system, and users shared their workflows for others to use. Our tool serves the purpose of converting natural language text to get workflow recommendation and our tool also provide ready-to-import workflow files to the users. Not only that, we evaluate our tool by conducting a user study. Our study results find that our developed tool significantly increases the Galaxy system's usability.

1.4 Publications

Below is the list of publications for accepted and to-be submitted papers:

- Shamse Tasnim Cynthia, Banani Roy, Debajyoti Mondal, Feature Transformation for Improved Software Bug Detection Models, 15th Innovations in Software Engineering Conference, February 2022, ACM Indexed, https://doi.org/10.1145/3511430.3511444. (Accepted)

- Shamse Tasnim Cynthia, Banani Roy, Improving The Usability of Software Systems Using Group Discussions: A Case Study on Galaxy, ACM Canadian Celebration of Women in Computing 2022. (Poster presentation)

- Shamse Tasnim Cynthia, Banani Roy, Improving The Usability of Software Systems Using Group Discussions: A Case Study on Galaxy, The 15th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. (to-be submitted)

1.5 Outline of the Thesis

The thesis contains six chapters in total. In order to investigate and improve the usability of software systems, we conduct three independent but interrelated studies and this section outlines the different chapters of the thesis.

- Chapter 2 discusses the background of the thesis such as scientific workflows, the life cycle of scientific workflow, scientific workflow management systems, Galaxy workflow management system, qualitative analysis and so on.

- Chapter 3 discusses the first study to investigate the user forum and categorize the issues users are discussing about.

- Chapter 4 discusses the importance of having tags in the questions, the investigation performed on the user forum to find out the relevant tags and the method we introduced to suggest proper tags.

- Chapter 5 discusses the tool we built to convert natural language queries to workflow recommendations and provide ready-to-import workflow files as well as the user study we conducted to evaluate our tool.

- Chapter 6 concludes the thesis with a list of directions for future work.

2 Background

In this chapter, we provide a short discussion of the background and technical preliminaries of the thesis. First, in Section 2.1, we discuss the scientific workflows, in Section 2.2, we give a general view of the SWfMSs and their working procedure. Next, in Section 2.3, we present a detailed overview of Galaxy. Then we present the types of qualitative analysis in Section 2.4. After that, we discuss the natural language processing methods in Section 2.5. Finally, we discuss the two measurement scales: Likert Scale and NASA-TLX in Section 6.8 and Section 2.7. A short description of the internal and external threats to validity has been discussed in Section 2.8. At last, Section 2.9 summarizes the chapter.

2.1 Scientific Workflows

Scientific workflow is the process of accomplishing a scientific goal or objective that is usually expressed in terms of tasks and their dependencies. Scientific workflows perform tasks or analyses for scientific simulations or data analysis. Workflows were mainly used in business process modeling as well as the database community. But the main difference between the workflows in the database community and in business process modeling captured the eyes very soon. It was understood at the earlier stage by the database community that traditional business data management is different from the characteristics of scientific data. The importance of workflow concepts was emphasized by WASA, a Workflow-based Architecture for Scientific Applications, where the term 'scientific workflow' has been introduced [66]. Scientific workflows have also been introduced by other roots, including problem solving environments. This came into light in the nineties as an intuitive tool for solving a target class of scientific computing-related problems [45]. However, with the advancement of e-science, scientific workflow research and development has been going through a major rejuvenation. Distributed and big computational techniques, tools from the computational sciences, databases, data analysis, visualization, sharing and collaboration of all sorts of works needs workflow management now.

Scientific workflows are mainly used for modeling and performing scientific experiments on a dataset [30]. The workflow steps are more commonly referred to as computational modules where the modules are engaged in performing independent tasks by manipulating and processing the data. The input and output formats of data is also handled by the modules. Computational modules are correlated with another forming of *Directed Acyclic Graph* of the modules presenting the dataflow relations among the modules [30] [8]. The entire DAG can be divided into multiple sub-graphs, which are then known as sub-workflows. Scientific workflows generally maintain data flow-oriented architecture for their execution. Thus, a workflow module

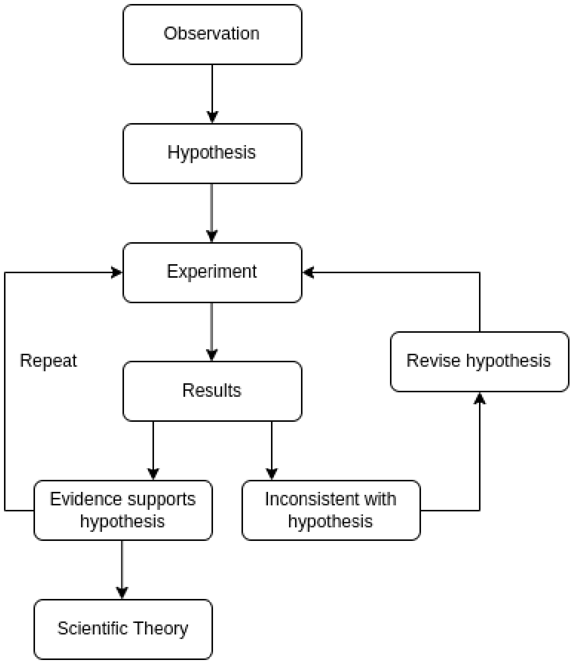


Figure 2.1: The basic steps of a scientific workflow that is dedicated to performing a series of analytical steps

tends to start its execution at the time of available corresponding input datasets.

There are four phases of a scientific workflow for completing its execution. They are:

1. *Design and Refinement:* The cycle starts with designing of a new or refining an existing workflow extracted from a repository. In this phase, the components needed for executing the workflow are selected and the establishment of these components is performed. The dependencies of data and components are precisely included in this step [43].

2. *Sharing and Planning:* In this phase, the designed workflow is expected to be shared with the e-science community infrastructure so that other researchers can access the workflow and share their ideas to run or extend them. In the planning phase, the designed workflow is expected to be turned into a concrete executable workflow. This is achieved by mapping abstract parts to concrete applications or algorithms. The parameters and sources of data are confirmed as well as the resources for execution are also selected. A thorough planning to particularly important for large-scale and compute intensive workflows as, in these cases, executions have to be mapped to high performance computing (HPC), Grid or Cloud resources [43].

3. *Workflow Execution:* A workflow engine manages the workflow execution. The work is done by mapping the execution to a proper environment or settings while extracting relevant information about suitable software, computing resources as well as data resources. The components in the workflow are executed in a structured manner and also get monitored by the workflow engine. The defined input data is

provided to the components and the results are sent back to the engine, and passed to the user [43] [25].

4. *Analysis and Learning:* An analysis and learning step successfully elaborates the last phase of a workflow execution. Here the workflow results are published and shared with other peers in the community. If the expected result has not been achieved, then the scientists execute the workflow again by tuning the parameters and other resources. In this way, the workflow is refined and improved. The iterative refinement process is very common in the daily life of a scientist which is also known as *trial and error.* Further analysis of the result indicates the completion of a scientific workflow for a certain task [43].

2.2 Scientific Workflow Management Systems

A Scientific Workflow Management System (SWfMS) is designed specially to construct and execute a series of computational or data manipulation steps for a scientific application. In this new era, conducting and managing scientific experiments has become extremely challenging. Because of the complex computational applications and the increasingly large volume of data, scientific experiments have become more resource-oriented and time-consuming. Moreover, scientific data and instruments such as machines and sensors have made scientific data management even more challenging [84]. In this situation, an integrated framework that can help conduct scientific experiments with less complexity and less computational time is strongly needed. For this purpose, workflow management systems have emerged in recent years to facilitate scientific experiments in various possible and renowned sectors such as biology, bioinformatics, geology, environmental science, eco-science and eco-informatics. The importance of workflow management systems can be understood even by following a simple example. We can think of a scientific experiment that detects fruits from captured crop field images by drone. Here several steps are essential to execute the experiment. Such as loading the images into the analysis software from drone storage, extraction, submitting to core detection modules, saving the detected areas into the database for future analysis etc. All these steps are needed to be done one by one for the successful execution of the scientific process. But these steps can be done better by a SWfMS in various ways and with less amount of time. That is why developing scientific workflow management systems with unique features has become a trendy topic for researchers and many types of research have been conducted over the year to reduce the complexities that a scientist can face while designing the workflows.

We can understand that scientific workflow systems basically automate the scientific workflow execution by assisting in workflow design, composing, managing and sharing of a workflow description. Some other cruicial functions are also supported by workflow management systems such as workflow optimization, execution monitoring, recording, querying provenance information etc. A great number of scientific workflow management systems have come into the light to control the completion of various workflows on heterogeneous computing resources [25] [32][5] [44] [100]. SWfMSs offers an integrated development environment for the e-science developers to program/specify workflows at a high level where a workflow can be treated as an algorithm with defined inputs and outputs. While SWfMSs has been gaining momentum to support software analytics, e.g., Mostaeen et al. [68] made use of SciWorCS to validate clone detection output, they have been heavily used in other domains as well, such as Bioinformatics [44][69].

2.3 Galaxy

Galaxy [34] has initially been introduced to examine and inspect the genomics HTS data that is a collection of free bioinformatics tools being immensely powerful, flexible, dynamic, easy to use and accessible using any web browser, including mobile devices. The visual web scripting language within Galaxy allows the user to create of an automated analysis process of workflow. The Galaxy workflow can be shared and executed with user defined data and parameters [99]. The Galaxy platform was designed and developed to explore the area of an open, web-based approach to address the scientific experimental challenges and to facilitate genomics research. It has gained popularity as a web based genomics workbench that can enable users to perform the computational analyses of genomics data [20].

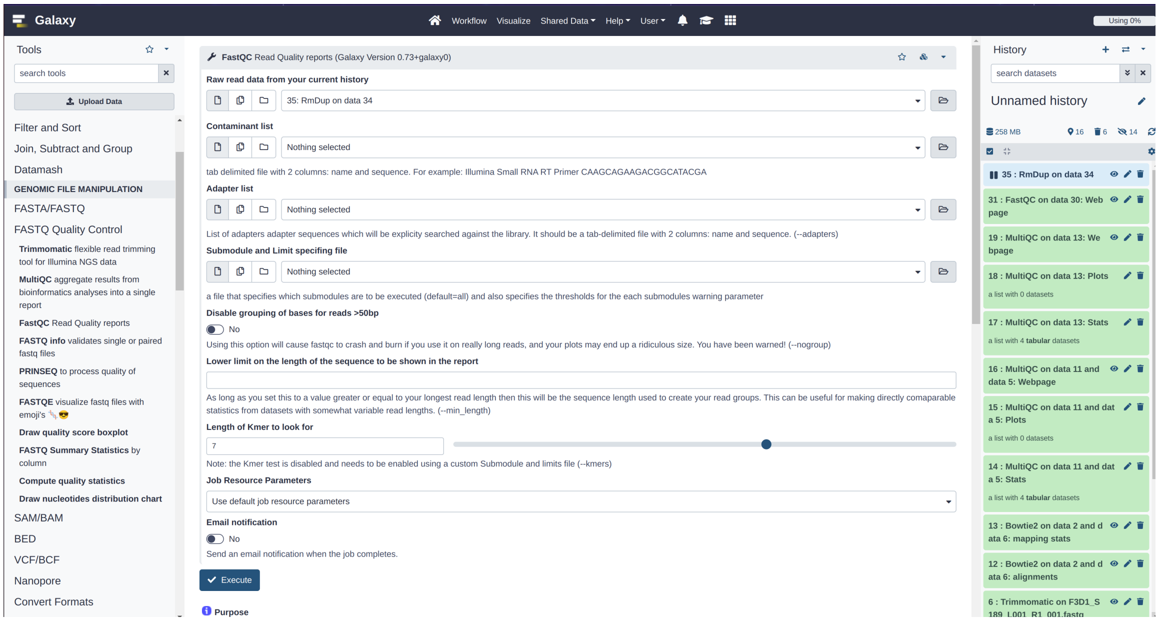
Galaxy is architecturally a modular python based web application. It was developed in 2005 and since then, it is enabling biologists who do not have programming knowledge and system administration expertise, they can perform computational analysis through this system. Galaxy provides several services, including analyzing tools, genomics data, tutorial demonstration, persistent workspaces and publication services available to any scenarists who have an internet connection with a consistent web interface. Galaxy allows its users to perform multi-step analyses by running these tools and it also stores the entire provenance of every analysis step. Galaxy tries to connect the tool developers and researchers by helping them in accelerating their research. It has several components [20]. They are: *Public Galaxy Server* - it is available since 2007 and anyone can use it to analyze their data without any cost. The site offers substantial CPU and disk space, which makes it possible to analyze large datasets, *Galaxy software framework* - it can be deployed in any Unix system and is highly customizable, *Galaxy Tool Shed* - in this place, tool developers upload and share their tools with the community for use. The tool shed is the central location for every tool. It also provides a description of how to install any tool and its necessary dependencies and *Galaxy Community* - The galaxy community consists of users, tool developers and administrators who maintain Galaxy instances.

Galaxy is driven by three of their major goals [20]. They are -

1. Galaxy explores to increase the access to complex computational analysis for all sorts of scientists, that includes both the users who have or no programming knowledge.

2. Galaxy provides a web based graphical user interface which makes it simple and easy to do everything required for analyzing relatively large datasets.

3. Galaxy allows its users to experience a collaborative and transparent analysis by enabling users to share and publish their analysis via the web.


Galaxy
Workflow
Visualize
Shared Data
Help
User
Using 0%
Tools
search tools
Upload Data
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
Trimmomatic flexible read trimming tool for Illumina NGS data
MultiQC aggregate results from bioinformatics analyses into a single report
FastQC Read Quality reports
FASTQ info validates single or paired fastq files
PRINSEQ to process quality of sequences
FASTQE visualize fastq files with emoji's 🧬😎
Draw quality score boxplot
FASTQ Summary Statistics by column
Compute quality statistics
Draw nucleotides distribution chart
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Execute
Purpose
History
search datasets
Unnamed history
258 MB
16
6
14

FastQC Read Quality reports (Galaxy Version 0.73+galaxy0)

Raw read data from your current history
35: RmDup on data 34

Contaminant list
Nothing selected
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list
Nothing selected
List of adapters adapter sequences which will be explicity searched against the library. It should be a tab-delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifing file
Nothing selected
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp
No
Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report
As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths. (--min_length)

Length of Kmer to look for
7
Note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

Job Resource Parameters
Use default job resource parameters

Email notification
No
Send an email notification when the job completes.

35 : RmDup on data 34
31 : FastQC on data 30: Webpage
19 : MultiQC on data 13: Webpage
18 : MultiQC on data 13: Plots
a list with 0 datasets
17 : MultiQC on data 13: Stats
a list with 4 **tabular** datasets
16 : MultiQC on data 11 and data 5: Webpage
15 : MultiQC on data 11 and data 5: Plots
a list with 0 datasets
14 : MultiQC on data 11 and data 5: Stats
a list with 4 **tabular** datasets
13 : Bowtie2 on data 2 and data 6: mapping stats
12 : Bowtie2 on data 2 and data 6: alignments
6 : Trimmomatic on F3D1_S189_L001_R1_001.fastq

Figure 2.2: The Galaxy interface consists of the tool menu (left panel), tool interface (center panel) and history (right panel)

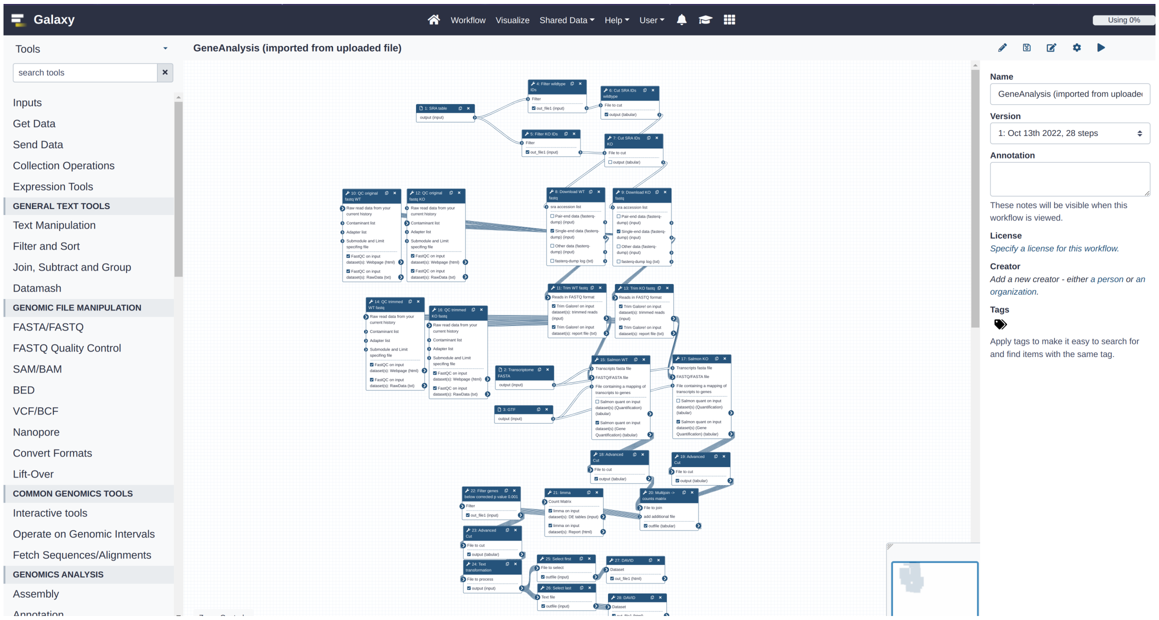
Galaxy provides some core features to its users. They are:

- *Analysis interface and history*: The Galaxy analysis interface consists of three panels. One is for showing available tools, one for running the tools and the last one is for viewing the dataset and history of datasets created by running the tools. The tools are categorized into their groups and one can search for any specific tool through the search option provided in the tool panel. When a tool is selected, it is shown in the main window, where its inputs and parameters are set and the tool is executed. After executing a tool, the output dataset is added to the history panel. Figure 2.2 shows the working interface of Galaxy.

- *Workflows*: Galaxy provides an easy to use workflow editor for creating multi-step analysis easily and efficiently by using the drag and drop functionality. Here in this editor, tools can be added and connected to each other to get the output of one tool and the output can be used as an input of another tool. The tool parameters can be set and changed according to the requirements of the experiment. Figure 2.3 shows the workflow editor with a sample workflow.

- *Sharing and publishing*: The datasets, created histories, workflows and visualizations can be shared with individuals or they can be shared via the web. The published items are listed together and are shared with all the users. One of the main objects of Galaxy is the pages. Pages are web interactive web documents that includes embedded and interactive Galaxy objects.

The Galaxy developer community has provided necessary and required tools for the users to perform their analysis. The user interface also provides a simple and efficient platform for large scale computing. Although



GeneAnalysis (imported from uploaded file)

Name: GeneAnalysis (imported from uploade

Version: 1: Oct 13th 2022, 28 steps

Annotation

These notes will be visible when this workflow is viewed.

License: *Specify a license for this workflow.*

Creator: *Add a new creator - either a person or an organization.*

Tags

Apply tags to make it easy to search for and find items with the same tag.

Figure 2.3: The central workflow editor, showing a sample workflow

one of Galaxy's main goals is to provide such a system that would help both the users with or without programming and expertise knowledge, there has not been any tool developed where people can express their requirements in natural language and they will be provided with necessary resources. Again the user forum of the Galaxy system lacks many issues and shows the problems users are facing while using this system.

2.4 Qualitative Analysis

Qualitative data analysis is a process where analysis is closely scrutinized. The analysis process aims to interpret qualitative data by transforming it into valuable findings and conclusions. The results are presented as thick descriptions, overarching themes, and detailed investigations of varying socio-cultural realities. A researcher needs to follow some certain steps before engaging in this process. They must have a dataset that qualifies for performing a qualitative research on it [39]. Some of the examples of this type of datasets are - non numeric data, interviews, written presentations of speech and activities, images and other textual data such as letters, emails, forums etc. Qualitative analysis is the directly opposite of quantitative research as in quantitative research; researchers need to perform statistical analysis and other varieties of numerical and predictive measurements. Qualitative data analysis is meticulous, complicated, repetitive, nonlinear and intellectually demanding [63]. Unlike the quantitative research, qualitative research focuses on exploring values, meanings, briefs, thoughts, experiences, and feelings characteristics of the phenomenon under investigation [105].

