Stacking Answers, GitHubing Solutions: Exploring Developer Challenges in Scientific Workflow Management Systems through Advanced Topic Modelling

Chi Vu, Khairul Alam, Banani Roy

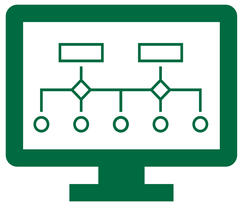
Department of Computer Science, University of Saskatchewan

INTRODUCTION

Scientific workflow management systems (SWfMSs) play a pivotal role in coordinating complex computational workflows within scientific research. However, developers often encounter challenges while working with these systems.

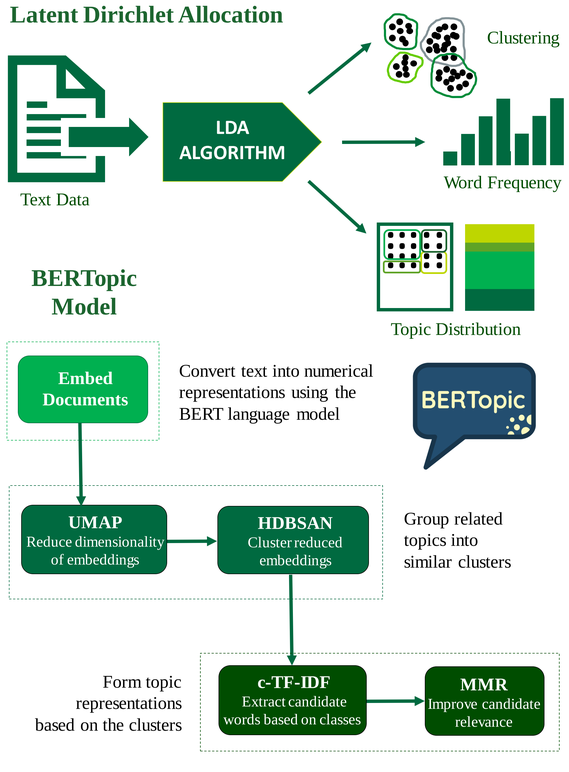
This project employs advanced topic modelling techniques, such as **Latent Dirichlet Allocation** (LDA) and the **BERTopic** model, to extract meaningful data from **Stack Overflow** posts and **GitHub Issues**.

The insights gained from this study addresses practical concerns of real-world projects, which can assist the development of **more efficient** and **user-friendly** SWfMSs.

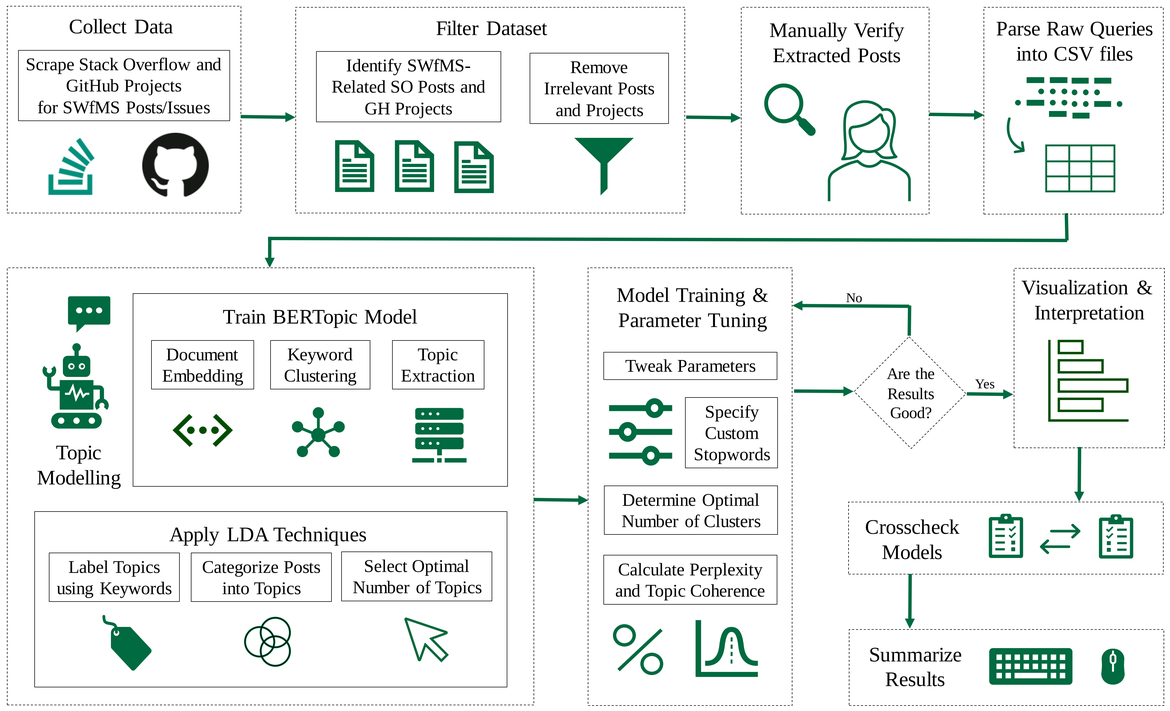


BACKGROUND

Topic modelling is a natural language processing technique used to extract and identify the main subjects discussed in a large amount of text. The goal is to discover hidden patterns and structures in a text corpus by grouping similar documents together based on the topics they cover.

