Analyzing User Needs at NERSC Using Natural Language Processing and Machine Learning Techniques

Martha Asare

Host Site : Lawrence Berkeley National Laboratory

Mentor(s): Dr. Lipi Gupta , Dr. Erik Palmer

Keywords: [User, Application Terms, Issue Terms, Regex Matching, Sentence embedding, Natural Language Processing, Machine Learning, IT Service Management, word frequency, topic clusters]

Abstract

The National Energy Research Science Computing Center (NERSC), a High Performance Computing facility, supports a large number of scientific projects whose users generate many support tickets due to the complexity of their tasks. Every year, NERSC receives over 6,500 user help tickets reporting issues and requesting services. In this study we explore the effectiveness of using Natural Language Processing and Machine Learning (ML) techniques on User ServiceNow ticket data to automatically identify frequent and repetitive problems that need attention. Starting with word frequency and moving to sentence embedding and clustering, we demonstrate the capability of each approach to generate actionable insights. Further development of these mechanisms will allow NERSC staff to make targeted enhancements and boost user research efficiency.

Introduction

The National Energy Research Scientific Computing Center (NERSC) is a leading high-performance computing (HPC) facility in the U.S. Department of Energy's Office of Science, making it a cornerstone of scientific research across a wide array of disciplines. NERSC supports more than 9,000 scientists globally, who are engaged in research ranging from climate science and renewable energy to materials science and astrophysics. The facility houses one of the world's most advanced supercomputers, Perlmutter, which is optimized for both traditional simulations and machine learning applications [1]. NERSC's user base is diverse, consisting primarily of researchers from universities, national laboratories, and industry partners, reflecting a wide demographic range of scientific disciplines. This vast user base benefits from NERSC's computational power, which accelerates its research by enabling complex simulations, data analysis, and artificial intelligence-driven studies that would otherwise be computationally infeasible [1].

Given the complexity and scale of projects at NERSC, users frequently encounter technical issues that necessitate support, leading to the submission of a significant number of help tickets each year. In 2023, over 6,944 tickets were submitted. To manage this influx and enhance user support Natural Language Processing (NLP) and Machine Learning (ML) techniques could be instrumental. These technologies could help identify common and repetitive issues within submitted tickets, allowing NERSC staff to make targeted improvements that enhance user satisfaction and overall research productivity. This approach may not only streamline the resolution of user issues but also help NERSC maintain its critical role in advancing scientific discovery.

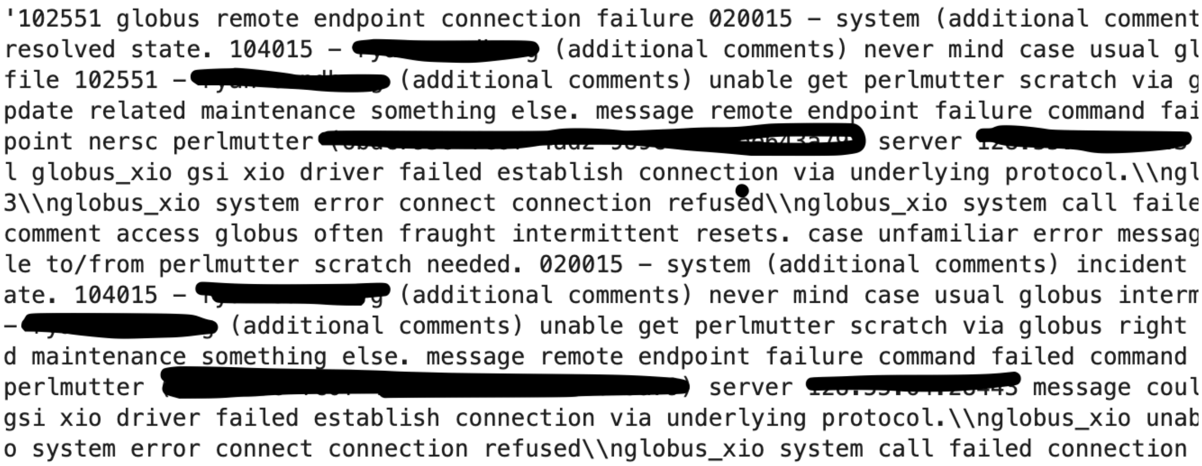
The motivation for analyzing user support tickets at NERSC stems from the need to enhance the efficiency and effectiveness of the support process for its diverse and growing user base. NERSC handles thousands of complex scientific projects annually, leading to a substantial number of user support tickets. The ability to address these tickets quickly and accurately is crucial for maintaining high user satisfaction and research productivity. However, the current system often results in delays because tickets are manually triaged and assigned to different consultants. By applying Natural Language Processing (NLP) and Machine Learning (ML) techniques, it is possible to automatically categorize tickets and direct them to consultants who have previously resolved similar issues. This not only speeds up the resolution process, but also leverages the expertise of consultants, ensuring that users receive the most relevant and informed assistance.