In qualitative research, data analysis is considered as the process of searching and arranging textual data systematically to increase the understanding of the phenomenon. The data analysis mainly involves categorization of the data according to the similar categories. Fundamentally, this process involves in making understanding of huge amounts of data by decreasing the raw information volume. Through this process, significant patterns can be identified, and meaningfully information can be drawn and logical chain of evidence can be built too [105]. The main purpose of doing qualitative analysis is to let the computer work better in doing the quantitative calculations. The reason is that computers are not going to analyze or create categories, code and make decisions based on patterns and draw meaningful findings from the data. Due to the nature of the qualitative research, the use of computers in this sector is very limited or close to none. There are five types of quality analysis methods [94]. They are:

1. *Content Analysis*: In this research method, examination and quantification of certain words, subjects and concepts are taken into account. Through this method, qualitative input is transformed into quantitative data, which helps to make reliable conclusions about how users perceive about the system or how they can improve their experiences and provide the opinion. In content analysis, it is not required to engage directly with the participants to gather the data. The process can be replicated easily when it is standardized and automation of the process can be done manually. However, this process is very time-consuming and the results can be easily affected by the subjective interpretation [77].

2. *Thematic Analysis*: In this analysis method, identification, analysis and interpretation of patterns take place in qualitative data. This method can be done using several tools. Thematic analysis is different from concept analysis in terms of identifying the frequencies and recurrence of words and subjects. Moreover, thematic analysis can only be performed in qualitative data and mainly aims to identify patterns [77].

3. *Narrative Analysis*: In this method, interpreting research participants' stories are taken into consideration. The data used in this method are like case studies, interviews, testimonials and other textual or visual data. In heavily structured interviews or in written surveys, the narrative analysis does not work as in those cases, and participants are not given much opportunities to narrate their stories in their own words [88].

4. *Grounded Theory Analysis*: In this method, qualitative research is conducted by examining real world data to develop theories. The technique aims to create hypotheses and theories by collecting and evaluating qualitative data. In contrast with other analysis methods, in this process, theories are developed from the data not the other way round [104].

5. *Discourse Analysis*: In discourse analysis, through concrete research, the underlying meaning is drawn from qualitative data. It includes the observation of texts, audio and videos to investigate the relationship between the information and its context. Unlike the content analysis, in this method, the main focus is given to the contextual meaning of language. It emphasizes the perception of a topic from the audience and what they feel about the topic [50].

In our study, we have performed the content analysis. As we tend to examine and quantify of certain concepts and make reliable conclusions about improving the users' experiences, thus we took content analysis into consideration.

2.5 Natural Language Processing

Natural language processing (NLP) is a branch of computer science and more specifically, it can be said that natural language processing is the branch of *Artificial Intelligence* or *AI* [62]. The main goal of the processing methods is to provide computers with the ability to understand text and spoken words in a similar way human beings can understand. The way humans communicate with each other is described as natural language. However, humans use a different modes of communication platforms among them such as signs, emails, SMS (Short Message Service), web pages etc. All these types of data are important and analyzing them opens new doors to research fields and also the findings from the analysis are beneficial to the mankind. But working with natural language has its own disadvantages too. For example, natural language is messy and primarily hard to interpret for any machine. In this context, methods to natural language processing bring high hopes and make the analysis easier. NLP makes it possible for computers to read text, hear speech, measure sentiment and determine which parts are important. In today's world, machines are capable of interpreting more language-based data than any human can do. Moreover, as we communicate with each other in various ways, the data quality reduces. Grammar and syntax errors, misspellings, using unknown terms and slangs - all of them are common in today's data source. NLP helps to resolve the ambiguity in human data and also adds useful numeric structure to the data for making it suitable for the applications [19], [62].

The two key factors that NLP combines are - computational linguistics and rule-based modeling of human language with statistical, machine learning and deep learning models [62]. The combination of these technologies brings the ability of the computers to process human language in text or voice data format and also tries to understand the intent and the sentiment of the writer. The most amazing feature of NLP is that it can drive computer programs to translate text from one language to another, it can give response to verbal commands and also the summarizing of large volume text data is possible with the help of this technology. NLP is not a new science to the mankind, but its popularity is gaining day by day. With the advancement of ongoing interest in converting human language to machine language and the availability of big data with powerful computational resources, NLP technologies are blooming.

NLP consists of many different techniques to interpret human language which range from statistical and machine learning methods to rule-based and algorithmic approaches. As there are many variations, remain in text and voice data, so several types of approaches are needed for processing the data. Tokenization and parsing are the methods of basic NLP techniques. After that, lemmatization or stemming, identification of semantic relationships, detection of language, part of speech tagging - all these techniques play a vital role in processing textual or spoken words data [70]. Generally, NLP breaks down the language into shorter forms and tries to understand relationships between the short forms. Meaningful findings are also achieved from the shorter pieces. The techniques are used in higher-level NLP applications such as content categorization, topic discovery and modeling, corpus analysis, contextual extraction, sentiment analysis, speech-to-text and text-to-speech conversion, document summarization and machine translation [41].

2.5.1 Natural Language Processing Pipeline

The NLP pipeline consists of some steps to read and understand human language [70](Figure 2.4). They are:

1. *Sentence Segmentation:* It is the first step in the NLP pipeline. The main task here is to divide an entire paragraph into several short sentences to understand the context in a better way. For example: “Saskatoon is the largest city in the Canadian province of Saskatchewan. It straddles a bend in the South Saskatchewan River in the central region of the province. It is located along the Trans-Canada Yellowhead Highway and has served as the cultural and economic hub of central Saskatchewan since its founding in 1882 as a Temperance colony”: source Wikipedia. Sentence segmentation will make the following result:

    - Saskatoon is the largest city in the Canadian province of Saskatchewan.
    - It straddles a bend in the South Saskatchewan River in the central region of the province.
    - It is located along the Trans-Canada Yellowhead Highway and has served as the cultural and economic hub of central Saskatchewan since its founding in 1882 as a Temperance colony.

2. *Word Tokenization:* In this process, the sentence is broken down into separate words or tokens. For a better understanding of the context, tokenization is really helpful. For example, when the sentence “Saskatoon is the largest city in the Canadian province of Saskatchewan” is tokenized, it will result in the following tokens: “Saskatoon”, “is”, “the”, “largest”, “city”, “in”, “the”, “Canadian”, “province”, “of”, “Saskatchewan”.

3. *Stemming:* Stemming turns the word into its root form. In other words, stemming helps to detect the parts of speech for each token. For example, writing, writes and written - these words have the same root word *write*.

4. *Lemmatization:* Lemmatization discards the inflectional endings and returns the canonical form of a word. Lemmatization is quite similar to stemming, except that lemma is an actual word. For example, ’eating’ and ’eats’ are forms of their root word ’eat’ and here, ’eat’ is a proper word. In stemming, it is not necessary that the stem word would be a proper word.

5. *Stop Word Analysis:* The importance of each and every word in a sentence is not the same. In English language, some of the words appear in a sentence frequently but carry less value to the context of the

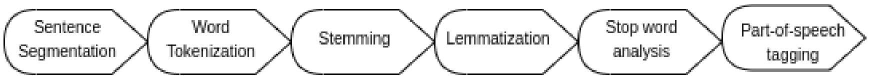


Figure 2.4: NLP pipeline

word. So these types of words need to be removed. They are filtered out so that NLP methods can give more focus on the other important words.

6. *Dependency Parsing:* In this process, the relation between the words is explored. A tree can be built to find out the dependency. Each of the single words can be assigned as a parent word and main verb is considered to be the root node always.

7. *Part-of-speech Tagging:* Parts-of-speech tags has consisted of verbs, adverbs, nouns and adjectives which help to indicate the meaning of words in a grammatically correct way in a sentence.

2.5.2 Cosine Similarity

Cosine similarity is popularly known as a metric in information retrieval and related studies [85]. This metric converts a text document as a vector of terms and with the help of this model, similarity between two documents' term vectors can be derived. The cosine value is calculated between two documents and the implementation can be applied to any two texts such as sentences, paragraphs or whole documents. Cosine similarity calculation is very well known to any search engine because the similarity score between the user's query and documents is checked. The score is then sorted to show the result from the highest one to the lowest one. The higher the similarity score between two texts, the higher is the relevance between them.

Cosine similarity is particularly used in high-dimensional positive spaces. The working procedure of cosine similarity involves measuring the similarity as the cosine of the angle between two vectors [107]. Two similar vectors are expected to have a small angle between them. In document-query cases, a document can be expressed as a term vector. The vector's dimensions refer to the terms available in the document. Inside a document, the dimension's value is considered as the occurrence of the term. Thus, a document can be presented as a vector form as:

$$\vec{d} = (w_{d0}, w_{d1}, ..., w_{dk}) \tag{2.1}$$

The query also can be expressed as a vector form as:

$$\vec{q} = (w_{q0}, w_{q1}, ..., w_{qk}) \tag{2.2}$$

where w_{di} and w_{qi} are float numbers that indicates the occurrence of each term inside a document. So the similarity between two vectors can be defined as:

$$Sim(\vec{q}, \vec{d}) = \frac{\sum_{k=1}^{t} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2} . \sqrt{\sum_{k=1}^{t} (w_{dk})^2}} \tag{2.3}$$

The similarity score is represented as the *cosθ* value between the document and query. Based on the cosine principle, cosine 0 is 1 and less than 1 to the value of another angle, then the value of the similarity of the two vectors is said to be similar when the value of the cosine similarity is 1 [59].

2.5.3 Fuzzy Matching

Fuzzy matching is also called approximate string matching. This is a technique that identifies two elements of text, strings or sentences that are mostly similar but not the same in an exact way. Its applications are mostly seen in search engines. Fuzzy matching follows an algorithm that is not only dedicated to looking for exact similar strings but also it quantifies the closeness of the two strings to each other. It mainly uses a distance metric known as *edit distance*. The metric calculates how close two strings are by calculating the minimum alterations needed to be done to convert one string into another. There are several types of edit distances to be used to calculate the closeness score such as Levenshtein distance, Hamming distance, Jaro distance etc. The fuzzy string matching technique uses Levenshtein distance in its calculation.

Fuzzy matching techniques can come real handy in many situations. For example, in ensuring data accuracy, fuzzy matching plays a vital role. In a recent study, it has been found that more than 60% of companies have implemented machine learning-based solutions. But in these cases, data accuracy became extremely crucial as the companies mostly rely on machine learning and artificial intelligence. There are groundbreaking researches ongoing to improve the performance of the neural network and machine learning technologies. But what really is missing here is that it has not been given proper notice on whether these models are fed with good quality data or not. For a machine learning model to work without accurate data is somewhat similar to launching a rocket to mars using compressed natural gas. In this context, fuzzy string matching comes with great help by improving the data quality and accuracy with the help of data duplication, identification of false positives etc. Moreover, in detecting fraud inside an organization, fuzzy string matching also comes in handy. There are several use cases for fuzzy string matching techniques such as spelling correctors, classification of genome data etc.

Now let's talk about the Levenshtein distance that has been used in fuzzy matching for calculating the closeness of two texts. It was introduced by Levenshtein in 1966. The Levenshtein distance is a metric that is used to compute the difference between two string sequences. The higher the distance score is, the more differences there are between two strings. The distance is considered as the number of deletions, insertions or substitutions necessary to convert the source string to the target string [37]. The procedure for computing the Levenshtein distance between two strings $X = x_1, x_2, ..., x_m$ of length m and $Y = y_1, y_2, ..., y_n$ of length n is to measure step by step in a matrix of order $m \times n$ edit distance between sub-strings of X and Y. The corresponding values are preserved in the matrix up to the box *(m,n)* that represents the minimum distance between the two matrix X and Y. Mathematically, the Levenshtein distance can be defined as follows:

$$lev(a,b) = \begin{cases} |a| & if|b| = 0, \\ |b| & if|a| = 0, \\ lev(tail(a), tail(b)) & if a[0] = b[0], \\ 1 + min \begin{cases} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b)) \end{cases} & otherwise, \end{cases} \tag{2.4}$$

Here, the Levenshtein distance between two strings *a*, *b* is calculated by *lev(a,b)*.

Now, *Fuzzywuzzy* is a Python library that uses Levenshtein distance to compute the difference between strings in a simple-to-use package [87]. It has been developed and open-sourced by SeatGeek, a service dedicated to finding sport and concert tickets.

2.6 Likert Scale

The Likert scale is one of the leading, most fundamental and most popular psychometric tools in educational and social sciences research. Because of its pattern of analysis and point inclusion, it is also one of the subjects of skepticism and debates [51]. This scale is well known for the measurement of attitudes and opinions with a greater degree of nuance. The Likert scale uses five or seven points which are sometimes referred to as a satisfaction scale ranging from one extreme attitude to another. Typically, this scale also includes a neutral or moderate opinion in its scale. In comparison with the binary surveys, Likert scale gives the participants more detailed options to express their opinions in the form of, e.g, “very satisfied” or “somewhat satisfied” or “excellent” etc [73]. Likert scales are mostly used when the questions are focused on one topic and the questions can keep the survey participants happy as they can select an option which is not very hard to think about. Also using this scale can improve the data quality. One of the use cases of Likert scale can be surveying about customer satisfaction. This scale can be used to measure the customer satisfaction with using the customer care service.

For example, there can be a question: *how much satisfied or dissatisfied are you with our customer care service?* The options for choosing an answer would be like this - 1. Very satisfied, 2. Satisfied, 3. Neither satisfied nor dissatisfied, 4. Somewhat dissatisfied, 5. Very dissatisfied.

Likert scales are used when we need to dig down into a topic where we need the opinion of the users on how they find the usability of that topic. That is why we used Likert scale in our study too to find out the user opinion.

2.7 NASA-TLX (Task Load Index)

NASA-TLX is a multi-dimensional rating scale that derives a sensitive and reliable estimation of the workload by combining the magnitude and six workload-related factors [38]. The workload score is measured based on a weighted average of ratings on the six subscales including mental demand, physical demand, temporal demand, performance, effort and frustration level [15]. The Human Performance Group at NASA' AMES Research Center first developed this index and it took over a three-year time frame and more than 40 laboratory simulations to accomplish the development cycle. The main research goal was to develop a workload rating scale that can provide a significant summarization of the workload variations within and between the tasks. This scale is measured in the range of 1 to 100 where 1 point increment suggests greater sensitivity to experimental manipulations. Table 2.1 gives a better understanding of the measurement criteria of the NASA-TLX scale.

Title	**Endpoints**	**Description**
Mental demand	Low / High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low / High	How much physical activity was required (e.g.. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low / High	How much time pressure did you feel due to the rate or pace at which the tasks occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Low / High	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low / High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration level	Low / High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table 2.1: NASA-TLX Rating Scale Definitions [38]

2.8 Threats to Validity

Checking the threats to validity holds a very significant place in research designs. An easy way to understand this is that a hypothesis may be tested in a different manner other than how the researchers have tested it. In our study, we discuss two types of validity - internal and external validity.

Internal validity checks whether the study design, conduct and analysis answer the research questions without having any bias. External validity examines whether the study findings can be conducted in a generalized way in other contexts [6]. Many factors can threaten the internal validity of a study which includes errors in measurements or the participation selection in the study. The researchers are expected to think about these errors and avoid them. When the researchers established the internal validity, they can proceed in the process of judgement regarding its external validity by asking whether the study results apply to similar participants in a different setting or not. External validity refers to the extent where the study result is generalizable, especially for the population that the sample is thought to represent. The lack of internal validity indicates that the study result can deviate from the truth. Nevertheless, the lack of external validity indicates that the result may not be applicable to different sets of population samples [78].

2.9 Summary

In this chapter, we introduced different terminologies and background concepts that would help one to follow the remaining of the thesis. We discussed the scientific workflow, its life cycle, the workflow management systems and a brief overview of Galaxy. We also described the natural language processing techniques and their methods. We discussed cosine similarity and fuzzy matching. We gave an overview of the description of internal and external threats to validity. At last, we gave some ideas about the Likert Scale and the NASA-TLX scale.

3 Related Studies

In this chapter, we discuss the related literature of my thesis. Section 3.1 shows the study done on finding usability problems through group discussion. Then, Section 3.2 reveals the usability challenges presented in SWfMSs. After that, Section 3.3 shows the application of content analysis in different forums to find out the first research problems, second, section 3.4 discusses how cosine similarity has been used for suggesting tags and last, Section 3.5 discusses the different natural language processing pipelines to find out significant information from user data and also converting natural language queries to usable tools for the users.

3.1 Finding Usability Problems Through Group Discussion

Finding usability through group discussion is not a new topic in the research field. Pyae et al. [81] investigated the usability and user experiences of voice user interface through a social-media based interest group. They reached out to the posts of 114 users who actively participate in the group discussion. Their findings turn out to be very insightful for the developers and researchers for future research in voice user interfaces. Folstad [33] found that group discussions can affect the output of a usability inspection. His study showed twenty five percent new usability issues were generated through group discussion compared to the individual evaluators. The importance of group discussion was perceived by Twidale et al. [101] where they examine several project's bug reports. Their aim was to find out developer's point of view in addressing and resolving the issues concerned with interfaces and interaction design. Their study also explored how open source projects addressed the usability issues through the bug reports and discussions. Chen et al. [18] analyzed more than three million user interface (UI) related reviews to understand the user's point of view towards the design of user interfaces. One of their study key findings was customization which indicated that if users can customize the user interface according to their personal preferences, then they can improve the usability of the UI. Discussion mining has been practicing in almost every popular web based forums for quite a long time. P.S. Kochhar [55] mined testing related questions asked on Stack Overflow to find out the common challenges and important topic of discussion. His study brought out some important topics which are discussed more frequently than the other topics. Similar study has been conducted by Bi et al. [11] where (QA-AT) Quality Attributes - Architectural Tactics knowledge is mined and analyzed using Stack Overflow as a source. The goal of the study was to mine the knowledge effectively and to help developers by identifying the relationship between ATs and QAs. So exploring and investigating group discussions plays significant role in finding out the usability issues of software systems.

3.2 Usability Challenges in SWfMSs

With the advancement of high-performance computational research, improving usability for scientific workflow management systems became high in demand. Sonntag et al. [96] found in their study that only a few researchers have been conducted focusing on the necessity of improving the conventional workflow technology as well as accommodating all the essential requirements of the scientists. Their study provides technical views on the problems that come up while deploying traditional workflow technology in the scientific area. They also recommended ways to resolve the barriers faced by the scientists in this aspect. The key findings of this study are the missing features of SWfMSs that are hindering the usability of these systems. Some of them are not having integrated tools, non-explicit deployment procedures, not being able to start the workflows within the modeling tool, lack of monitoring, lack of flexibility and inefficient provenance tracking system. All of these features are needed for improving the usability according to the study.

Another study by Gŏrlach et al. [35] focused on the ability of conventional workflow technology to fulfill requirements of scientists and scientific applications. Their study suggested that scientists can be benefited with the features of conventional workflows but some improvements are also necessary. The authors showed that the usability of SWfMS is strongly dependent on the supported tools that are needed for the realization of the whole workflow life cycle. Usability hinges on easy-to-use tools in combination with wide-ranging automation. But the automation in workflow technology needs some improvements also to meet the requirements of usability. Furthermore, some other issues are also discussed in this study to improve the usability of SWfMSs. Implementing expressive modeling language, achieving convenient data handling technologies, improving the flexibility mechanism are dominant among them.

A formative usability study investigating the usability of different WfMSs in the biomedical data analysis field has been conducted by Jelonek et al. [49]. They performed the study on one graphical user interface based workflow system and another on one script-based workflow system. The study addresses certain questions about the resource utilization and management of the usability in these systems. Many usability problems have been revealed by the study either in tools or different levels of severity. The most dominant among them was the barrier faced by the non-programmers when they start using a workflow management system. Script-based workflow systems are not useful for the life science group. Moreover, the novice users face difficulties even with the simplest of the changes in the scripts and they were only be able to solve the issue with the assistance of the expert users. Hence, the authors suggested to provide tutorials or training for them. Another usability challenge that came into light is that the users face difficulties while understanding or interpreting the result of the experiment. Through the study, it is quite evident that specific points of improvement are necessary for different workflow management systems.