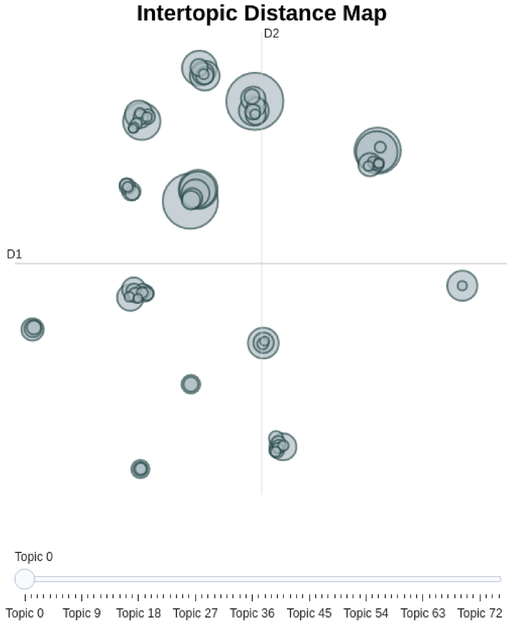


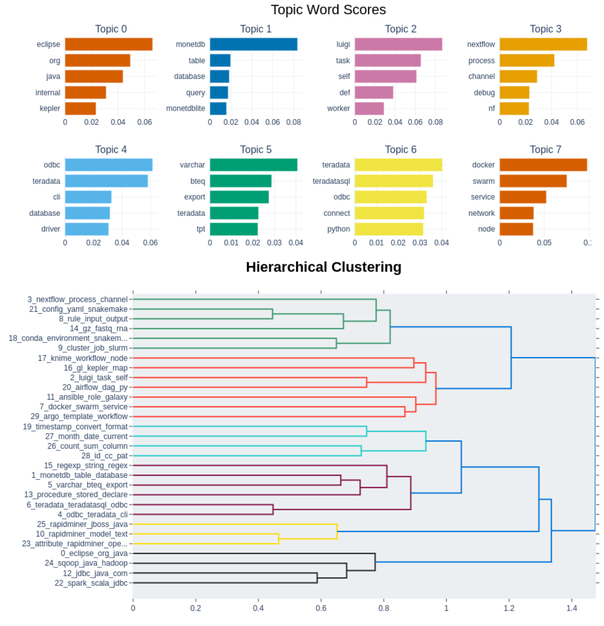
METHODOLOGY AND ANALYSIS



RESULTS

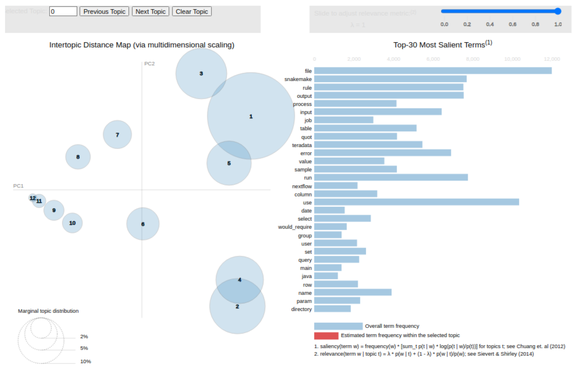
The BERTopic model identified distinctive topics from our dataset, which were characterized by clusters of related terms. The topics extracted encompassed a diverse range of issues. From these, we were able to discern overarching patterns and commonalities within the challenges scientists encounter when working with scientific workflow management systems.





EVALUATION

Furthermore, a comparison was made between BERTopic's results and a previous analysis using LDA techniques. Cross-referencing both models for consistency and overlap enhanced the robustness of our findings, contributing to a comprehensive understanding of SWfMS challenges.



DISCUSSION

TAXONOMY OF QUESTION CATEGORIES

Category	Description	Freq
Learning	Inquiries of group revolve around questions such as "How to do a task?", "How to manage/create/classify?", "Can someone explain?", or "Is it possible?"	125
Errors	This category seeks explanations and solutions for SW-related errors and exceptions.	80
Installation Issues	Questions in this category include phrases like "Cannot install," "Installing problem," "How to install?", or "Issues with install".	30
Discrepancy	Questions in this category aim to find explanations or solutions for unexpected outcomes, such as "What is causing the issue?", "Why is it not functioning correctly?", "Why is there a disparity in output?", or "Why is the execution not meeting expectations?"	28
Review	These questions generally seek improved alternatives in evaluating existing solutions to address a problem.	25
Theoretical	This category of questions asks about the theoretical explanations of SWs-related topics.	25
Version Issue	This category discusses the issues due to the versions of plugins, tools, or applications.	12
Database Issue	In this category, the questions are like "How to connect the database?", "How to transfer database?" or "Facing issues with particular database operation".	11
Lack of Facilities	Inquiries within this category addresses the absence of various functionalities in SWfMSs. For instance, a question highlights the lack of a Try Catch process in the Nextflow SWfMS.	9
Integrating Applications	Questions within this group revolve around combining distinct applications to serve their specific objectives.	8
Performance	Inquiries within this category focus on matters concerning performance, such as "How to improve performance?", "Issues with slow performance", or "Optimizing an operation."	7
Conceptual	Questions of this category pertain to topics like comparing SWfMSs or exploring the limitations of SWfMSs.	5

Our research reveals that the primary challenges developers face when working with SWfMSs are:

- Lack of documentation
- Third-party tool integration
- Configuration difficulties

Community support plays a significant role in mitigating these challenges. The main goals are to:

- Ensure consistent and thorough documentation
- Create user-friendly interfaces
- Foster collaboration & community-building efforts

ACKNOWLEDGEMENTS

This research was possible thanks to the NSERC Discovery Grant and the Undergraduate Student Research Award funded by the Computer Science Department at the University of Saskatchewan.

University of Saskatchewan | Computer Science

NSERC CRSNG