The analysis techniques presented in this paper have the potential to improve the NERSC user experience in other ways as well. For example, implementing ML-driven predictions could provide users with an estimate of how long it will take to resolve their issues, based on the analysis of similar past tickets. A system could inform a user, “Tickets like this usually require 4-5 days to resolve,” even before a consultant responds. This transparency helps manage user expectations and allows users to plan their work accordingly. These improvements not only streamline the support process but also empower NERSC to handle the increasing demand for computational resources more effectively, thereby reinforcing its role as a critical infrastructure for scientific research.

Data Source and Preprocessing

The study utilized user ticket data from the NERSC’s ServiceNow system covering the period from January 2023 to January 2024. The dataset included 6,944 tickets with 24 fields, out of which three text-based fields were selected for analysis:

- ***Help Ticket Title:*** A summary of the ticket.
- ***Additional Comments:*** Information provided by users to help resolve their issues.
- ***Comments and Work Notes:*** Interaction logs between users and staff, including internal staff notes.



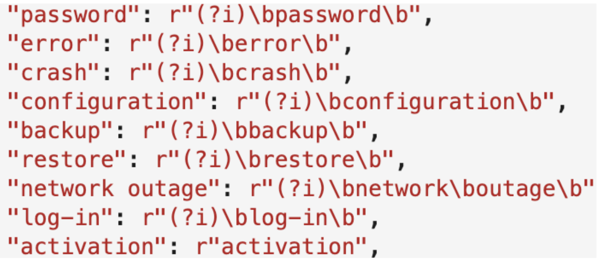
```
'102551 globus remote endpoint connection failure 020015 – system (additional comment
resolved state. 104015 – [redacted] (additional comments) never mind case usual gl
file 102551 –[redacted] (additional comments) unable get perlmutter scratch via g
pdate related maintenance something else. message remote endpoint failure command fai
point nersc perlmutter [redacted] server [redacted]
l globus_xio gsi xio driver failed establish connection via underlying protocol.\\ngl
3\\nglobus_xio system error connect connection refused\\nglobus_xio system call faile
comment access globus often fraught intermittent resets. case unfamiliar error messag
le to/from perlmutter scratch needed. 020015 – system (additional comments) incident
ate. 104015 – [redacted] (additional comments) never mind case usual globus interm
–[redacted] (additional comments) unable get perlmutter scratch via globus right
d maintenance something else. message remote endpoint failure command failed command
perlmutter [redacted] server [redacted] message coul
gsi xio driver failed establish connection via underlying protocol.\\nglobus_xio unab
o system error connect connection refused\\nglobus_xio system call failed connection
```

Figure 1:A single sample of the processed ticket-text data from the 6,944 dataset.

The preprocessing stage involved combining all text fields into a single field for analysis (see Figure.1 above). Stop words were filtered out using the Python libraries *"stopwords"* and *"WordNetLemmatizer."* The preprocessing phase also ensured the retention of the ticket creation timestamps, which are crucial for time-based analysis.

The first step involved merging the selected text fields (Help Ticket Title, Additional Comments, and Comments and Work Notes) into a single text column for each ticket. This consolidation was essential for treating the entire textual content as a single entity, thus enabling a more effective analysis.

The text was then cleaned by removing special characters, numbers, and unnecessary whitespaces. Standard NLP techniques, including tokenization and lemmatization, were applied to normalize the words. *Lemmatization* helped reduce words to their base forms, thereby minimizing redundancy in the analysis. Common stop words (e.g., "hello," "hi," "thanks") were removed to focus on the significant terms. Personal information such as user email addresses, phone numbers, and names were removed. The *NLTK library's* predefined stop word list was used, with additional customization to remove domain-specific stop words that were irrelevant to the analysis. Regex Matching was implemented using the python library *"re"* to concurrently match Issue terms and Applications terms as shown in figure 2 below.



```
"password": r"(?i)\bpassword\b",
"error": r"(?i)\berror\b",
"crash": r"(?i)\bcrash\b",
"configuration": r"(?i)\bconfiguration\b",
"backup": r"(?i)\bbackup\b",
"restore": r"(?i)\brestore\b",
"network outage": r"(?i)\bnetwork\boutage\b"
"log-in": r"(?i)\blog-in\b",
"activation": r"activation",
```

Figure 2: Partial list of regex expressions used in Jupyter Notebook for the study.