A comparative study has been done by Bhagwanani [10] over the user interfaces between graphical user interfaces and the other available solutions. The study finds that despite putting a lot of efforts from the workflow systems to build a user friendly system, there are overwhelming and unavoidable complexities remaining in these systems. Their study showed that a visual component of a SWfMS is very significant. Scientists will be able to achieve their desired output of the experiments when they are provided with ease-of-use and intuitive visual elements. The SWfMSs are already very complex in nature because of the heterogeneous data and tools and their interactions with them. The study finds that with the increasing rate of complexity of the underlying architecture, the number of components in the user interface increases exponentially which makes the system more difficult to use.

3.3 Content Analysis of User Data

Context analysis of user data is a very renowned and old-practiced form of analysis. In many research cases, context analysis has been the first task to perform. Xiao et al. [108] used context analysis to precisely provide recommendation for personalized mobile services that ensures privacy concerns. In their study, they introduced six dimensions of privacy concerns to propose a novel approach for personalized mobile recommendation service based on relevant contextual information by performing a context analysis. Qian et al. [82] performed content analysis to find out the features of social support from online smoking cession communities to provide better user communication and interaction. They collected 2758 replies from 29 active replies and 408 correlated posts from a Chinese online smoking cessation forum, Baidu Tieba. Their analysis found that in most cases, emotional support is the most typical feature or support that the community provides. Other than that, informational support is also very beneficial for users to quit smoking.

To find out the relationship between data science to its library and information science, Virkus et al. [103] performed a content analysis of 80 publications. Six broad categories were included in those publications. They found several categories addressed by the scientists such as tools, techniques and application of data science, followed by data science from the health science perspective, education and training and skill and knowledge of data professionals. Not only data science, to find out the current research methods in machine learning, Kamiri et al. [52] performed a content analysis of a total of 100 articles published since 2019 in IEEE journals. Their study found that the use of quantitative research methods with experimental research designs is now mostly used by the machine learning practitioners. Their study also revealed that optimal feature selection has become one of the key method to the researchers to optimize the machine learning models performance. Arya et al. [7] emphasized on the comments that are posted on the issues of the different open source software projects. They find it very significant that those posted comments are embedded with rich information containing information about software projects that can possibly help the open source software stakeholders to meet their diverse needs. In their study, they performed a qualitative content analysis of 15 complex issue threads from three projects in GitHub. They uncovered 16 information types and also created a labeled corpus which contains 4656 sentences. Their work represented a significant first step towards tools and methods to identify the rich information stored to support the open-source software stakeholders.

3.4 Cosine Similarity and Tag Suggestion

The application of cosine similarity in automatic tag suggestion is very renowned. Sriharee [97] proposed an auto-tagging methodology using tags defined in ontology. Classification process and tag selection process are two of the processes for auto-tagging methodology. The classification process considers semantic analysis including the term-weight matrix and cosine similarity. Al-Attar et al. [4] proposed a domain specific algorithm to suggest Arabic tags. Their algorithm is implemented by using three similarity measures and they are: Cosine, Dice and Jaccard. For building tag based recommendation system, using cosine similarity is getting popularity. Sen et al. [92] designed and evaluated algorithms that could predict the users' ratings on the basis of their inferred tags for movies. The study showed that the application of cosine similarity in recommending tags has always been on top of the list where the cosine score between the user preferred tags and the most representative tags.

3.5 Natural Language Processing In User Data Analysis

NLP methods are being used in analyzing user data for quite a long time. Many researches have been conducted in this field to acquire significant and useful findings from the user data. Setlur et al. [93] proposed a natural language interface, Eviza, which provides an interface with an interactive query box and visualization system to build an effective system for the users' analysis. Eviza allows the user having interaction with their data by make it able to revise and update their queries repeatedly. They found it promising that natural language interfaces are a groundbreaking approach to interact with data and make analytics accessible to a greater audience.

Sentiment analysis based works have been done with the help of NLP methods in a lot of research studies. Ke et al. [53] proposed a NLP based framework to analysis sentiment that provides enough domain knowledge to generate additional labeled data to answer the question of why a user's message produces certain emotions. They followed a rule based semantic parser that converts the explanations into programmatic labeling functions. Their study proved the potential efficiency and effectiveness of semantically augmented data in combating the labeled data scarcity and class imbalance problems of publicly available sentiment analysis datasets. Some other researches have been conducted in sentiment analysis with the help of NLP methods in different sectors such as in clinical data analysis [86], in health care system [2], on Twitter data [40] and so on.

Some other researches have been done in turning natural language queries in technical terms. Xu et al. [110] found the challenge for the users where they face difficulties in turning their concepts into code syntax. They tried to find the answer of the question that whether the current state of technology improves developers' productivity and whether there are any remaining gaps or challenges. Based on the answer they developed a PyCharm IDE plugin which implements a hybrid code generation and code retrieval functionality. They also conducted a user study with or without the plugin and found that the developers are largely positive about the increased productivity by the use of that plug-in. Similar kinds of researches are done in [54], [111], [71] etc.

Another very promising field of research with NLP methods using user data is developing chatbots. A great number of researches followed the idea of implementing an auto question answering interface which is very efficient for improving usability of a system. Chao et al. [16] studied emerging technologies for NLP based intelligent chatbot development using a systematic patent analysis approach. In their study, they implemented text mining techniques which also includes document term frequency analysis for extracting key terminologies, clustering method to identify the subdomains and LDA [48] (Latent Dirichlet Allocation) to find the key topics of patent set. A literature survey was done by Caldarini et al. [14] to review the recent progress on chatbots and the challenges and limitations persisting in this area. They made noteworthy recommendations for future research investigation. Another study was conducted by Abdellatif et al. [1] about the comparison of natural language understanding (NLU) platforms for chatbots in software engineering. They indicated the open challenge of selecting the best natural language understanding components for developing chatbots. In their study, they evaluated four most commonly used NLUs which are Google Dialogflow, IBM Watson, Rasa and Microsoft LUIS. They used two datasets for their study that includes the commonly performed tasks by the software engineering practitioners. Their findings showed that IBM Watson performed the best in intents classification having a F1-measure greater than 84% although Rasa shows with a higher median confidence score. The key finding of the study is to provide guidance to the software engineering practitioners on deciding which NLU to use in their chatbots. Some other studies have been done in the same field such as developing cloud based chatbots [80], whether chatbots are useful for human resource managements [67], their efficiency in generating multiple replies [17] and so on.

3.6 Summary

From our above discussion, we see that many researches have been conducted to find out the usability of the software systems. But to the best of our knowledge, improving the usability in SWfMS is still a new area to be explored for the researchers. Also, the automatic tag suggestion methods are proposed for similar type of platforms or for building recommendation systems. Not only that, we did not find any researches done on natural language query to workflow recommendation. In contrast to the former related studies, in this thesis, we find the usability problems users are facing in the scientific workflow management systems. We propose a better tag suggestion model based on the previously posted questions for domain specific platform. Moreover, we are introducing a new tool to convert natural language query to workflow recommendation system. We believe our study will help the users of the SWfMS to experience a better usable and effective environment while conducting their studies.

4 Investigating the Galaxy User Forum and Categorizing The Problems Users Are Facing

In this chapter, we discuss our first study that categorizes the issues users are facing mostly while conducting their experiments in the Galaxy platform. In Section 4.2, we explain the methodology of this study that we follow. Section 6.5 shows the result we get from our study. Section 4.4 discusses the threats to the validity of this study and Section 4.5 summarizes the chapter.

4.1 Introduction

The Galaxy community and the research community aim to provide remarkable support through a dedicated web forum[1] to the users' who are posting their concerns on various topics. This forum is known as community help for Galaxy users. But what type of posts users are sharing here and what type of problems they are facing frequently, are not known to the community. We aim to analyze the posts to classify them into several categories and search for similar patterns of the posts. Among the posts, there are several issues that have surged up showing the lack of efficient usage of the services. Users struggle to resolve some of the issues because of the scarcity of proper resources and expert help. Addressing these issues can definitely increase the site's usability and users' interest to work with more enthusiasm. We can find the problems and difficulties users are facing while using the system. They post their inquiries about the errors they faced, ask to get some suggestions from the experts and also ask for solutions for problems that occur while conducting their experiments. So we can get insights into the usability of the system by analyzing these posts.

The word 'usability', does only mean that the website is interacting with the target users in an acceptable manner [57]. The usability of a web application depends on the usage of an average person whether they can achieve specific goals and get complete satisfaction. Usability implies that a specific web service can deliver its contents to the users in an efficient and ordered way. A renowned web-usability consultant, Jakob Nielson, considered usability as a combination of different contributing aspects. These factors are - efficiency of use, ease of learning, navigational efficiency, subjective satisfaction and error frequency and severity [74]. In compliance with that, in this study, we want to find out whether Galaxy also focuses on the usability of the applications and whether the users of this platform feel the same. We provide the first attempt to understand the challenges of Galaxy users face by investigating what users are asking on the forum. Our

[1] https://help.galaxyproject.org/

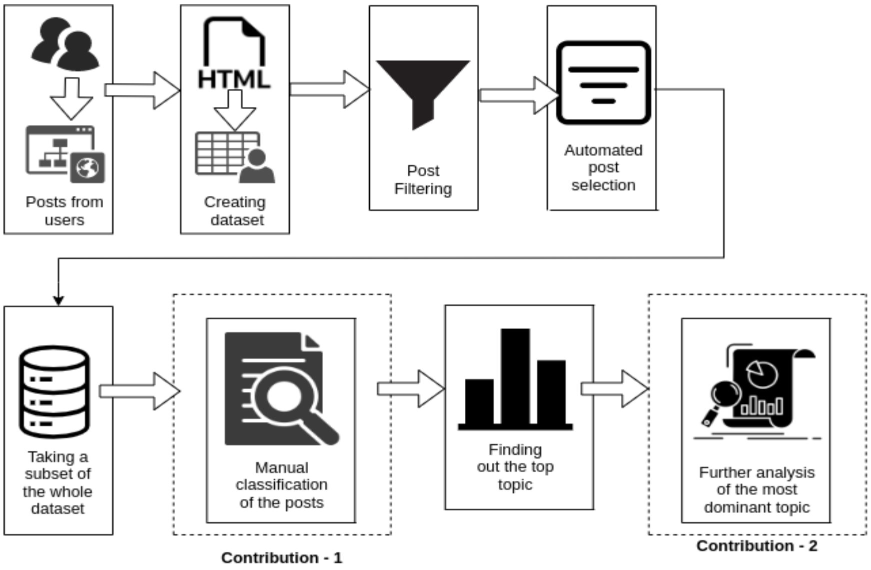

Posts from users
HTML
Creating dataset
Post Filtering
Automated post selection
Taking a subset of the whole dataset
Manual classification of the posts
Contribution - 1
Finding out the top topic
Further analysis of the most dominant topic
Contribution - 2

Figure 4.1: Flow chart of the proposed method

investigation drives to dig more into the most dominant topic-related posts for further analysis. Lastly, we want to provide some research implications and recommendations for the stakeholders of the web service. Specifically, our work raises the following research questions:

- **RQ1: Is there any possibility to categorize the topics discussed in the user forum?** We find that the users are facing problems in several topics. Thus categorizing them into some specific categories might come in handy. So we group the posts according to their subjects and posts in the same group are kept into a specific category.

- **RQ2: Which topic is the most dominant one among the other topics and can this topic be categorized again?** We want to find which topic is the most dominant one in our total group of topics. We focus to find out the reason why that topic is discussed more frequently and whether it can also be categorized into several groups.

- **RQ3: How can we help the Galaxy community through our analysis?** We provide some implications and recommendations for the Galaxy community. We expect that our study findings can solve the usability issues Galaxy have. Following the findings, we believe that other software systems can look for these usability issues in their system and take action accordingly.

From our study, we found some noteworthy research implications. Our study opens opportunities for the development community and the expert community to focus more on the classified problems. Extensive analysis of the most presiding topic brings out remarkable discoveries about the issues which might have been unknown to the development community. From our conclusion, the research community can explore more on this domain and recommend interesting findings too. Fig 4.1 illustrates the flow chart of the proposed method applied in our study.

4.2 Methodology

In this section, we describe the methodology of the study. We first describe the data extraction process, data preparation process and finally how we classified the posts according to similar topics.

4.2.1 Data Extraction

We collected data from the Galaxy support community hub. On this web page, users post their problems and concerns about the usage of the Galaxy workflow management system. The other users or the experts try to solve them with their knowledge base. The website contains almost 4000 posts. We tried to extract as much as we could by running a web scrapping script. Firstly, we downloaded support Galaxy's data dump in HTML format. We stored the user id of the posts, the post title and the post description for each of the users' posts. To remove the redundant information, we filtered the HTML tags through a python script. As we intended to store only the post-related information, we searched for the table where the link to all the posts was stored. From surfing through the links, we selected those tags which contained the user id, title and description of the posts. Therefore, we managed to get 3,032 posts at the time of creating the dataset.

4.2.2 Data Preparation

We created a CSV file from the dataset for manageable analysis. The description of the posts was converted into a single corpus and removed the common sentence at the beginning of each of the posts was. A total of 244 users posted in that forum and among them, 149 users posted single posts and 95 users posted multiple posts. We randomly selected 400 posts to perform the manual analysis.

4.2.3 Classification

Manual or qualitative analysis is genuinely important in supporting a researcher in producing a profound and extensive understanding of a given phenomenon [60]. Manual classification has been used in numerous research studies such as in medical science [72], text analysis [47], soil science [24], etc. We observed the importance of manual classification and carried out a similar process in our study too. In this study, each post is classified two times: at first, we tried to group the posts into similar categories as far as we can, and second, we made an abstract classification of those outlined categories. We tried to focus on the problem statement of each of the posts and kept them in a common category. When we found a similar kind of discussion on another post, we assigned the post to the previously labelled category. Otherwise, a new category has been created. After labelling all the posts, we searched for the smaller cluster of posts related to one category and merged them into a bigger category cluster. If the number of posts in a category is comparatively higher than in any other category, we split the group into multiple categories.

Below we present which types of posts were included in the specific categories and were classified them accordingly.

- *Tools issues*: Here the users talk about issues related to tools and their usages. How to use the tools, how to install new tools, and what to do when tools are showing various kinds of errors - these types of issues are discussed in these posts.

- *Galaxy account/server/installation issues*: Users raised questions about the problems they faced while installing the local Galaxy application, the problems appearing with the users' Galaxy accounts and also the issues in the Galaxy server. Posts related to the environment setup and maintaining users' requirements, were also included in this category.

- *Upload/input/output issues*: Here users post those issues where users failed to upload a dataset or the upload system did not work according to their expectations. The input of the data in a workflow, input criteria mismatching and confusing or non-explainable output results have been discussed in these classified posts.

- *Asking workflow/task-based suggestions*: Users described different task-based problems and asked for expert opinions. The most frequent suggestions they wanted were on the workflows and how to build an effective workflow to perform a certain experiment.

- *Job running issues*: The posts contain the problems users faced while running a job and their waiting time. Users experimented with different tools and datasets but they faced problems when the job was running for a long period of time or did not finish at all.

- *Miscellaneous*: In this section, we kept all those posts contained which are not similar to the other specified categories. Categorizing them into multiple labels would result in smaller clusters of posts to define as a different category. Users mainly talked here about the workflow-related issues, tutorial-related issues, specific experimental issues and so on.

We again analyzed the post percentage of all the categories and tried to figure out the most repeated topic presented in the posts. From that, we found that the most frequent topic posted in the forums is about the tools. So we further analyzed the posts related to tools and categorized them again into 7 categories. Listed below are the label names of the categories and the type of posts discussed according to the labels.

1. *Tool usage problem*: A huge number of posts we found, are about tool usage problems. Users faced inconsistent types of issues while working with the tools. Most of the posts were about the error users got while running a tool on a dataset and the experiment failed to execute because of the error. A big portion of the total posts is related to this issue. Users predominantly complained about the lack of explainability of a tool's error. Also, proper classification and description of these errors created more difficulties as users failed to seize the idea of where the main problem is taking place.

2. *Tool usage suggestion*: The next most repeated topic we found is asking suggestions of using a tool or which tool to use in a particular case or what kind of dataset is applicable to use a certain tool. Users find it challenging when they need to run a new experiment, they know about the process, the required inputs and the expected outputs but they fail to realize which tools are more appropriate to perform that experiment. An example of posts related to this topic is: " Hi all. I am trying to use GATK Mutect2 to do somatic variant calling. Not like VarScan Somatic, which allows the user to input normal and tumor bam files. Mutect2 only has one input field. I read through its example but still found nowhere to input the paired bam files. Can anyone help me on this issue? Thank you ahead"

3. *Missing tool information*: Galaxy provides the users with a great number of useful tools and also allows the users to develop new tools but many of the tools lack proper information about their usages. So users find it challenging when they are unknown of the working details of those tools and fail to make proper use of them. A post under this topic is asked by a user: "On my tool box I am not able to find either velveth or velvetg. Regarding velvet the only tool I can find is VelvetOptimiser Automatically optimize Velvet assemblies (Galaxy Version 2.2.6)."

4. *Tool update/new tool develop request*: Users asked about updating the tools or developing new tools from time to time in these posts. New emerging analysis of data and high throughput experiments need compatible tools for extensive investigation. If the tools are not up to date following the surging and more resourceful experiments, then users may find it tough to get the benefits from those tools. A user asked this following question related to this topic: "Hello,, I am wondering whether it would be possible to update the 'Trim Galore' to the latest version. It would be appreciated in advance., Regards"

5. *Tool unavailability*: Users have complained in some of the posts that they could not find the necessary tools to run specific analyses because of the unavailability of the tools. Users faced problems while looking for a certain tool following a tutorial or experts' recommendations but they failed to find the tool on the Galaxy platform or Galaxy toolshed[2]. A post under this topic is asked by a user: "On my tool box I am not able to find either velveth or velvetg. Regarding velvet the only tool I can find is VelvetOptimiser Automatically optimize Velvet assemblies (Galaxy Version 2.2.6).Can you please help with this."

6. *New tool developing issue*: Galaxy allows its users to develop new tools for the ever-growing demand for high throughput analysis so that the other users can keep themselves up-to-date with the newest tools. But while developing those tools, tool developers faced several issues to make them ready to install and use effectively. A post related to this topic is: "I am new to Galaxy. I developed a tool wherein I am storing the output (multiple files) of the tool after it finishes running into a directory which I create during run time, Now my main aim is, how can I get the name of the directory created

[2] https://toolshed.g2.bx.psu.edu/

during run time in my XML code. Also, how can I set the output of the tool in the XML file to point to the contents of the directory."

7. *Tool installation issue*: While installing the tools, users faced difficulties and could not complete the installation process as needed. These types of posts are kept in this category. A post under this topic is asked by a user: "I just set up a fresh galaxy installation, following the instructions on: https://galaxyproject.org/admin/get-galaxy/ 5 . But once I try to install a tool I get the following error:, Error cloning repository: [Errno 2] No such file or directory: 'hg': 'hg', It seems to always be the same error with every tool. Can someone here help me out?"

We developed these categories based on the context of the posts. For evaluation purposes, we contacted one experienced researcher who has two years of experience in the domain of SWfMSs. We selected 5% of the posts by random sampling from the previously labelled 400 posts and asked him to comment on the categorization that we furnished. We also provided him with the category list and a proper explanation of each of the categories.

At first, the researcher classified 70% of the provided posts according to our classification labels. We intended to make the classification more accurate and comprehensive. Thus, we went for a discussion with the researcher about which type of posts should be kept in each of the categories. We revised the posts again and provided another subset (5%) for his evaluation. This time, the researcher classified 90% of the posts according to our classification. We also calculated Cohen's Kappa [58] to measure the inter-rater agreement. This coefficient is used to calculate the inter-rater reliability for qualitative items. We got the score $K = 0.86$ i.e., near-perfect agreement. Therefore, the final abstract classification resulted in 6 categories and the most repeated topic was categorized into 7 categories.

4.3 Result

In this section, we try to answer the research questions by explaining our experiment results.

4.3.1 RQ1: Is there any possibility to categorize the topics discussed in the user forum?

Table 4.1 shows the post number and their percentage values for each of the categories. We can see from the table data that tool-related posts hold 44.25% of the total posts in the whole dataset. The less-numbered posts that appeared in the dataset are about jobs running issues while executing a workflow. Miscellaneous classified posts hold a significant percentage (20.75%) in the classification. The reason behind it is users face various complications in their specific domain of interest. Thus, it is not possible for any service provider to cover all the requirements according to the users' preferences. So multiple posts from various domains can occur while performing analysis.

Category Label	Post number	Percentage
Tools	177	44.25%
Galaxy account/server/installation issues	52	13.00%
Upload/input/output issues	36	9.00%
Asking suggestions	32	8.00%
Job running issues	20	5.00%
Miscellaneous	83	20.75%

Table 4.1: Abstract categorization of the posts

Answering RQ1: The Galaxy users ask about every possible topic they can face while using the Galaxy system. We can group them into several categories such as tool usage, galaxy account or server or installation issue, upload or input or output issue, suggestions, job running issue and miscellaneous issues. The most popular topic among them is tool usage problems.

4.3.2 RQ2: Which topic is the most dominant one among the other topics and can this topic be categorized again?

Fig 4.2 illustrates the classification of the tool-related posts as it was found to be the most frequently occurring topic users talked about. From the figure, we can see that more than half of the posts were submitted asking for tool usage errors. Missing information about the tools comes next. The rest of the categories clearly depict the importance level of each of the categories.

Answering RQ2: The Galaxy users are more concerned about the tool installation issues among the other problems. Users find the tool installation process very difficult because of the lack of proper explanation and guidance on how to install and use them.

4.3.3 RQ3: How can we help the Galaxy community through our analysis?

Our study raised some of the very important issues Galaxy SWfMS users are facing while using the service. Based on the findings, we can recommend the features which might be implemented on the platform for improving usability and also for the other existing SWfMSs. Here we describe our findings from the manual analysis:

- **For the developers:** Tool usage problem is the most discussed one in our findings and it is one of the most concerning facts too. We found a great number of posts where people are complaining about errors in tool applications. When users get errors at the time of performing a task, they get restless and try to find solutions immediately. Again, missing proper information about tools becomes a hindrance in that pathway. Also, users complained about the unavailability of some tools in the toolbar and in

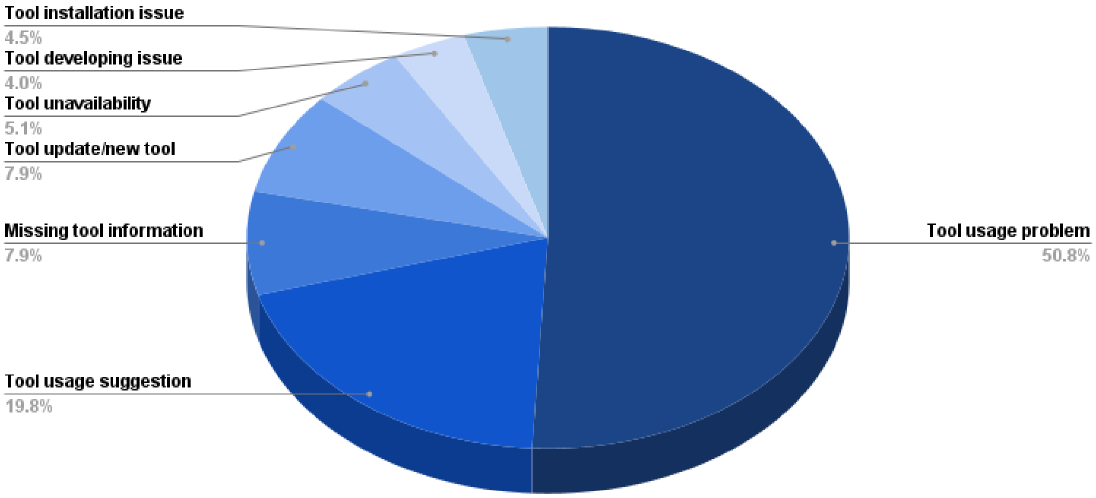


Figure 4.2: Categorization of the 'Tool usage' related posts

the tool repository. In addition, some tools being old versioned and not compatible with the recent implementations of the experimental analysis created problems too. Moreover, some of the posts show that tool development and maintenance of them became a demanding issue for providing better services to the users. From our implication, we can recommend to the developer community that they should investigate more on these problems and develop the tools according to the users' requirements. The other SWfMSs developers can utilize our study results and focus on improving their tools services to meet the users' requirements.

- **For other users:** Our findings implied that only a few of the Galaxy users post about their problems on the forum. So there are still a large number of other users who might have faced similar in the same categories or divergent problems which have not been categorized yet, but they did not reach out to seek help from the forum. Those users can try to explore the other platforms as well for getting their jobs done. When users are not satisfied with the usability of the service, they lose attraction and interest to work even on the most popular platform. So the Galaxy site owner and development community can take the necessary initiatives so to build a more appropriate user-friendly environment with proper resources and defined instructions for using them.

- **For researchers:** The research community can be greatly benefited from our findings. The categorization made it easy for them to focus on specific topics to look at. They can go for a deeper analysis of a certain category and try to find out additional useful information about the users' problems. They can suggest effective ways to resolve those issues and provide significant research findings on increasing the usability of the site as well as the other SWfMSs.

- **For active users:** Our proposed category can be beneficial to the active users too. When they will have the proper categories of the posts, they could easily find out the section to post for a specific topic. If a user face problem with a specific issue, they can directly go to the particular category and surf through the posts to get their answer. The expert community will find it more efficient when they want to help the users having problems in their particular area of expertise.

Answering RQ3: We find that our proposed categories find out the usability problems Galaxy system has and solving those issues would definitely benefit the stakeholders in long run.

4.4 Threats To Validity

The main concern of our manual analysis is that we did not explore the other existing SWfMSs community support hubs. Therefore, it is quite possible that we might not have covered all the topics and some major points might have been missed from our analysis.

4.5 Conclusion

Our main focus, in this study, was to analyze the posts Galaxy users shared on the forum and also find out meaningful ways to improve the usability of the other existing software systems. When services from a website is popular with users' satisfaction, it will draw attention to more users who might have explored the other available options. Our initial analysis showed the type of problems users face in working on this platform are very noteworthy. While we performed further analysis, we found out the most repeated topic users discussed was about the tools and their usage. We also evaluated our analysis with one experienced researcher and his opinion clarifies that the categorization is appropriate and suitable for further analysis. We believe that exploring user forums for other software systems will definitely bring out significant information about users' dissatisfaction, requirements and improvements they want to see in that particular platform.

5 Ensuring proper tags for the users to improve the usability of the forum

In this chapter, we discuss our second study which proposes a method for automatically suggesting tags for the users' posts. In Section 5.3, we explain the methodology of this study that we follow. Section 5.4 shows the result we get from our study. Section 5.5 discusses the threats to the validity of this study and Section 5.6 summarizes the chapter.

5.1 Introduction

Chapter 4 shows that users post their problems on the Galaxy user forum to get recommendations, suggestions and the solution to the problem they faced while conducting their experiences. But we found that most of the tags of the posts do not correspond to the context or users do not even use any tags at all. What happens, in this case, is that not every user likes to post their problems and they go for searching posts related to their problems. But they will not find the appropriate posts that will help them in solving their problems. Let us consider these two examples. In the first case, *user1* is searching for a post using particular tags. But they are getting some of the answers matching the searched tags. Also, they find that the context of the posts does not match the tags. So *user1* is now confused and also irritated as they are not getting the appropriate answer that they are looking for. In another case, *user1* posted a question but did not use any tags. They got some good and effective answers which can be of use to other users. Now, *user2* is looking for posts which are similar to user1's. But as *user1* has not included any tags in the searched result, that post is not going to appear. Thus *user1* looked for all the other questions but did not get any appropriate answers they were looking for. But the answer in *user1*'s post might come in handy for *user2*. As *user1* has not included any tags, so *user2* is not getting the answer they are looking for. Thus usability issue raises here. To overcome this problem, in this thesis, we propose a method to automatically suggest appropriate tags to the users so that users can find the posts according to the context of the posts and users who are not sure which tags to use, they can get some recommendations too.

Several studies found the importance of automatic tag suggestion for users' posts [98, 4, 89]. Roy et al. [90] described in their studies the reason behind why most of the questions do not receive any answer and the absence of proper tags is one of the key reasons for that. As the tagging process is manual and inconsistent by nature, so they designed an automatic tag suggestion technique for Community Question Answering (CQA) sites that could suggest tags to the users based on their posts. Al-Attar et al. [4] found the importance of implementing domain-specific automatic tag-suggesting systems for Arabic tags. It is well known to the researchers that domain-specific software systems suffer from usability problems because normal approaches are not always applicable to those systems targeted for specific users. Along with that, suggesting tags based on the previous content is very popular in recent studies. Wu et al. [106] proposed a generative model for recommending tags on textual content. They observed that the tags and their similar tags most of the time appeared in the corresponding contents. Galaxy, being a very domain specific software system, its tagging system is also manual and not consistent in nature. Moreover, most of the posts contain error descriptions so it is quite difficult to fit this type of data into the normal pre-trained word vocabulary. But previously posted questions can give us insights on which tags should be used and users can always evaluate them before using them.

Thus, to address the problem, we consider getting similar questions to the new question a user is going to post. If the tags used in similar questions are provided to the user, then they can get some insights about which of the appropriate tags they need to use. We apply cosine similarity to the dataset and find the related questions for the newly posted question. To train the method, we used our previous dataset extracted from the Galaxy user forum and we find the cosine similarity with any new questions posted after the extraction date. To evaluate our proposed method, we perform a user study to see the relevance of the tags suggested by our method compared to the originally posted tags. We also look for the relevance rate of the tags where users did not consider putting tags at all.

5.2 Research Questions

- **RQ1: Do the tags represent the context of the posts?** To answer this research question, we did our investigation on the Galaxy user forum and found a relationship of the posts where tags are not used and where improper tags have been used. We try to understand the problems or issues users are facing when tags do not represent the context of the posts.

- **RQ2: How relevant our suggested tags are compared to the tags that were originally posted?** Here, we evaluate our proposed method by conducting a research study. We provided ten posted questions where proper tags have not been used or the posts contained no tags at all. We provided our participants with our model's suggested tags and asked them to rate the relevance level. Our findings show that our method's suggested tags are more relevant than the already posted tags.

5.3 Research Methodology

Our previous study (Chapter 4) finds that the posts in the user forum are not categorized accordingly and many of the posts did not include appropriate tags. Thus we decided to propose a method to suggest tags automatically to represent the context of the post and also recommend tags to those posts with no tags.

5.3.1 Experimental setup

Data preprocessing

We used the dataset that was previously used in Chapter 4. So it was necessary to remove the HTML tags again. Then we started preprocessing steps by converting the texts into lowercase data. In natural texts, the uppercase and lowercase letters do not hold significant differences from the posts. Next, numbers and punctuation have been removed from the post texts. Numbers and punctuation are both necessary to understand the context of the questions, but in our study, to suggest the tags, these two features do not play any role particularly. There remains a chance that the posts' texts have some unnecessary white spaces. So we remove the white spaces from the text. Next, tokenization is performed. In this stage, the text is broken down into smaller "tokens". They can be words, sentences, characters or paragraphs. But in our study, we used word tokenization as we extracted one word at a time from the text. Next, the removal of stop words is performed. These are the words that do not add up any significant information to the system but raise the possibility of skewing the analysis procedure. In this English language, examples of stop words are "the", "an", "and" etc. If the stop words are not removed from the text, there might be a chance for those words to be treated as a feature of the model because of the high frequency. In the last phase of the preprocessing steps, we performed stemming and lemmatization to finish the procedure. Stemming is the procedure of converting a word into its stem. For example, "writing" and "wrote" are the same two different tenses for the same word, which actually points to the action "write". Stemming does this process by reducing the text words into its step. We used a library called porter-stemmer which identifies and then removes the suffix and affix of a word. Next, we used WordNetLemmatizer to make lemmatization of a word. It is a process to convert the word to its root synonym. This process ensures that the reduced word is a dictionary word.

Training and word embedding

For training the dataset, we use a pre-trained model from tensorflow hub [1]. It is a repository containing trained machine learning models ready for fine-tuning and deployable anywhere. In our method, we used a token-based text embedding trained on English Google News 200B corpus. We created embedding both for the training data and the test queries. After that, a recommendation function has been used to recommend the similar questions to the users and their used tags also. In the recommendation function, the new question has been passed as the parameter. The function takes the question and preprocess it. Then the preprocessed question is compared with the dataframe and with the help of cosine similarity value, the recommendation function returns the matched questions posted before. A threshold value (0.5) has been used to identify

[1] www.tensorflow.org/hub

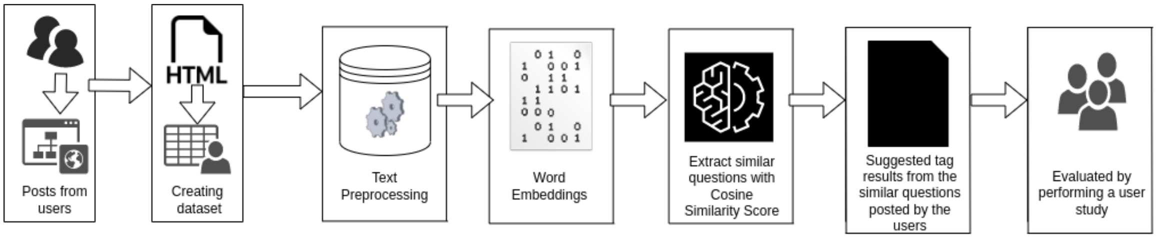
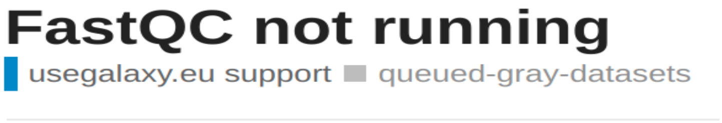

Posts from users
HTML
Creating dataset
Text Preprocessing
Word Embeddings
Extract similar questions with Cosine Similarity Score
Suggested tag results from the similar questions posted by the users
Evaluated by performing a user study

Figure 5.1: Flow chart of the proposed method for suggesting tags



FastQC not running

usegalaxy.eu support queued-gray-datasets

Figure 5.2: Tags not representing the context of the post, e.g., usegalaxy.eu support and queued-gray-datasets.

the most relevant posts according to the new question. The tags of the matched questions are taken as our suggested tags. Fig 5.1 shows the methodology of this study.

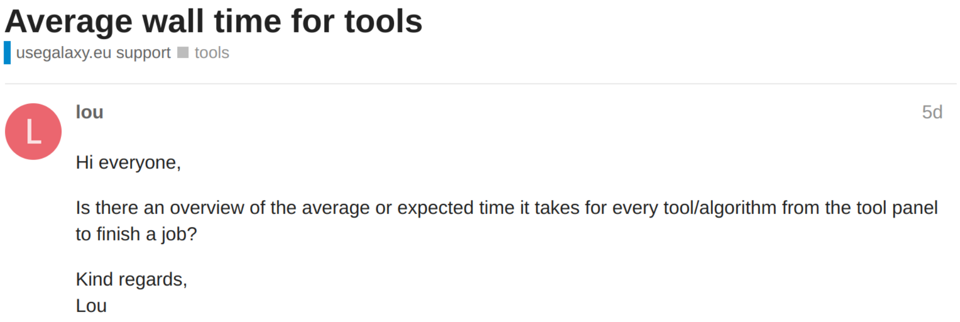
5.4 Result and Analysis

In this section, we are going to explain our result from the experiment and show the analysis based on it. We present the answers to the research questions too.

5.4.1 RQ1: Do the tags represent the context of the posts?

While doing our manual analysis in chapter 4, we found several posts where tags are not appropriate or are not representing the context properly. Let us view some of the examples from the forum. In Fig 5.2, we can see that a user has posted a question stating that their FastQC tool is not running. They posted screenshots of the error and asked for help. But the tags they used do not go with the context of their posts. *Queued-gray-dataset* does not align with the post. In another case, like shown in Fig 5.3, 5.4, 5.5 and 5.6, we can see that the tags used here are not properly representing the topics of the posts. For Fig 5.3, the post is about the average wait time for tools in general. But the user used only *tools* tag. So whenever another user is going to search using the *tools* tag, they will get this post as a result too. Similar in Fig 5.4 the user used *input* as one of the tags but the post says nothing about the inputs. Again in Fig 5.5 and Fig 5.6, in both of the cases, only using *error* tag does not properly present the context of the posts. These two posts have replies and solved answers based on the question. But if the users used more specific tags, then the other users might get these posts easily and promptly. But using only the *error* tag, these posts might get lost in the loads of other posts where *error* tag has been used too.

In addition, we try to find the posts where tags have not been used at all. We find 859 of the total posts



Average wall time for tools

usegalaxy.eu support ▪ tools

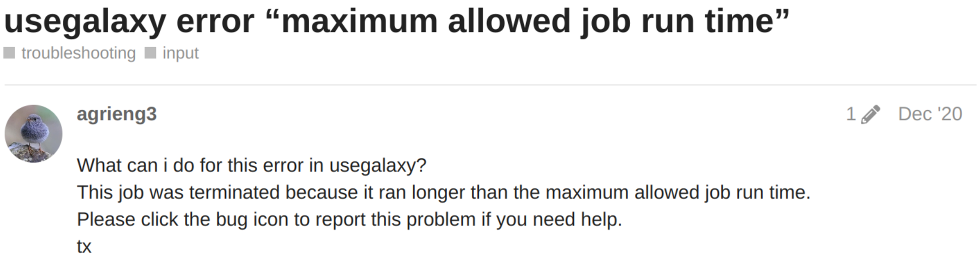
L **lou** 5d

Hi everyone,

Is there an overview of the average or expected time it takes for every tool/algorithm from the tool panel to finish a job?

Kind regards,
Lou

Figure 5.3: Not appropriate tags: Example:01



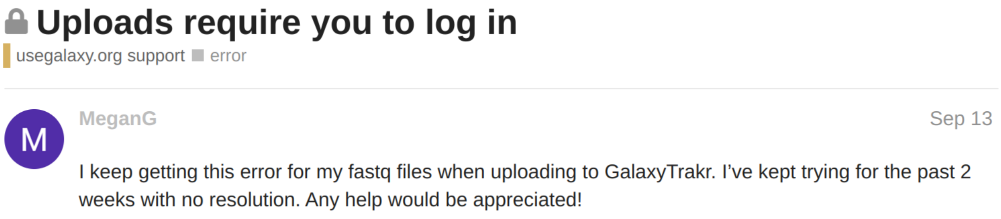
usegalaxy error “maximum allowed job run time”

▪ troubleshooting ▪ input

agrieng3 1 ✎ Dec '20

What can i do for this error in usegalaxy?
This job was terminated because it ran longer than the maximum allowed job run time.
Please click the bug icon to report this problem if you need help.
tx

Figure 5.4: Not appropriate tags: Example:02



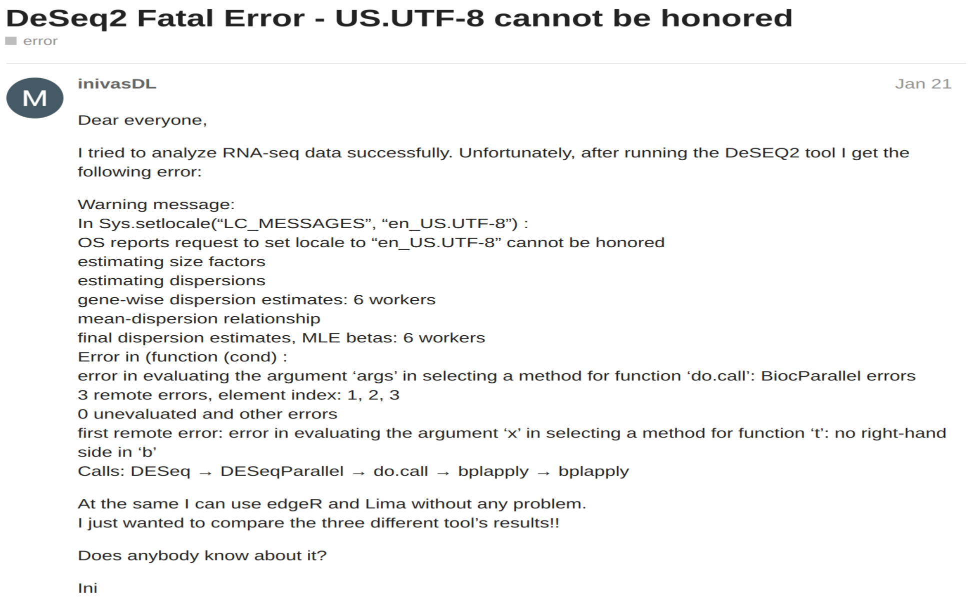
🔒 Uploads require you to log in

usegalaxy.org support ▪ error

M **MeganG** Sep 13

I keep getting this error for my fastq files when uploading to GalaxyTrakr. I’ve kept trying for the past 2 weeks with no resolution. Any help would be appreciated!