Analysis Techniques

Word Frequency Count

We examined the frequency of occurring terms in the user tickets, providing a straightforward measure of common issues. We wrote code in Jupyter Notebooks running on NERSC's Perlmutter to process the text fields exported from the ticket database. From there, we constructed a custom Python kernel which allowed us to integrate several popular data analysis libraries, such as counter, matplotlib, sklearn.feature_extractor, etc . Using these tools we counted the occurrence of issue terms and application terms in each user ticket received.

The frequencies of specific terms related to software applications and common user issues were plotted to identify patterns and prevalent problems. Beyond basic word counts, the Term Frequency-Inverse Document Frequency (TF-IDF) approach was employed to highlight words that were not just frequent, but also significant in distinguishing one ticket from another. This helped identify terms that were particularly informative and could indicate specific issues or topics. The results of word frequency analysis were visualized using bar charts and word clouds. Word clouds provided a quick visual summary of the most prominent terms, whereas bar charts offered a more detailed breakdown of the frequencies.

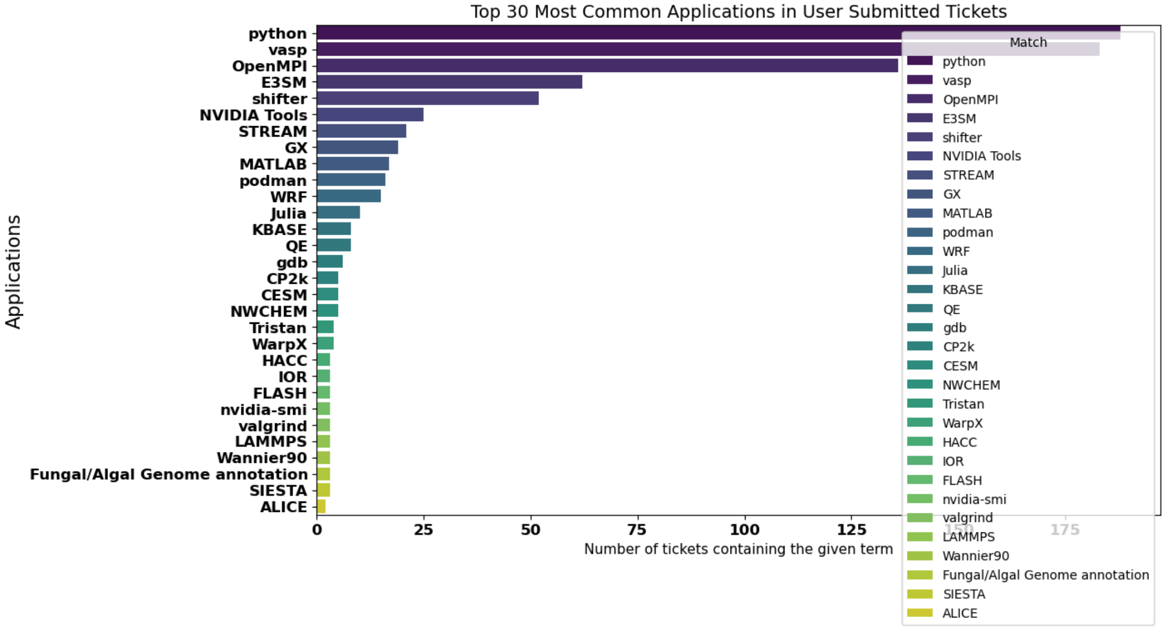


Figure 3: Bar plot displaying the occurrence of top 30 issue terms in a user submitted ticket. Each term is only counted once per ticket.