Figure 5.5: Not appropriate tags: Example:03



DeSeq2 Fatal Error - US.UTF-8 cannot be honored

error

inivasDL Jan 21

Dear everyone,

I tried to analyze RNA-seq data successfully. Unfortunately, after running the DeSEQ2 tool I get the following error:

```
Warning message:
In Sys.setlocale("LC_MESSAGES", "en_US.UTF-8") :
OS reports request to set locale to "en_US.UTF-8" cannot be honored
estimating size factors
estimating dispersions
gene-wise dispersion estimates: 6 workers
mean-dispersion relationship
final dispersion estimates, MLE betas: 6 workers
Error in (function (cond) :
error in evaluating the argument 'args' in selecting a method for function 'do.call': BiocParallel errors
3 remote errors, element index: 1, 2, 3
0 unevaluated and other errors
first remote error: error in evaluating the argument 'x' in selecting a method for function 't': no right-hand
side in 'b'
Calls: DESeq → DESeqParallel → do.call → bplapply → bplapply
```

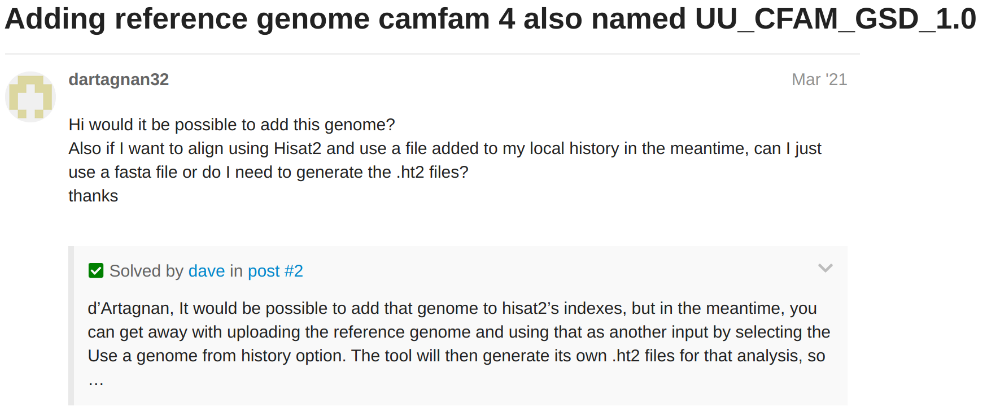
At the same I can use edgeR and Lima without any problem.
I just wanted to compare the three different tool's results!!

Does anybody know about it?

Ini

Figure 5.6: Not appropriate tags: Example:04

such as that. We again analyzed those posts and we found that because of the absence of the tags, those posts received fewer replies compared to the posts with tags. Moreover, in some of the questions, there are solved answers but as those posts do not fall into any groups of the tags, it is difficult to find the question for other users. For example, in Fig 5.7, this post does not contain any tags. It got a solved answer but when we try to find this post using the keywords used in this post, we did not find the post. So it is expected that posts with no tags do not come helpful for other users as these post cannot be found later or at a quick attempt.



Adding reference genome camfam 4 also named UU_CFAM_GSD_1.0

dartagnan32 Mar '21

Hi would it be possible to add this genome?
Also if I want to align using Hisat2 and use a file added to my local history in the meantime, can I just use a fasta file or do I need to generate the .ht2 files?
thanks

✓ Solved by dave in post #2

d'Artagnan, It would be possible to add that genome to hisat2's indexes, but in the meantime, you can get away with uploading the reference genome and using that as another input by selecting the Use a genome from history option. The tool will then generate its own .ht2 files for that analysis, so …

Figure 5.7: Solved post with no tags

Answering RQ1: We can see that the tags used in the user forum posts do not always represent the topic or the context of the posts and are even inappropriate in some contexts. Also, the posts with no tags cannot be found easily later and sometimes they get lost when searched with keywords.

5.4.2 RQ2: How relevant our suggested tags are compared to the tags that were originally posted?

To find the relevance of our suggested tags, we tried to apply some evaluation methods and metrics such as accuracy, precision, and recall. But the result we obtained from those experiments was very skewed. We investigated the reason behind this result and our finding indicates some significant information about the result. As the extracted data is very less in number and we are training our dataset with a huge pre-trained model, there might be a risk that preprocessing of the dataset does not reflect the real meaning of the corpus and thus the obtained results are skewed. Thus, to evaluate our model, we performed a user study. In the following sections, we describe the study methodology.

Survey design

We designed the survey as follows.

- We ask the participant to give a rating on the relevance level. Our survey included a total 10 questions (Appendix A) that were posted on the forum. We collected the tags used in those posts.
- In the survey, we provided the original tags and also the tags suggested by our proposed method.
- We then asked the users to rate the relevance level of the suggested tags on the basis of the context of the questions.
- Among the 10 questions, 3 of the questions were without tags in the forum. So we asked the participants to investigate whether the suggested tags are appropriate to the topic of the posts.

We explained to the participants about our research goal and ensured that their information will be treated confidentially. We informed the participants that the survey would take a total of 10 to 15 minutes to complete. We asked the participants to give their confirmation whether they consent to proceed with the survey.

We used Google forms to collect the users' opinions. We sent out the survey forms to the participants by emailing and contacting them directly. Participants took part in the survey by clicking on a link attached to the emails or by the links we shared. As mentioned earlier, the participants were provided with 10 questions with originally posted tags and suggested tags by our method. We asked the users to rank the relevance of the tags to the questions on a scale of 1 to 5. Here 5 indicates the suggested tags being most relevant and 1 being not relevant at all. The survey also included opinion-based open-ended text boxes for the participants to share their opinions.

5.4.3 Demography of Participants

Next, we tried to collect the demographic information of the participants. For example, the age of the participants, their gender, their years of experience and so on. We had 12 participants who participated in our study. Among the participants, eight of them were male and four of them identified as female. The mean age of the participants was 27 years. Each of the participants had a minimum of 2 years to 18 years of experience in software engineering research and six of them have experience in working with SWfMSs and two of them are now doing research in the field of bioinformatics. We provided our survey to the lab members and also to the internet community. We asked for their time to fill out the survey. We made the survey open for any users who have a similar domain can participate and give their feedback. We received very little response from the other community. We did not promise any payments for the participants to fill out the survey.

Survey result analysis

The survey result is depicted in Fig 5.8. The average score provided by the participants for each of the questions is illustrated here. We can see that, among the 10 questions, 5 of them got an average score greater than 4.5. We asked the participants behind the reasons for which they gave more relevance scores to these questions. They pointed out several tags from our proposed method and explained why the suggested tags are better representing the context of the posts. For example, in 5.1, the first question was asked decribing the problem a user had faced while installing the Galaxy on their local machine. But they were mentioning the issue related to the operating system. There, the user has used only two tags. But our method suggested some more relevant tags. The model has suggested the 'ubuntu' 'conda' and 'python3' tags too so users facing the same problem can find this post if they use any of the previously mentioned tags. Then, for the rest of the questions, 4 of them got an average score greater than 3. When we asked the participants about their reasoning for giving this score, they said in these cases, both the posted tags and the suggested tags are relevant. The last question, got the least score. We investigated the one question which got less relevant scores from the users. The question was about text manipulation, which is not directly related to the usual problems that are posted on the user forum. So this might be the reason that our proposed method could not find similarity with other questions from the training set. That is why it could not suggest relevant tags.

Table 5.1 shows three of the questions that were presented in the survey for the participants to evaluate the suggested tags. The table also includes the posted tags and suggested tags from the proposed method. We can see from the table that for the first two questions, our model actually suggested more relevant and appropriate tags than the posted ones. For the third question, the user did not provide any tags, but our method can suggest tags for that question too.

Answering RQ2: Based on the user study, we can answer the second research question - how relevant our suggested tags are compared to the tags that were originally posted. We saw that users' posted tags are not always relevant to the posted questions and they do not represent the actual topics many of the times. With the help of our method, we can suggest to the users more relevant and appropriate tags for their posted queries. Also, for the questions where users are not sure of which tags should be used, our method can suggest proper tags for representing the topic properly.

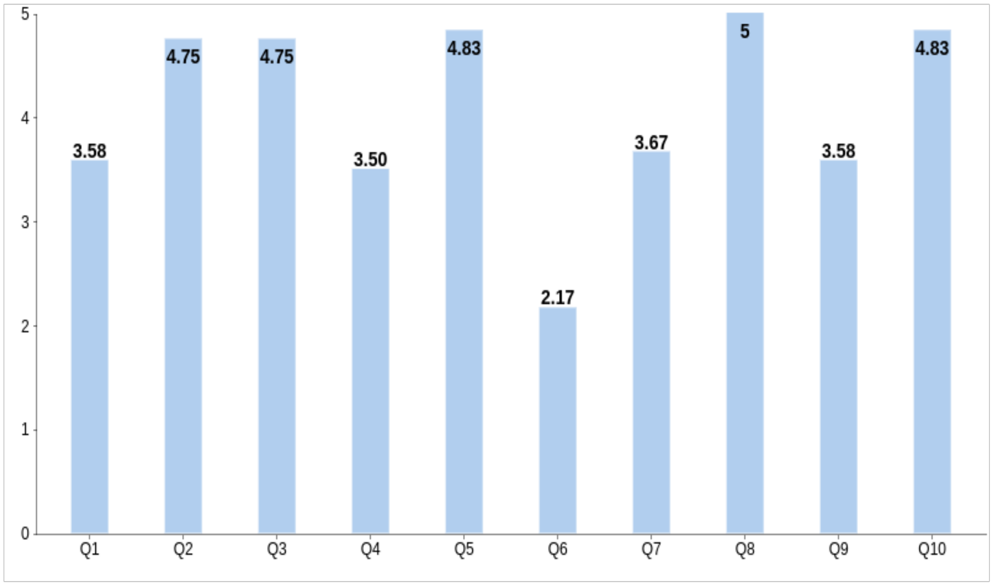


Figure 5.8: The result of user study

Table 5.1: Comparison of posted and suggested tags

Question asked	Posted Tags	Suggested Tags
"Hi everyone I am trying to install locally galaxy, I am using conda 4.14.0 with ubuntu (20.04.4 LTS). The installation seems correct after running ./run.sh it seems everything went on fine, and after a while the installation freeze at "solving enviroment", i left it like that for hours and nothing happened. I saw in other post that this problem was fixed using virtualenv: I installed virtualenv, desactivate conda, activate virtualenv, and the run the ./run.sh and I get this error.	'admin', 'galaxy-local'	'tool-install', 'galaxy-local', 'ubuntu', 'conda', 'error', 'python3'
Hello, I am trying to analyze a ChIPseq experiment for the first time. I have a collection of ten fastq files, and I queued a series of jobs on this collection: (1) FastQC, (2) Trimmomatic, (3) FastQC on the Trimmomatic output, and (4) BWA on the Trimmomatic output. The first FastQC job stalled with only 8/10 files completed. The Trimmomatic and subsequent FastQC jobs completed. The BWA appears stalled on the first 2/10 files. These have been stalled since Friday. Wondering if this is a typical run time and I should keep waiting, or if this is a sign of a problem? Very new to Galaxy. Thank you!	'queued-gray-datasets', 'server-side-delay'	'jobs', 'workflow', 'public-galaxy-server', 'server-side-delay', 'queued-gray-datasets', 'dataset', 'fastq', 'bowtie2', 'fastq-format-error'
Hello, I am doing an analysis of previously published ChIP-seq data for a bioinformatics class that I am taking. When I try to use the plotFingerprint tool to compare the IP strength between two samples (treatment/control) it is giving me the following message: "An error occurred with this dataset: format png database mm10. Job failed." I'm not sure how to correct this issue so if anyone has any suggestions that would be great. Thanks,Anthony	*No tags*	'error', 'toubleshooting', 'workflow'

5.5 Threats to validity

Threats to internal validity involve the errors and biases that originated from the experimental setup. Another threat to internal validity relates to the design of the study. We sent out our study to two of the experienced users who had experience in using the SWfMSs. After getting their feedback, we modified our study accordingly and then sent it out to the participants.

The threat to external validity relates to the experience level of the participants. Our participants have experience ranging from novice to experienced and constitute mostly of researchers from the software research field. But the questions were chosen in such a way that the participants could easily understand the context and answered confidently. We provided guidelines and asked them to clear out any confusion if they have to understand the questions.

5.6 Conclusion

Users post their queries in the forum but not all users like to post on the forum. They want to look for similar problems and get solutions from it. But if the posts are not tagged appropriately or even missing tags, then it gets difficult for other users to find out the appropriate post they were looking for. To solve this problem, we propose a method to automatically suggest tags to the users based on the previously asked similar questions. We provide them with the tags that were used in the similar questions posted before and from the tag suggestion, users can choose any tags they think are most relevant. For post without any tags, our proposed method suggests relevant tags too. We evaluate our method with a user study where participants were told to rate the relevance level between the originally posted tags and the suggested tags. We find that 50% of the suggested tags received the highest relevance score and 40% of them received a similar relevance score to the originally posted tags. Thus our proposed method can be effective to group the posts under the same tag for the users to find them in a very short period of time.

6 Workflow recommendation from natural language queries

6.1 Introduction

While we were exploring the usability problems in 4, we not only came across the tagging issue (5) of the post, but also we found out that users have a hard time while designing a workflow in different scenarios. One of the major difficulties a user can face while designing a workflow is to generate the idea of effectively designing it. A user can think of performing a task but to shape his idea into the design can be a challenge for the novice or less expert users. All the researchers who use SWfMSs to conduct their scientific experiments, they understand by concept what they need to do and what they need to do next, but they often get confused or do not get many ideas of how to create a solid implementation of their idea. As different users have different requirements, so they cannot follow one single implementation and complete their tasks by following it. The Galaxy users can get many solutions to implement their ideas into workable workflows, but many times we saw posts where the users ask questions about how to transform their ideas into workflows, for example,

"May I ask a question to create the reference from my own WGS data using Galaxy. I have my WGS data that was already mapped with the native reference (BAM file). Then I want to use this kind of data as a reference for further analysis with another dataset to identify SNPs"[1],

"We have WGS PacBio sequenced data of environmental samples. As the data is huge, we don't have any high-spec machine to analyze it. Therefore, we thought we should try the Galaxy platform. Is there any specific tool in Galaxy that can do the assembly of PacBio WGS metagenomic reads? Or any other online tool / pipeline that may do it?"[2] etc.

Some of the posts do not often get replies or the answers are not a very effective one. This creates frustrations and lowers the confidence of the users. As the SWfMSs are very domain specific and there are not a lot number of forums exist, so users either have to find out their solution by themselves or they have to completely rely on the platform to find an answer. We estimate that this type of need is more likely to be continued in future too, as the experiments are getting complicated day by day and developing new tools are getting popularity too [79],[13],[34], so users find it difficult to match their ideas completely with others.

Recent studies show that the popularity of natural language processing tools is growing at a faster rate

[1] *https://help.galaxyproject.org/t/create-reference-genome-from-my-wgs-data-custom-reference/1051*

[2] *https://help.galaxyproject.org/t/metagenomics-analysis-on-pacbio-sequenced-data/8808*

[22, 46, 75]. In [75], the authors built a tool to automatically extract temporal information for natural language understanding. The tool performs two basic tasks, one is to understand explicitly mentioned time expressions and the other one is to understand the temporal information which is conveyed implicitly via relations. Another tool is developed in [76] to facilitate the phenotyping of cognitive status in medical science. Similar kind of natural language processing-based tool was also built in [29] to monitor suicidal ideation smartly. Several studies shown the development of tools by processing natural language such as for searching research projects [23], for composing phrase [3], for detection of late life depression [27] and so on. So we get the idea of building a tool by processing natural language which can recommend workflows according to the users' requirements.

In this study, we develop and integrate a tool on the Galaxy platform for natural language to workflow recommendation (NL2WF). We apply Fuzzy matching technique as the background algorithm of our developed tool. To evaluate the tool we carry out a human study with 15 participants who were assigned to use the tool and provide their opinion to us. We also asked for their recommendations to modify our tool according to their requirements.

6.2 Overview of our study

The goal of our study is to explore to what extent and in what ways natural language to workflow recommendation can be useful to the users and researchers of SWfMSs. Our aim is to evaluate the tool with human opinions. Given the importance, first, we built a tool for Galaxy platform for users to put their queries in natural language and the tool can supply proper and relative workflows to them. Those provided workflow files can be directly imported to the local or main server of Galaxy. Second, we recruit 15 participants with diverse experience in conducting scientific experiments on the SWfMSs. Finally, we analyze the data to answer three research questions as follows.

- **RQ1: Do the users feel the necessity of having a system that can recommend workflows according to their requirements?** In this research question, we intend to find the reason why users are not getting good recommendations of workflows and whether there exists a necessity to have such a system or platform for the users to get a proper recommendation of workflows.

- **RQ2: How does the NL2WF tool differ from the normal search in the repository?** Here, we recognize the traditional result provided by the search option and try to find out the difference between that result and the result we get from our developed tool. We find that the tool's generated output is more efficient than the traditionally achieved output by the search option.

- **RQ3: How do users assess the usefulness and effectiveness of the NL2WF tool?** Here, we perform a user study to assess the usefulness and effectiveness of the tool. We take two scales, i.e., Likert Scale and NASA-TLX, into consideration to measure the users' opinions. The result from the

Likert scale shows that the participants take the idea of the tool in a very positive way and most of the participants are very satisfied with the tool. From the NASA-TLX scale, we find the task load score of the tool is very low which implies that the participants could easily perform and get their desired answers with less effort.

6.3 NL2WF Tool Design

We designed and built a tool, NL2WF (natural language to workflow) for Galaxy platform. The tool takes an English input query from the user and gives a list of five workflows containing their names, annotations of each workflow and the ready-to-import file. In this way, the human-computer interaction can be ensured as a natural task to study participants.

6.3.1 The underlying structure of Workflow Recommendation System

For recommending the workflow, we took the help of the Fuzzy matching natural language processing technique (Section 2.5.3). The steps are as follows.

1. We generated a new dataset for this study. We found that Galaxy has a shared workflow repository where Galaxy users provide their workflows for other users to import and use. So we collected our data from that repository. We collected the workflow title, workflow annotation and the workflow file (with .ga extension) for each shared workflow.

2. We used the *fuzzywuzzy* NLP library[3] for the purpose of matching the strings between queries and the dataset. The method calculates the similarity index by providing a score out of 100. As mentioned in 2.5.3, this method uses *Levenshtein Distance* to find the differences between the sequences. We imported the *fuzz* and *process* functions from the *fuzzywuzzy* library.

3. We again preprocessed our data by following similar steps from 5.3.1. We imported the *NLTK* python library which is one of the leading platforms to build python programs for working with human data. It supplies very effective and useful interfaces over 50 corpora and lexical resources such as WordNet, along with various text processing libraries for classification, tokenization, stemming, tagging, parsing and semantic reasoning, wrappers for industrial-strength NLP libraries [4]. We used the *WhitespaceTokenizer()* function to perform tokenization from the *NLTK* library. The *WhitespaceTokenizer()* is dedicated to extract tokens from strings without whitespaces, new lines and tabs [5].

4. Next, we intended to lemmatize the query and dataset strings by using the WordNetLemmatizer again[6]. Then we removed the stop words from both the queries and the dataset strings too.

[3] https://github.com/seatgeek/fuzzywuzzy
[4] https://www.nltk.org/
[5] https://www.nltk.org/api/nltk.tokenize.html
[6] https://www.nltk.org/_modules/nltk/stem/wordnet.html

5. Finally, we implemented a dedicated function to calculate the *token_set_ratio* between each of the strings of the dataset and the query that the user provided. *token_set_ratio* is one of the very useful module of *fuzzywuzzy* library. It performs a set operation that considers the common tokens between two string sequences instead of just tokenizing the strings. It does not consider the repeated words and for this reason, we get a good score while computing the set ratio. So, when we got the score of each of the dataset instances and the query, we performed a sorting operation and we saved the highest top five instances as the result.

6.3.2 How the tool works

1. As we mentioned earlier, the tool takes a natural language query input from the user and gives a list of workflows as a result. Our tool takes the position in the tool panel under the category of *MyTools* (Fig 6.1). This tool is developed in the local server, which is why we separated the category from the other ones.

2. Once the tool is selected, an interface with a text-box is going to appear. The user can put his query in the text-box and select the *Run Tool* button. Fig 6.2 shows the screenshot of the interface for the users to write their queries in the text box.

3. After the user has provided his query and run the tool, they can see the job running in the history panel. Then they can check how the tool is working and also can check whether it gives an error or not.

4. Once the job is finished successfully, then they can go to the provided link to a web page and there they can see the result of the tool. Fig 6.3, 6.4 and 6.5 show the demonstration of the tool for the earlier described procedures.

5. When the user is provided with the result, they can check for the workflow name, and annotation according to their query. They can select the best one that matches their requirements and download the workflow file by following the provided link. The workflow files are ready to import to any local or global server of Galaxy. The user can get their ideas by viewing those recommended workflows and designing their own accordingly.

6.4 Result and Analysis

6.4.1 RQ1: Why is there a necessity to have a tool for recommending workflows to the users?

We study the user forum in a rigorous way to find out the necessity of having a tool that can recommend workflows to the users. We found several posts where people asked questions about designing a workflow

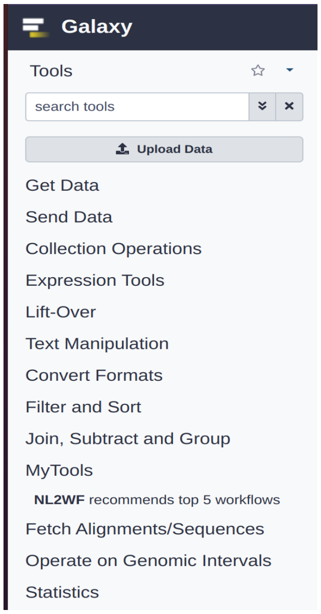
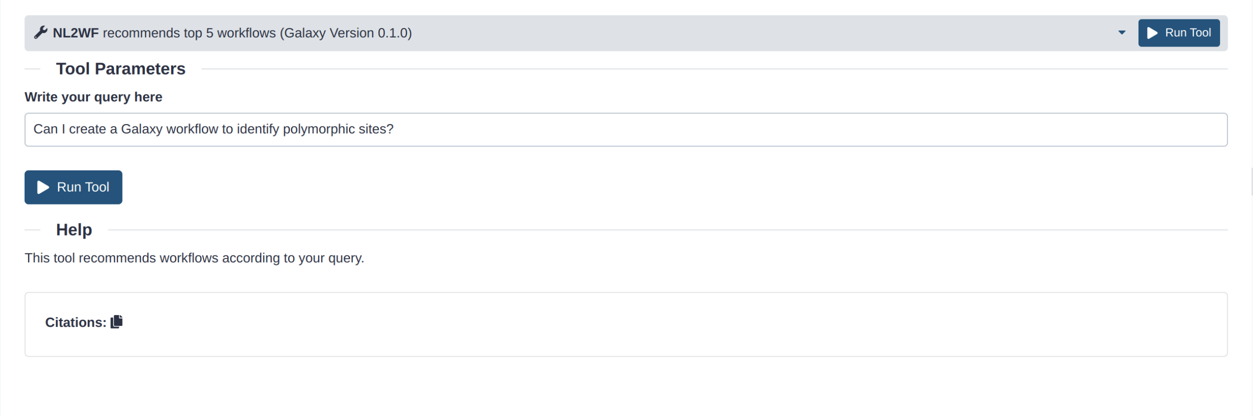


Figure 6.1: Screenshot of the position of our developed tool in the Galaxy platform



Write your query here: Can I create a Galaxy workflow to identify polymorphic sites?

This tool recommends workflows according to your query.

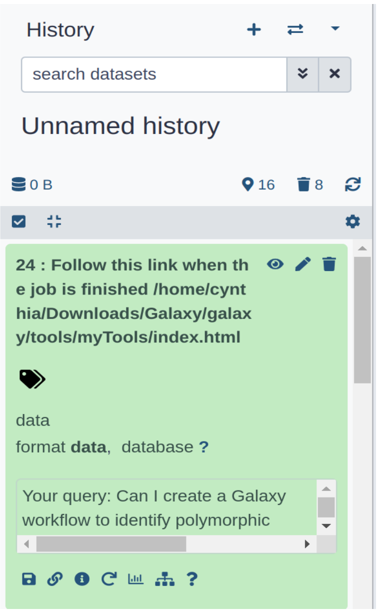
Figure 6.2: Screenshot of the interface for providing query to the NL2WF tool

and also asked for recommendations on how to implement their ideas. One of the users of the forum asked "How to calculate the differential expression without replicate data?" He tagged the post within 'workflow' groups. Some other posts that we found on the forum implies the necessity of having a platform for getting workflow-related suggestions. For example,

"I would like to perform paired-end RNA seq analysis for 21 samples with all of them having a biological replicate. I would like to know how should build a dataset list for such a cohort. Should one data list have all R1 and the other datalist have all R2 to be able to run commands on all sample at once? And how should I account the biological replicates"[7]

"Silly, simple question, but I cannot find a good solution. I have 2 VCF files being produced in workflow,

[7] https://help.galaxyproject.org/t/paired-end-rna-seq-with-biological-replicates/7434



Unnamed history

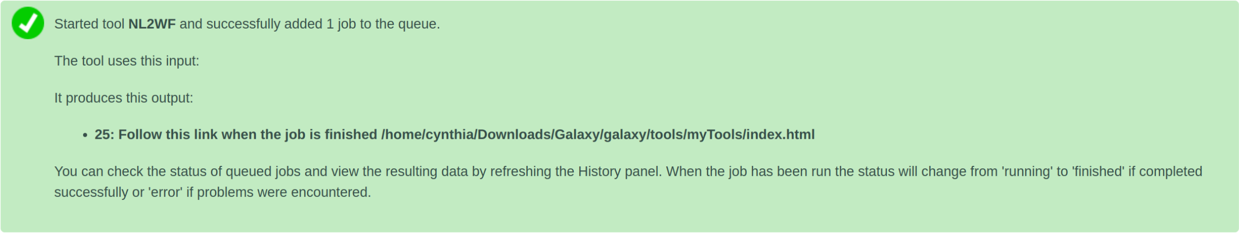
0 B 16 8

24 : Follow this link when the job is finished /home/cynthia/Downloads/Galaxy/galaxy/tools/myTools/index.html

data
format **data**, database **?**

Your query: Can I create a Galaxy workflow to identify polymorphic

Figure 6.3: Screenshot of the history panel after running the tool



Started tool **NL2WF** and successfully added 1 job to the queue.

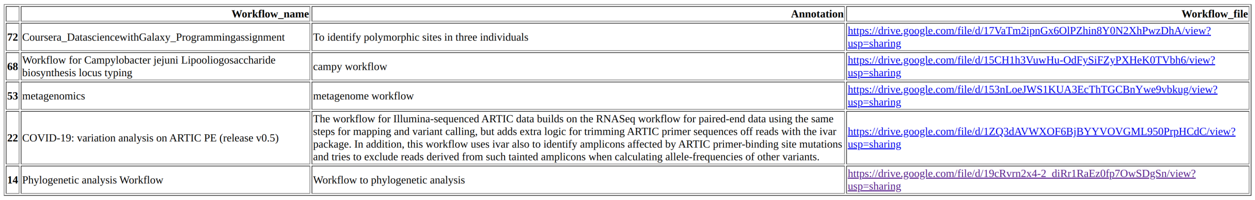
The tool uses this input:

It produces this output:

- **25: Follow this link when the job is finished /home/cynthia/Downloads/Galaxy/galaxy/tools/myTools/index.html**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Figure 6.4: Screenshot of the link provided to find the result



	Workflow_name	Annotation	Workflow_file
72	Coursera_DatasciencewithGalaxy_Programmingassignment	To identify polymorphic sites in three individuals	https://drive.google.com/file/d/17VaTm2ipnGx6OlPZhin8Y0N2XhPwzDhA/view?usp=sharing
68	Workflow for Campylobacter jejuni Lipooligosaccharide biosynthesis locus typing	campy workflow	https://drive.google.com/file/d/15CH1h3VuwHu-OdFySiFZyPXHeK0TVbh6/view?usp=sharing
53	metagenomics	metagenome workflow	https://drive.google.com/file/d/153nLoeJWS1KUA3EcThTGCBnYwe9vbkug/view?usp=sharing
22	COVID-19: variation analysis on ARTIC PE (release v0.5)	The workflow for Illumina-sequenced ARTIC data builds on the RNASeq workflow for paired-end data using the same steps for mapping and variant calling, but adds extra logic for trimming ARTIC primer sequences off reads with the ivar package. In addition, this workflow uses ivar also to identify amplicons affected by ARTIC primer-binding site mutations and tries to exclude reads derived from such tainted amplicons when calculating allele-frequencies of other variants.	https://drive.google.com/file/d/1ZQ3dAVWXOF6BjBYYVOVGML950PrpHCdC/view?usp=sharing
14	Phylogenetic analysis Workflow	Workflow to phylogenetic analysis	https://drive.google.com/file/d/19cRvrn2x4-2_diRr1RaEz0fp7OwSDgSn/view?usp=sharing

Figure 6.5: Screenshot of the result based on the user query

and I'd like to combine these files as a single VCF. The bcf_contact tool will do it, but how do I feed both files to the tool in a workflow?"[8]

Hello, May I ask a question to create the reference from my own WGS data using Galaxy. I have my WGS data that was already mapped with the native reference (BAM file). Then I want to use this kind of data as a reference for further analysis with another dataset to identify SNPs. Previously, I try to create it by converting the BAM file to BED and then using Bedtools: getfastabed to generate the FASTA file. However, this FASTA seems not the reference from my own data. Could you recommend the step to generate the FASTA file for reference? I try to file the protocol in Custom Reference Genome topic, but there is no mention of the way to generate the FASTA file from the individual data. Only the FASTA format is mentioned.[9]

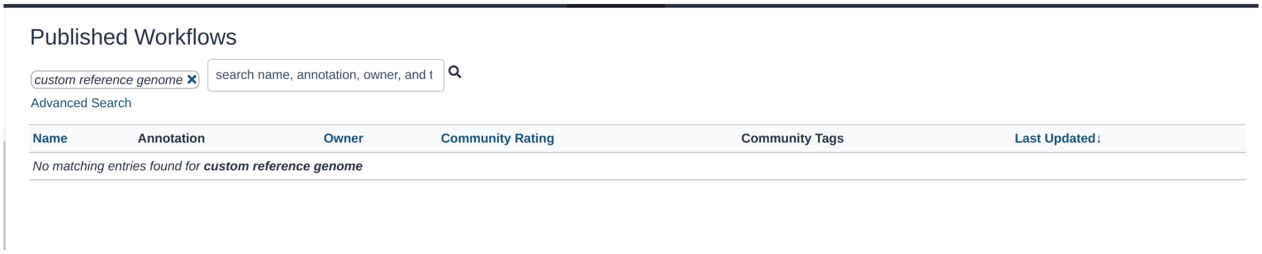
These posts encourage us to investigate why users are facing difficulties in designing workflows according to their requirements. We find the reasons behind that

1. **Lack of existing helping forums:** The workflow management systems are very domain specific. Users conducting experiments in these systems, need to have prior knowledge of the working procedures and terminologies. Users cannot find any solutions from any other generic question-answer sites. They have to rely on the forums offered by the workflow management systems. Otherwise, they need to contact the experts manually which is not always possible or the contact information of the experts may not always be available. Moreover, against 100 users there might be 1 expert so this option is not a feasible one. Again, when users post their queries about designing workflows on a certain topic, they simply might not get any answers from the other users or experts. Because there might be a shortage of answer providers who are experts in that certain topic. So frustration is bound to grow up inside the users as there are not many services where they can put their queries and the existing one is not coming to any of their help.

2. **Lack of provided features to help:** When users need to look for a workflow for getting ideas to implement in their proposed workflow, they need to search for it. But extensive search and a lack of annotations of the shared workflows exist in this case. Most of the time they do not provide insightful information about the workflow. Thus, users get confused about whether they can use the existing workflows or not for meeting their requirements.

Answering RQ1: We find the necessity of building a workflow recommendation tool from the users' posts. We investigate the reason behind the cause of not getting proper workflow recommendations and find that the system does not provide any such platform for the users where they can get the workflows without any prior knowledge or having to wait for the experts to answer their queries. This is the reason, a tool or platform to serve these purposes is very necessary.

[8] https://help.galaxyproject.org/t/how-to-combine-2-vcf-files-in-a-workflow/2595
[9] https://help.galaxyproject.org/t/create-reference-genome-from-my-wgs-data-custom-reference/1051



Published Workflows

custom reference genome ✕ search name, annotation, owner, and t

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated↓
No matching entries found for ***custom reference genome***					

Figure 6.6: Found no result in the shared workflow repository when searched using two or more keywords

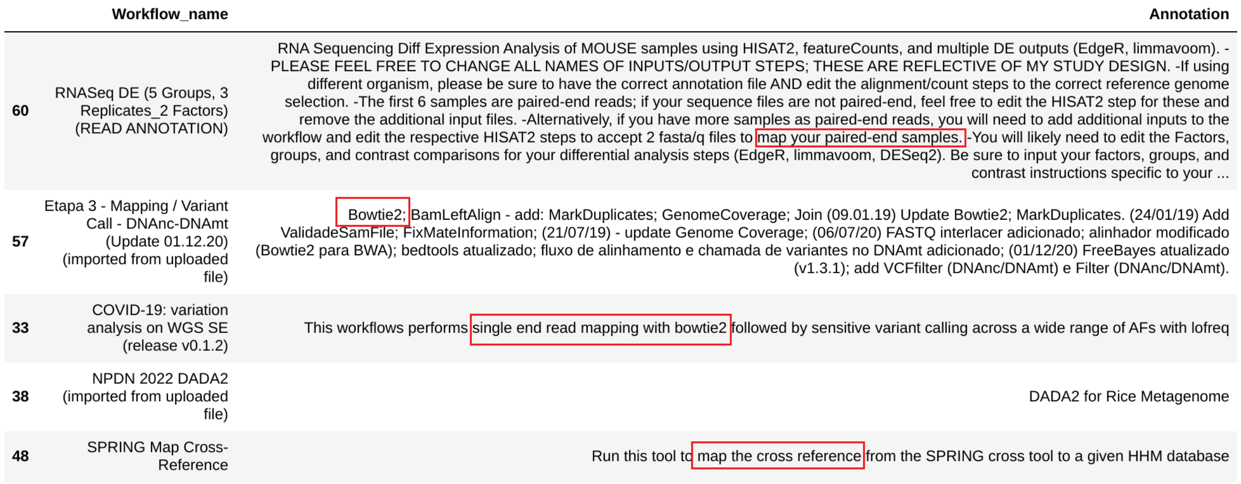
6.4.2 RQ2: How does the NL2WF tool differ from the normal search in the repository?

While developing our tool, we emphasized the fact that the tool should serve the unique purpose of increasing the usability of SWfMSs. So one question can come to the reader's mind that, we can search for the workflows in the repositories, but what is new in our developed tool? To answer this question, before developing the tool, we investigated and explored the shared workflow repository by submitting queries and we found that the search mechanism only works by matching keywords. And if any user gives more than one keyword, then the search mechanism cannot provide any result. For example, we try to search workflows according to the user query, "I want to map my reads with Bowtie 2 to a custom reference genome"[10], but as the search option only works for titles or annotations or tags, we tried searching with only three keywords. And the result is shown in Fig 6.6.

Also, when searched using one keyword, the result is shown providing the list of the workflows containing only that particular matched keyword. But through our developed tool, even if it could not find the exactly matched workflow, it will possibly show the workflows for all the keywords. For any user, getting the list of workflows where all the keywords are matched or most of the keywords are matched, is far better than searching workflows keywords by keywords. For the above-mentioned example, the output result of our method is shown in Fig 6.7, where we can see that the recommended workflow is a list of collective workflows matching the most number of keywords.

Answering RQ2: We can see that when searched with multiple keywords, the searched result contains no outputs. But following our method, NL2WF can recommend several workflows according to the user query. Thus our tool is different from the normal search in repository.

[10] https://help.galaxyproject.org/t/bowtie-2-in-workflow-no-custom-reference-genome-available-solution-create-a-custom-build/1985



	Workflow_name	Annotation
60	RNASeq DE (5 Groups, 3 Replicates_2 Factors) (READ ANNOTATION)	RNA Sequencing Diff Expression Analysis of MOUSE samples using HISAT2, featureCounts, and multiple DE outputs (EdgeR, limmavoom). -PLEASE FEEL FREE TO CHANGE ALL NAMES OF INPUTS/OUTPUT STEPS; THESE ARE REFLECTIVE OF MY STUDY DESIGN. -If using different organism, please be sure to have the correct annotation file AND edit the alignment/count steps to the correct reference genome selection. -The first 6 samples are paired-end reads; if your sequence files are not paired-end, feel free to edit the HISAT2 step for these and remove the additional input files. -Alternatively, if you have more samples as paired-end reads, you will need to add additional inputs to the workflow and edit the respective HISAT2 steps to accept 2 fasta/q files to map your paired-end samples. -You will likely need to edit the Factors, groups, and contrast comparisons for your differential analysis steps (EdgeR, limmavoom, DESeq2). Be sure to input your factors, groups, and contrast instructions specific to your ...
57	Etapa 3 - Mapping / Variant Call - DNAnc-DNAmt (Update 01.12.20) (imported from uploaded file)	Bowtie2; BamLeftAlign - add: MarkDuplicates; GenomeCoverage; Join (09.01.19) Update Bowtie2; MarkDuplicates. (24/01/19) Add ValidadeSamFile; FixMateInformation; (21/07/19) - update Genome Coverage; (06/07/20) FASTQ interlacer adicionado; alinhador modificado (Bowtie2 para BWA); bedtools atualizado; fluxo de alinhamento e chamada de variantes no DNAmt adicionado; (01/12/20) FreeBayes atualizado (v1.3.1); add VCFfilter (DNAnc/DNAmt) e Filter (DNAnc/DNAmt).
33	COVID-19: variation analysis on WGS SE (release v0.1.2)	This workflows performs single end read mapping with bowtie2 followed by sensitive variant calling across a wide range of AFs with lofreq
38	NPDN 2022 DADA2 (imported from uploaded file)	DADA2 for Rice Metagenome
48	SPRING Map Cross-Reference	Run this tool to map the cross reference from the SPRING cross tool to a given HHM database

Figure 6.7: Result of the user query

6.4.3 RQ3: How do users assess the usefulness and effectiveness of the NL2WF tool?

For our NL2WF tool, we designed and conducted a user study with 15 participants. We took survey as our study method. The participants were assigned to complete tasks on their choices. In this study, our goal is to evaluate the effectiveness and usefulness of the tool. That is why our end users are the scientists, researchers and students who are from the bioinformatics and software research domain.

Survey Design

We designed the survey as follows:

1. To emulate a real-world experience, we provided the participants with some sample queries to test the tool. The queries were collected from the user forum posted by real users. We tried to investigate whether the tool's result of the real user queries can make sense to the participants. We showed the sample queries to help the users so that they can get a clear idea about what types of queries are posted in the forum, and what type of sentences they can write in the query box etc.

2. We provided our local machine where Galaxy local system has been installed. As our tool has been developed on the local server, so the results were also given on the local machine.

3. The participants were told to run the tool following some steps. The steps were described in the survey and a screenshot of the galaxy system explaining the working procedure of each of the sections was provided too. The participants had the independence of writing their own generated queries or they could also test the tool by using the sample queries.

4. Then we attached a link to the video of the tool demonstrating how it works with an example so that participants can get a clear image of the working procedure of the system.

5. After that, we included some of the sample queries and their corresponding answers for the participants who want to evaluate the tool by understanding the tool demonstration. The results were shown on the survey so that the participants can get some ideas about how the result appears, which portion indicates what and what they can gain from the showed result.