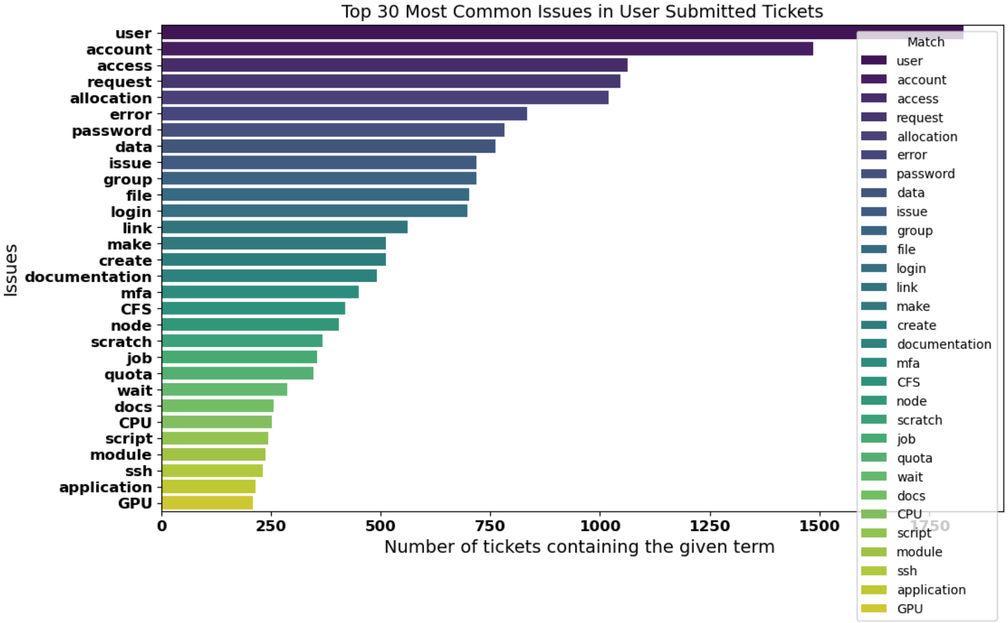


Figure 4: Bar plot showing occurrences of top 30 issue terms.

By examining the correlation of words among application terms and issue terms we can generate a 2-D matrix that represents the amount of tickets with both terms appearing together. The goal of this representation is to gain insight into what type of issues users have with each application.

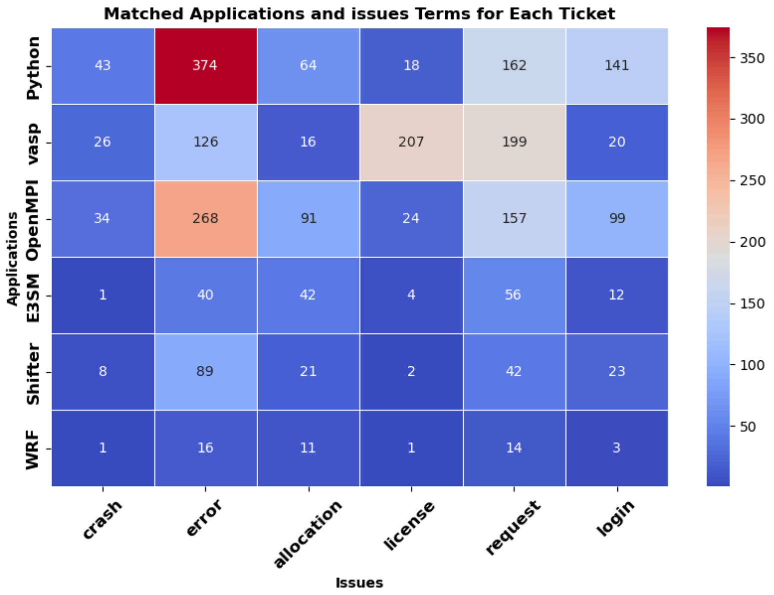


Figure 5: Heatmap displaying the correspondence between application and issue terms for the most frequent applications and issues.