6. Finally the participants were given the questions according to the Likert Scale to evaluate the tool. After that, they were given a question set to evaluate the task load according to the NASA-TLX (Task Loading Index) scale.

7. The participants had full authority over whether they wanted to test the tool with their own queries or from the displayed sample queries.

Video Demo and the Experiment

We recorded a video demo for the participants to show how the tool works. We at first recorded one video for the participants who were from the bioinformatics domain and another video for the participants with a general software research background. In the first video, we showed an example of writing a query to get related workflows for *phylogenetic* analysis. We showed the imported workflow in the video demonstration. In the second video, we showed the example of *iris* data analysis. We put our query in natural language which asks for showing workflows to analyze the *iris* data. Then we showed the result of the analysis through the video demonstration.

Survey Questions

After showing the video or when the user has tested the tool, we asked the users some questions based on the Likert Scale and NASA-TLX. In the Likert scale measurement, we provided 8 statements and asked the users to rate the statements according to their judgment. The questions covered the participant's opinion or likelihood towards our tool such as *"Rate the ease of use of the tool"*, *"Rate the effectiveness of the tool"* and *"I am likely to recommend this tool to other"*. The score of the answer ranged from 1 to 5, while 1 being the *extremely dissatisfied* and 5 being the *extremely satisfied.* Next, to asses the subjective mental workload of a participant when they were given to perform on the tool, we used the NASA-TLX scale. We asked the users to answer the standard 6 questions to measure the task load index. The schema of the survey is presented in Appendix A.

Demographic of Participants

The participants were selected aiming that they have diverse technical backgrounds as well as experience in software research and bioinformatics research. We advertised our study in two ways: 1) inside the university community by sending personal emails or contacting them directly, 2) outside the university community by analyzing their background and willingness to participate in this study. We send the survey to software researchers and bioinformatics researchers. In total, we got 15 individuals to participate in our study. 12 of them are graduate students and the rest of them are researchers working on different projects. The average age of the participants is 29 years old. They have experience in the software or bioinformatics research field ranging from 2 to 18 years. 90% of the users had experience in using various types of software systems. 40% of them had used different kinds of SWfMSs for their research purpose.

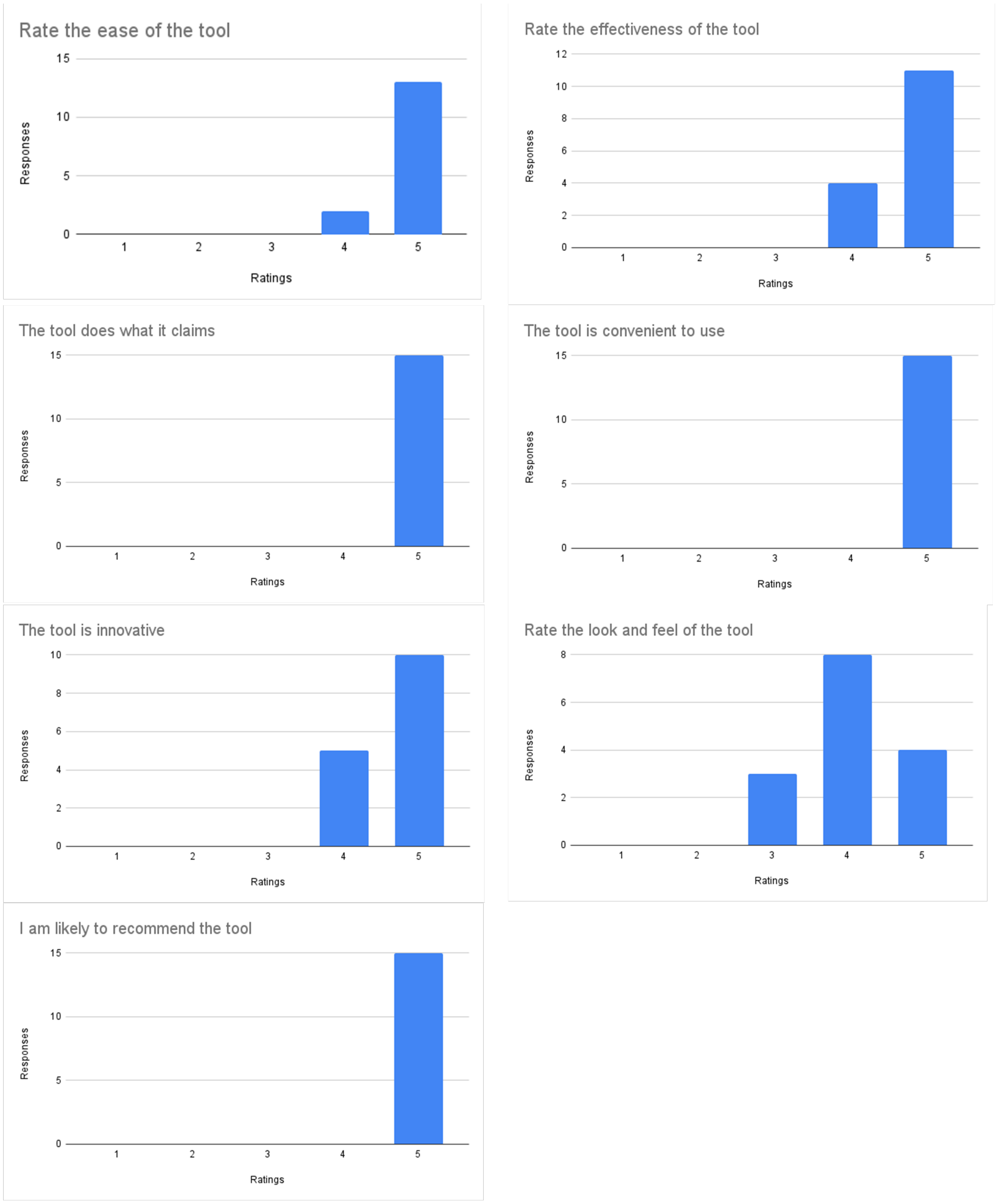
Study Result Analysis

We get the study result from two of our evaluation scales. Fig 6.8 show that participants accepted our tool in a very positive manner. Overall 79% of the participants are extremely satisfied with our tool and 18% of them are satisfied. 87% of the participants stated that the ease of the tool is extremely satisfying. When they were asked to rate the effectiveness of the tool, 73% of the participants gave a score of 5 because they are satisfied with the tool's efficacy. We asked the participants to let us know whether the tool can do what it claims to do and all of the participants agreed to the fullest in this regard. We received the highest score from all of the participants. Not only that they were also satisfied with the tool's convenience. However, the look and feel of the tool got mixed reviews. As the tool is still in the development phase and we were trying to get users' feedback regarding the efficiency of this idea, so we did not manage to provide a good look and feel for the tool. Thus 53% of the participants are somewhat satisfied with the look and feel of the tool. But at the end, all of the participants agree that they will recommend the tool to others.

Category	Total	Average
Mental Demand	103	6.87
Physical Demand	8	0.53
Temporal Demand	19	1.27
Performance	1412	94.13
Effort	48	3.2
Frustration	5	0.33

Table 6.1: The Result of NASA-TLX Survey Score

The total respondent is 15. The interpretation score of the data using NASA-TLX has been shown in Table 6.1. From the table, we can see that Mental Demand has an average score of 6.87 which implies that the participants did not need any high mental and perceptual activity to perform the task. As the tool is a very basic one containing just one text box to write queries so the average score of Physical Demand score also came very low, i.e. 0.53. Temporal Demand has an average score of 1.27 because the pace of the work requires a very less amount of time and effort. The tasks required no rushed movements from the participants. Performance has the highest score of 94.13. It indicates that all of the participants were very successful in doing what they were asked to do. So the task pressure was very low and the tool operation is



Rate the ease of the tool

Rate the effectiveness of the tool

The tool does what it claims

The tool is convenient to use

The tool is innovative

Rate the look and feel of the tool

I am likely to recommend the tool

Figure 6.8: Result analysis from the Likert Scale

very light in manner. The average score of Effort is 3.2 which describes that the participants did not need big effort to accomplish the tasks and they could do it with ease. Lastly, a very low average score of 0.33 for Frustration indicates that the participants have not felt insecure, desperate, offended or disturbed while performing the tasks.

So if we want to find the answer that how effectively the users can perform the task and what their opinions are towards the tool, we can easily say that users find the tool very effective and easy to use. The result we got from Likert Scale shows that users appreciated our tool and all of them most likely recommend the tool to others. The result from NASA-TLX survey illustrated that the task loading demand was very low and the performance score is very high. So the effectiveness of the tool is very remarkable.

We also asked the participants an open-ended question on whether they had any recommendations or suggestions for our tool. Many of the participants provided us with their views, e.g:

"The tool is of great use. Mostly for the novice users without prior domain knowledge."

"I think the tool is going to help a lot of the scientists. I can see people posting their concerns about workflow designs. So this tool will help I believe"

With appreciation, some of the participants showed us the ways we can improve the tool. For example: *"The tool could express the information more dynamically"*,

"The tool needs to be more dynamic. Otherwise the purpose of the tool is good. If the result could be evaluated by an expert then it would be more reliable for the particular domain.",

"Re-checking the results get through the tool by an expert in domain, optimal parameter selection for the tools that are suggested in the workflow" are some of them.

So we again talked with the participants for the detailed recommendation of the tool and wanted to give them the explanation why the tool was not covering all those features. As of the tool's being dynamic in nature, the main reason is that we are still developing the tool. We wanted to view the participant's perception and the effectiveness of the tool. That is why we emphasized more on the tool's concept. Secondly, we really like the idea of one participant's who stated that along with the workflows, if the tool could provide the optimal parameters too, then it would have been a great feature to use for. Some of the participants raised their concerns about evaluating the tool with the experts. We are also aware of this concern and still finding the way to do so in near future.

Answering RQ3: Our tool is being appreciated by the users and the effectiveness of the tool is proven through the scores we got from both Likert and NASA-TLX scales. The recommendations we got from the users imply that users want to see more features added to this tool to make it more powerful and they are realizing the impact of having this kind of system in their workspace.

6.5 Threats To Validity

6.5.1 Internal Validity

Threats to internal validity refers to the user study design. To ensure that the study is designed appropriately, we evaluated the study design with two of the professionals in the field of software research and bioinformatics research before sending the study form to the participants. We have modified our study after getting their feedback.

6.5.2 External Validity

The user study has been done with 15 participants and a lot of the participants are from the software research area. We tried to engage users from bioinformatics research as much as possible but we could not ensure that. But as the tools working procedure is very related to any general software systems' tools, we showed a relative workflow according to their field of study for their better understanding. Through this, we mitigate the threat of this study being biased. Moreover, we chose researchers from different parts of the world. So the generalizability of this study is also ensured.

6.6 Conclusion

The need for having a system which can recommend workflows has always been prevailing. But how the system is going to understand the users' queries turns out to be a challenging task. Thus, in this study, we analyze and explore the Galaxy user forum deeply and find prevailing problem of users' asking for recommendations of relative and proper workflows. We also discover the main constraint of the existing search option and why this method is not providing the expected outcome. Hence, we develop and integrate a tool to the Galaxy platform showing the concept of how a workflow recommendation system should work. We perform a user study to evaluate the purpose of the tool. We receive positive scores from the users having low scores on the task load. The workload is below 20% for almost all the participants. Also, there is almost 100% satisfaction rate for tool's usage and effectiveness. We find several recommendations to add and integrate into the tool.

7 Conclusion and Future Work

In this thesis, we conducted three studies to improve the usability issues of a software system using group discussion forums. Usability issues prevail in a software system and it is one of the major reasons why a software system can become unpopular among users. The main reason for identifying usability problems is to discover the users' preferences and identify the concerns they are raising about the system. So usability test brings out many significant issues regarding the system. But reaching out to the field users is always not possible because of the scarcity of their contact information, the users' interest and hindrances from the systems themselves. However, the users interact with other users and the experts of the domain through group discussion forums. Some of the systems provide such group discussion forums for the users. So from this type of forum, the usability issues of the system could be discovered as users post their concerns or requirements here. Therefore, the goal of this thesis is to investigate and explore user forums to find out the usability issues a software system is suffering from. We take Galaxy as a use case and we believe that our methods and findings will help the other software systems to resolve their usability issues too.

7.1 Categorizing The Users' Posts for Better Understanding

We find the forums rarely categorize the users' posts and users cannot find their desired posts when they want to get all the posts from a certain group. Thus, in our first study, we explore the Galaxy user forum and categorize the user posts into *six* main categories. We again further investigate the most discussed topic in the forum and find more categories for that topic. We found *seven* more categories of the most dominant topic discussed in the forum. Our categorizations are expected to help the developers to find the system's problems and resolve them according to the users' preferences.

7.2 Providing Better Suggestions for Tags

Analyzing the user forum, we find that users do not use proper tags for their posts. They often use unnecessary and irrelevant tags for their posts which do not include their posts in the expected category. Thus many of the users do not include tags at all. So we propose a method for providing a better suggestion for tags. We evaluate our method by performing a user study. They were provided with 10 sample questions from the user forums. Both the originally posted tags and our method's suggested tags were also given with the survey. We asked the users to rate the relevance level of the tags with the posted questions. Our study results show that in 5 out of the 10 cases, our method's suggested tags got more relevance scores from the users. Among the rest of the questions, 4 of the questions got the same relevance score as the posted ones. Among the questions, there were questions where tags were not been provided at all and our proposed method suggested relative tags in that case too.

7.3 Recommending Workflows From Natural Language Text

Our in-depth investigation brought up more usability problems that Galaxy users are facing. We find that the users who do not have much prior knowledge are struggling to get proper recommendations for workflows. The search option in the workflow repository is not providing the expected output. Moreover, when the users are asking for guidelines, most of the time they are not getting sufficient and fruitful answers. To solve this issue, in our third study, we developed and integrated a tool into the Galaxy platform which can recommend workflows based on the users' queries. The main advantage of our tool is that it can take natural language as a user query and provide related workflows to the users. We evaluated our tool by performing a user study with 15 participants from the software and bioinformatics research domain. We used both Likert Scale and NASA-TLX scale to asses the tool. We find that overall 79% of the users are very appreciative towards our tool and the workload seemed to be very less for all the participants. Not only that, almost all of the participants want to recommend our tool and they expect to see more features added to the tool to serve their requirements.

7.4 Future Work

In this thesis, we explore the Galaxy user forum to investigate the usability issues and improve them as well. However, our research does not investigate the other user forums for the different software systems. In the future, we plan to explore more user forums and investigate the usability issues there. We want to apply our proposed methods to those forums and check for the success rate in finding out the usability issues through these methods. A comparison of the results from different software systems' group discussion forums could bring out significant findings on how to deal with usability issues.

Along with that, we want to build adaptive software systems. We found several user concerns where the users expressed that they are not happy with the adaptiveness of the system. So in our future plan, we want to extend our research in that direction too. We want to analyze user logs and their event logs to analyze their preferences. In this way, the users regardless of any expertise in the domain knowledge can find the system very preferable to work. If we build a system which can provide features according to the users' preferences, then the system can be more approachable for any user to perform their tasks. We also plan to integrate this work in other SWfMSs, e.g., VizSciFlow [44].

References

[1] Ahmad Abdellatif, Khaled Badran, Diego Elias Costa, and Emad Shihab. A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*, 48(8):3087–3102, 2022.

[2] Laith Abualigah, Hamza Essam Alfar, Mohammad Shehab, and Alhareth Mohammed Abu Hussein. *Sentiment Analysis in Healthcare: A Brief Review*, pages 129–141. Springer International Publishing, Cham, 2020.

[3] Mayur Ahirrao, Yash Joshi, Atharva Gandhe, Sumeet Kotgire, and Rohini G Deshmukh. Phrase composing tool using natural language processing. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–4. IEEE, 2021.

[4] Bourair Al-Attar, Ali Thoulfikar A Imeer, Ahmed J Allami, Abdul Amir H Kadhum, Murtadha Yahia Abdulshaheed Altufaili, Hussein Ali Al-Bahrani, Rahem M Rahem, Thoulfikar Al-Bassam, and Ahmed A Al-Amiery. A domain-specific algorithm for arabic tag suggestion. *Webology*, 19(2):1515–1525, 2022.

[5] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: an extensible system for design and execution of scientific workflows. In *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, pages 423–424, 2004.

[6] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine*, 40(5):498–499, 2018. PMID: 30275631.

[7] Deeksha Arya, Wenting Wang, Jin L.C. Guo, and Jinghui Cheng. Analysis and detection of information types of open source software issue discussions. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 454–464, 2019.

[8] Adam Barker and Jano van Hemert. Scientific workflow: A survey and research directions. In *PPAM*, 2007.

[9] Nigel Bevan, James Carter, and Susan Harker. Iso 9241-11 revised: What have we learnt about usability since 1998? In *International conference on human-computer interaction*, pages 143–151. Springer, 2015.

[10] Sangeeta Bhagwanani et al. An evaluation of end-user interfaces of scientific workflow management systems. 2005.

[11] Tingting Bi, Peng Liang, Antony Tang, and Xin Xia. Mining architecture tactics and quality attributes knowledge in stack overflow. *Journal of Systems and Software*, 180:111005, 2021.

[12] Stefan Blomkvist. Towards a model for bridging agile development and user-centered design. In *Human-centered software engineering—integrating usability in the software development lifecycle*, pages 219–244. Springer, 2005.

[13] Simon Bray, Matthias Bernt, Nicola Soranzo, Marius van den Beek, Bérénice Batut, Helena Rasche, Martin Čech, Peter Cock, Anton Nekrutenko, Björn Grüning, and John Chilton. Planemo: a command-line toolkit for developing, deploying, and executing scientific data analyses. *bioRxiv*, 2022.

[14] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. A literature survey of recent advances in chatbots. *Information*, 13(1):41, 2022.

[15] Alex Cao, Keshav K Chintamani, Abhilash K Pandya, and R Darin Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods*, 41(1):113–117, 2009.

[16] Min-Hua Chao, Amy JC Trappey, and Chun-Ting Wu. Emerging technologies of natural language-enabled chatbots: A review and trend forecast using intelligent ontology extraction and patent analytics. *Complexity*, 2021, 2021.

[17] Eason Chen. The effect of multiple replies for natural language generation chatbots. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, 2022.

[18] Qiuyuan Chen, Chunyang Chen, Safwat Hassan, Zhengchang Xing, Xin Xia, and Ahmed E. Hassan. How should i improve the ui of my app? a study of user reviews of popular apps in the google play. *ACM Trans. Softw. Eng. Methodol.*, 30(3), apr 2021.

[19] KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[20] The Galaxy Community. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345, 2022.

[21] Larry L Constantine and L Lockwood. Process agility and software usability: Toward lightweight usage-centered design. *Information Age*, 8(8):1–10, 2002.

[22] Scott A Crossley, Laura K Allen, Kristopher Kyle, and Danielle S McNamara. Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, 51(5-6):511–534, 2014.

[23] Felipe Cujar-Rosero. Nature: a tool resulting from the union of artificial intelligence and natural language processing for searching research projects in colombia. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 12(4), 2021.

[24] Wana Maria de Souza, Antonio Júnior Alves Ribeiro, and Carlos Augusto Uchôa da Silva. Use of ann and visual-manual classification for prediction of soil properties for paving purposes. *International Journal of Pavement Engineering*, 23(5):1482–1490, 2022.

[25] Ewa Deelman and Yolanda Gil. Managing large-scale scientific workflows in distributed environments: Experiences and challenges. In *2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, pages 144–144, 2006.

[26] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J. Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, 46:17–35, 2015.

[27] Danielle D DeSouza, Jessica Robin, Melisa Gumus, and Anthony Yeung. Natural language processing as an emerging tool to detect late-life depression. *Frontiers in Psychiatry*, page 1525, 2021.

[28] Camila Rodrigues Dias, Marluce Rodrigues Pereira, and Andre Pimenta Freire. Qualitative review of usability problems in health information systems for radiology. *Journal of biomedical informatics*, 76:19–33, 2017.

[29] Evandro JS Diniz, José E Fontenele, Adonias C de Oliveira, Victor H Bastos, Silmar Teixeira, Ricardo L Rabêlo, Dario B Calçada, Renato M Dos Santos, Ana K de Oliveira, and Ariel S Teles. Boamente: A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. In *Healthcare*, volume 10, page 698. MDPI, 2022.

[30] Xubo Fei, Shiyong Lu, and Jia Zhang. A granular concurrency control for collaborative scientific workflow composition. In *2011 IEEE International Conference on Services Computing*, pages 410–417, 2011.

[31] Jennifer Ferreira, James Noble, and Robert Biddle. Agile development iterations and ui design. In *Agile 2007 (AGILE 2007)*, pages 50–58. IEEE, 2007.

[32] Rosa Filguiera, Iraklis Klampanos, Amrey Krause, Mario David, Alexander Moreno, and Malcolm Atkinson. dispel4py: A python framework for data-intensive scientific computing. In *2014 International Workshop on Data Intensive Scalable Computing Systems*, pages 9–16, 2014.

[33] Asbjørn Følstad. The effect of group discussions in usability inspection: A pilot study. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, NordiCHI '08, page 467–470, New York, NY, USA, 2008. Association for Computing Machinery.

[34] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):r86, 2010.

[35] Katharina Görlach, Mirko Sonntag, Dimka Karastoyanova, Frank Leymann, and Michael Reiter. *Conventional Workflow Technology for Scientific Simulation*, pages 323–352. Springer London, London, 2011.

[36] Cigdem Altin Gumussoy. Usability guideline for banking software design. *Computers in Human Behavior*, 62:277–285, 2016.

[37] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*, 2011.

[38] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[39] Tabitha Hart and Peg Achterman. Qualitative analysis software. *The international encyclopedia of communication research methods*, pages 1–12, 2017.

[40] Md. Rakibul Hasan, Maisha Maliha, and M. Arifuzzaman. Sentiment analysis with nlp on twitter data. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pages 1–4, 2019.

[41] Barbora Hladka and Martin Holub. A gentle introduction to machine learning for natural language processing: How to start in 16 practical steps. *Language and Linguistics Compass*, 9(2):55–76, 2015.

[42] Rune Thaarup Høegh. Usability problems: Do software developers already know? In *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, OZCHI '06, page 425–428, New York, NY, USA, 2006. Association for Computing Machinery.

[43] Sonja Holl, Olav Zimmermann, and Martin Hofmann-Apitius. A new optimization phase for scientific workflow management systems. In *2012 IEEE 8th International Conference on E-Science*, pages 1–8, 2012.

[44] Muhammad M. Hossain, Banani Roy, Chanchal K. Roy, and Kevin A. Schneider. Vizsciflow: A visually guided scripting framework for supporting complex scientific data analysis. *Proc. ACM Hum.-Comput. Interact.*, 4(EICS), jun 2020.

[45] E. Houstis, E. Gallopoulos, R. Bramley, and J. Rice. Problem-solving environments for computational science. *IEEE Computational Science and Engineering*, 4(3):18–21, 1997.

[46] Carlos Huertas and Reyes Juárez-Ramírez. Nlare, a natural language processing tool for automatic requirements evaluation. In *Proceedings of the CUBE International Information Technology Conference*, pages 371–378, 2012.

[47] Sara R. Jaeger and Morten A. Rasmussen. Importance of data preparation when analysing written responses to open-ended questions: An empirical assessment and comparison with manual coding. *Food Quality and Preference*, 93:104270, 2021.

[48] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[49] Markus Jelonek, Arne Peter Raulf, Eileen Fiala, Joshua Butke, Thomas Herrmann, and Axel Mosig. A formative usability study of workflow management systems in label-free digital pathology. *F1000Research*, 11(192):192, 2022.

[50] Barbara Johnstone. *Discourse analysis.* John Wiley & Sons, 2017.

[51] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015.

[52] Jackson Kamiri and Geoffrey Mariga. Research methods in machine learning: A content analysis. *International Journal of Computer and Information Technology (2279-0764)*, 10(2), 2021.

[53] Zunwang Ke, Jiabao Sheng, Zhe Li, Wushour Silamu, and Qinglang Guo. Knowledge-guided sentiment analysis via learning from natural language explanations. *IEEE Access*, 9:3570–3578, 2021.

[54] Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*, 2020.

[55] Pavneet Singh Kochhar. Mining testing questions on stack overflow. In *Proceedings of the 5th International Workshop on Software Mining*, SoftwareMining 2016, page 32–38, New York, NY, USA, 2016. Association for Computing Machinery.

[56] Ross Koppel, Joshua P Metlay, Abigail Cohen, Brian Abaluck, A Russell Localio, Stephen E Kimmel, and Brian L Strom. Role of computerized physician order entry systems in facilitating medication errors. *Jama*, 293(10):1197–1203, 2005.

[57] Ratnakar Kumar and Nitasha Hasteer. Evaluating usability of a web application: A comparative analysis of open-source tools. In *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, pages 350–354, 2017.

[58] Tarald O Kvålseth. Note on cohen's kappa. *Psychological reports*, 65(1):223–226, 1989.

[59] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6. IEEE, 2016.

[60] Jessica Nina Lester, Yonjoo Cho, and Chad R. Lochmiller. Learning to do qualitative data analysis: A starting point. *Human Resource Development Review*, 19:106 – 94, 2020.

[61] Xiu Li, Jingdong Song, and Biqing Huang. A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics. *The International Journal of Advanced Manufacturing Technology*, 84(1):119–131, 2016.

[62] Elizabeth D Liddy. Natural language processing. 2001.

[63] Thomas R Lindlof and Bryan C Taylor. *Qualitative communication research methods.* Sage publications, 2017.

[64] Ji Liu, Esther Pacitti, Patrick Valduriez, and Marta Mattoso. A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, 13(4):457–493, 2015.

[65] Mingwei Liu, Xin Peng, Qingtao Jiang, Andrian Marcus, Junwen Yang, and Wenyun Zhao. Searching stackoverflow questions with multi-faceted categorization. In *Proceedings of the Tenth Asia-Pacific Symposium on Internetware*, pages 1–10, 2018.

[66] Bertram Ludäscher, Shawn Bowers, and Timothy McPhillips. *Scientific Workflows*, pages 3320–3324. Springer New York, New York, NY, 2018.

[67] Soumi Majumder and Atreyee Mondal. Are chatbots really useful for human resource management? *International Journal of Speech Technology*, 24(4):969–977, 2021.

[68] Golam Mostaeen, Banani Roy, Chanchal Roy, and Kevin Schneider. Designing for real-time groupware systems to support complex scientific data analysis. *Proc. ACM Hum.-Comput. Interact.*, 3(EICS), jun 2019.

[69] Golam Mostaeen, Jeffrey Svajlenko, Banani Roy, Chanchal K. Roy, and Kevin A. Schneider. [research paper] on the use of machine learning techniques towards the design of cloud based automatic code clone validation tools. In *2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation (SCAM)*, pages 155–164, 2018.

[70] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[71] AmirHossein Naghshzan. Towards code summarization of apis based on unofficial documentation using nlp techniques. *arXiv preprint arXiv:2208.06318*, 2022.

[72] Kheline FP Naves, Adriano A Pereira, Slawomir J Nasuto, Ieda PC Russo, and Adriano O Andrade. Assessment of inter-examiner agreement and variability in the manual classification of auditory brainstem response. *Biomedical engineering online*, 11(1):1–10, 2012.

[73] Tomoko Nemoto and David Beglar. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–8, 2014.

[74] Jakob Nielsen. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing, USA, 1999.

[75] Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. Cogcomptime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, 2018.

[76] Ayush Noori, Colin Magdamo, Xiao Liu, Tanish Tyagi, Zhaozhi Li, Akhil Kondepudi, Haitham Alabsi, Emily Rudmann, Douglas Wilcox, Laura Brenner, et al. Development and evaluation of a natural language processing annotation tool to facilitate phenotyping of cognitive status in electronic health records: Diagnostic study. *Journal of medical Internet research*, 24(8):e40384, 2022.

[77] O Obono and K Obono. Analysis of qualitative data. 2008.

[78] Cecilia Maria Patino and Juliana Carvalho Ferreira. Internal and external validity: can you apply research study results to your patients? *Jornal brasileiro de pneumologia*, 44:183–183, 2018.

[79] Marco Enrico Piras, Luca Pireddu, and Gianluigi Zanetti. wft4galaxy: a workflow testing tool for galaxy. *Bioinformatics*, 33(23):3805–3807, 07 2017.

[80] D Sai Pranav, Mehar Mutreja, Devansh Punj, and Pronika Chawla. Natural language processing in chatbots. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 1*, 491:87, 2022.

[81] Aung Pyae and Tapani N. Joelsson. Investigating the usability and user experiences of voice user interface: A case of google home smart speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '18, page 127–131, New York, NY, USA, 2018. Association for Computing Machinery.

[82] Yuxing Qian, Wenxuan Gui, Feicheng Ma, and Qingxing Dong. Exploring features of social support in a chinese online smoking cessation community: A multidimensional content analysis of user interaction data. *Health Informatics Journal*, 27(2):14604582211021472, 2021.

[83] Yuan Fu Qiu, Yoon Ping Chui, and Martin G Helander. Usability analysis of mobile phone camera software systems. In *2006 IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6. IEEE, 2006.

[84] M. A. Rahman. Scalable scientific workflows management system swfms. *International Journal of Advanced Computer Science and Applications*, 7, 2016.

[85] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1, 2012.

[86] Adil Rajput. Chapter 3 - natural language processing, sentiment analysis, and clinical analytics. In Miltiadis D. Lytras and Akila Sarirete, editors, *Innovation in Health Informatics*, Next Gen Tech Driven Personalized Med&Smart Healthcare, pages 79–97. Academic Press, 2020.

[87] G Appa Rao, G Srinivas, K Venkata Rao, and PVGD Prasad Reddy. A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. *IJSC—ICTACT J Soft Comput*, 8(4):1728–1732, 2018.

[88] Catherine Kohler Riessman. *Narrative analysis*, volume 30. Sage, 1993.

[89] N Roopak and Gerard Deepak. Knowgen: a knowledge generation approach for tag recommendation using ontology and honey bee algorithm. In *European, Asian, Middle Eastern, North African Conference on Management & Information Systems*, pages 345–357. Springer, 2021.

[90] Pradeep Kumar Roy and Jyoti Prakash Singh. A tag2vec approach for questions tag suggestion on community question answering sites. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 168–182. Springer, 2018.

[91] Idafen Santana-Perez and María S Pérez-Hernández. Towards reproducibility in scientific workflows: An infrastructure-based approach. *Scientific Programming*, 2015, 2015.

[92] Shilad Sen, Jesse Vig, and John Riedl. Tagommenders: Connecting users to items through tags. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 671–680, New York, NY, USA, 2009. Association for Computing Machinery.

[93] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. Eviza: A natural language interface for visual analysis. UIST '16, page 365–377, New York, NY, USA, 2016. Association for Computing Machinery.

[94] Lea Sgier. Qualitative data analysis. *An Initiat. Gebert Ruf Stift*, 19:19–21, 2012.

[95] Prabhnoor Singh, Rajkanwar Chopra, Ojasvi Sharma, and Rekha Singla. Stackoverflow tag prediction using tag associations and code analysis. *Journal of Discrete Mathematical Sciences and Cryptography*, 23(1):35–43, 2020.

[96] Mirko Sonntag, Dimka Karastoyanova, and Frank Leymann. The missing features of workflow systems for scientific computations. *Software Engineering 2010–Workshopband (inkl. Doktorandensymposium)*, 2010.

[97] Gridaphat Sriharee. An ontology-based approach to auto-tagging articles. *Vietnam J. of Computer Science*, 2(2):85–94, may 2015.

[98] Juan Tan, Xiaohui Gao, Qiong Tan, and Hongwei Zhao. Multiple time series perceptive network for user tag suggestion in online innovation community. *IEEE Access*, 9:28059–28065, 2021.

[99] William H Thiel. Galaxy workflows for web-based bioinformatics analysis of aptamer high-throughput sequencing data. *Molecular Therapy - Nucleic Acids*, 5:e345, 2016.

[100] Daniele Turi, Paolo Missier, Carole Goble, David De Roure, and Tom Oinn. Taverna workflows: Syntax and semantics. In *Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007)*, pages 441–448. IEEE, 2007.

[101] M.B. Twidale and D.M. Nichols. Exploring usability discussions in open source development. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 198c–198c, 2005.

[102] Gias Uddin, Fatima Sabir, Yann-Gaël Guéhéneuc, Omar Alam, and Foutse Khomh. An empirical study of iot topics in iot developer discussions on stack overflow. *Empirical Software Engineering*, 26(6):1–45, 2021.

[103] Sirje Virkus and Emmanouel Garoufallou. Data science and its relationship to library and information science: a content analysis. *Data Technologies and Applications*, 2020.

[104] Diane Walker and Florence Myrick. Grounded theory: An exploration of process and procedure. *Qualitative health research*, 16(4):547–559, 2006.

[105] LP Wong. Data analysis in qualitative research: A brief guide to using nvivo. *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia*, 3(1):14, 2008.

[106] Yong Wu, Shengqu Xi, Yuan Yao, Feng Xu, Hanghang Tong, and Jian Lu. Guiding supervised topic modeling for content based tag recommendation. *Neurocomputing*, 314:479–489, 2018.

[107] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information Sciences*, 307:39–52, 2015.

[108] Liang Xiao, Qibei Lu, and Feipeng Guo. Mobile personalized recommendation model based on privacy concerns and context analysis for the sustainable development of m-commerce. *Sustainability*, 12(7):3036, 2020.

[109] Bowen Xu, Thong Hoang, Abhishek Sharma, Chengran Yang, Xin Xia, and David Lo. Post2vec: Learning distributed representations of stack overflow posts. *IEEE Transactions on Software Engineering*, 2021.

[110] Frank F. Xu, Bogdan Vasilescu, and Graham Neubig. In-ide code generation from natural language: Promise and challenges. 31(2), mar 2022.

[111] Jiansong Zhang and Nora M El-Gohary. Integrating semantic nlp and logic reasoning into a unified system for fully-automated code checking. *Automation in construction*, 73:45–57, 2017.

[112] Z. Zhao, A. Belloum, A. Wibisono, F. Terpstra, P.T. de Boer, P. Sloot, and B. Hertzberger. Scientific workflow management: between generality and applicability. In *Fifth International Conference on Quality Software (QSIC'05)*, pages 357–364, 2005.

Appendix A

Survey Questions

The questions asked on the Google form for Study 2 are listed below:

- In which country do you currently reside?
- What category below induces your age?
- To which gender identity do you most identify?
- How long do you have experience in software engineering and scientific workflow management system field?
- Question - ”Fastp jobs that were queued over 16 hrs ago have still not started running.” Posted Tags: ’queued-gray-dataset’, ’server-side-delay’ Suggested Tags: ’admin’, ’jobs’, ’workflow’, ’galaxy-local’, ’tutorial’
- Question - ”Hi everyone

  I am trying to install locally galaxy, I am using conda 4.14.0 with ubuntu (20.04.4 LTS). The installation seems correct after running ./run.sh it seems everything went on fine, and after a while the installation freeze at “solving enviroment”, i left it like that for hours and nothing happened.

  I saw in other post that this problem was fixed using virtualenv:

  I installed virtualenv, desactivate conda, activate virtualenv, and the run the ./run.sh and I get this error..”

  Posted Tags: ’admin’, ’galaxy-local’

  Suggested Tags: ’tool-install’, ’galaxy-local’, ’ubuntu’, ’conda’, ’error’, ’python3’
- Question - ”Hello, I am trying to analyze a ChIPseq experiment for the first time. I have a collection of ten fastq files, and I queued a series of jobs on this collection: (1) FastQC, (2) Trimmomatic, (3) FastQC on the Trimmomatic output, and (4) BWA on the Trimmomatic output. The first FastQC job stalled with only 8/10 files completed. The Trimmomatic and subsequent FastQC jobs completed. The BWA appears stalled on the first 2/10 files. These have been stalled since Friday. Wondering if this is a typical run time and I should keep waiting, or if this is a sign of a problem? Very new to Galaxy. Thank you!” Posted Tags: ’queued-gray-datasets’, ’server-side-delay’ Suggested Tags: ’jobs’, ’workflow’, ’public-galaxy-server’, ’server-side-delay’, ’queued-gray-datasets’, ’dataset’, ’fastq’, ’bowtie2’, ’fastq-format-error’
- Question - ”Hello siam here. I am a newbie in terms of MD simulation. For my research purpose, I need to simulate 100ns. how do I incorporate that into the galaxy server? how many steps do I need to include for a 100ns simulation? Do I need to step of ps? thank you.” Posted Tags: ’tutorial’, ’comp-chemistry’ Suggested Tags: ’workflow’, ’tutorial’, ’galaxy-local’, ’error’
- Question - ”This is the full error: /mnt/home/galaxy/database/jobs_directory/000/104/tool_script.sh: line 9: wine64_anyuser: command not found If I navigate to the manage dependencies view, I can see that wine64 is not installed, and msconverrt is listed in the section ‘no requirements defined’, so for some reason it isn’t telling galaxy that it needs wine64 to run. However I have used msconvert in other (admittedly much more mature) galaxy instances. Does anyone have any ideas?”

  Posted Tags: ’tool-install’ Suggested Tags: ’tool-install’, ’tool-dev’
- Question - ”Hi, I have some NGS data for a study on microbes but my samples do not match the fastq sample identifier. They are all mixed up and I need to swap them so they match up to my actual samples. Is there a way to actually do this? best wishes”

  Posted Tags: ’text-manipulation’ Suggested Tags: ’fastq’, ’dataset-collection’, ’collections’

- Question - ”Hello, I am doing an analysis of previously published ChIP-seq data for a bioinformatics class that I am taking. When I try to use the plotFingerprint tool to compare the IP strength between two samples (treatment/control) it is giving me the following message: “An error occurred with this dataset: format png database mm10. Job failed.” I’m not sure how to correct this issue so if anyone has any suggestions that would be great. Thanks, Anthony”

  Posted Tags: N/A Suggested Tags: ’error’, ’toubleshooting’, ’workflow’

- Question - ”I would like to fully delete a user so I go into the admin panel and delete them, while the wording is not ideal (“delete” is more like deactivate since nothing actually gets removed. . .), fine, I get that they are now unable to login. Now I want to really totally nuke the account, so I purge them. Why do they still exist in the GUI as purged. They are still in the database so that email cannot be reused. Can someone explain why purging does not actually fully purge them? Is there a way to completely delete a user?

  Also, if I do not create an account for someone manually, but they have a valid local user and login via PAM, why does their account then get listed as purged. . . ”

  Posted Tags: ’admin’, ’galaxy-local’ Suggested Tags: ’admin’, ’galaxy-local’

- Question - ”I keep getting this error for my fastq files when uploading to GalaxyTrakr. I’ve kept trying for the past 2 weeks with no resolution. Any help would be appreciated!”

  Posted Tags: ’error’ Suggested Tags: ’public-galaxy-server’, ’upload’, ’uploaded-by-url’

- Question - ”Hello everyone! I am working on developing some Galaxy tools.

  A problem I have come across a couple of times is programs requiring input files with specific file extensions e.g. .JPG.

  This is challenging for me as it is my understanding that the Galaxy representation of uploaded files is arbitrarily named with the .dat extension.

  Is there a way I can control the extension of uploaded files? Can I rewrite my tool XML to work around this?

  I would really appreciate any thoughts or suggestions :slight_smile:

  All the best and many thanks,

  Oliver”

  Posted Tags: ’admin’, ’tool-dev’ Suggested Tags: ’release’, ’tool-panel’, ’collections’, ’tool-dev’, ’upload’

The questions asked for Study 3 are given below.
Questions asked in Likert Scale survey are:

- Rate the ease of use of the tool
- Rate the effectiveness of the tool
- The tool does what it claims
- tool is convenient to use
- The tool is innovative
- Rate the look and feel of the tool
- I am likely to recommend this tool to others
- Are there any improvements that you feel we could make to the tool?

Questions asked in NASA-TLX survey are:

- How mentally demanding was the task?

- How physically demanding was the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed and annoyed were you?

Appendix B

Dataset

The dataset we have used in our thesis can be found here: https://usaskca1-my.sharepoint.com/:f:/g/personal/uji657_usask_ca/Eg86jRBTho9Ove2uEuhlBnQBtxck82qb8_8gY9g-dyD7RQ?e=5yLv9K

The replication package can be found here: https://github.com/cynthia247/Workflow-Recommendation-Tool