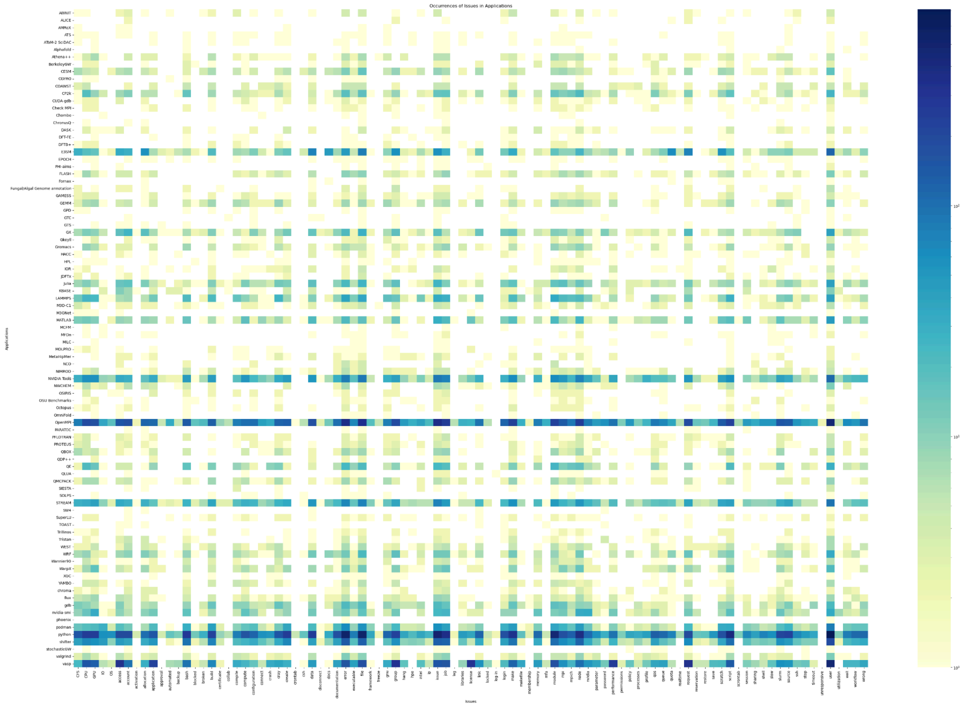
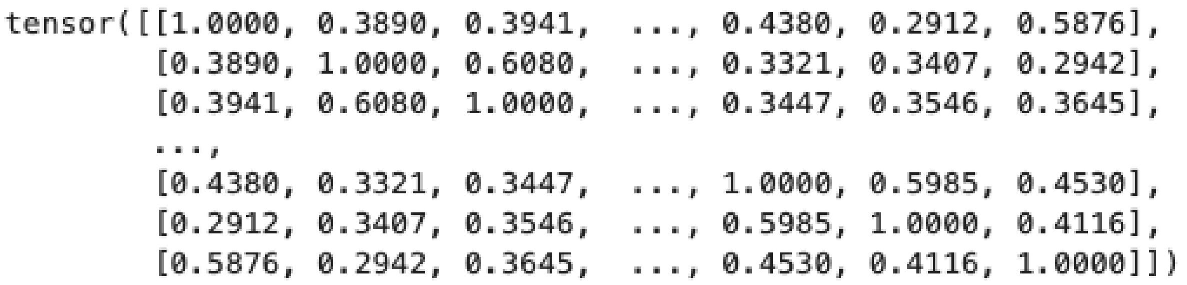


Figure 6. Heatmap of the entire data set (6,944) including all application and issue terms. Issue terms are listed along the horizontal axis, application terms are listed on the vertical axis. Darker areas represent a higher number of tickets where both the issue and the application term are contained in the ticket text. Darker lines that cross the entire plot horizontally indicate an application term that correlates with a wide range of issue terms. Darker lines that cross the entire plot vertically indicate an issue term that appears in a wide range of application terms.

Machine Learning (ML) Techniques and Natural Language Processing (NLP)

The implementation of the NLP and ML techniques required several steps, beginning with preprocessing of the data, followed by the application of word frequency analysis, sentence embedding, and topic modeling. We also explored if ML could effectively group tickets automatically. The implementation process is as follows. Sentences from the combined text fields were transformed into 384-dimensional dense vectors using the *"all-MiniLM-L6-v2"* language model[2]. This allowed for the creation of a similarity matrix using cosine distance metrics, which helped assess the similarity between different tickets. Cosine similarity was used to measure the closeness between the vectors representing different tickets. Tickets with higher cosine similarity values were considered similar in content, which helped group them into clusters using k-means.



```
tensor([[1.0000, 0.3890, 0.3941,  ..., 0.4380, 0.2912, 0.5876],
        [0.3890, 1.0000, 0.6080,  ..., 0.3321, 0.3407, 0.2942],
        [0.3941, 0.6080, 1.0000,  ..., 0.3447, 0.3546, 0.3645],
        ...,
        [0.4380, 0.3321, 0.3447,  ..., 1.0000, 0.5985, 0.4530],
        [0.2912, 0.3407, 0.3546,  ..., 0.5985, 1.0000, 0.4116],
        [0.5876, 0.2942, 0.3645,  ..., 0.4530, 0.4116, 1.0000]])
```

Figure 7: 6,944 tickets were encoded as 384-dimensional vectors using a sentence transformer with the pre-trained "all-MiniLM-L6-v2" language model. The values in the matrix represent similarly between ticket-texts.

Topic Modeling with Latent Dirichlet Allocation (LDA)

This study applied Latent Dirichlet Allocation (LDA) through the "*pyLDA*" library to generate topic clusters. This method provides a visual representation of how different tickets are grouped based on their content, offering insights into recurring themes and issues [4]. LDA was configured to generate a predetermined number of topics (30) with the number of topics tuned based on coherence scores and domain knowledge. Each topic generated by the LDA model was interpreted by examining the top associated keywords. These keywords provide insights into the underlying themes present in user tickets, allowing for manual labeling of the topics[3,4].

After calculating the similarity matrix, hierarchical clustering was applied to group the similar tickets.

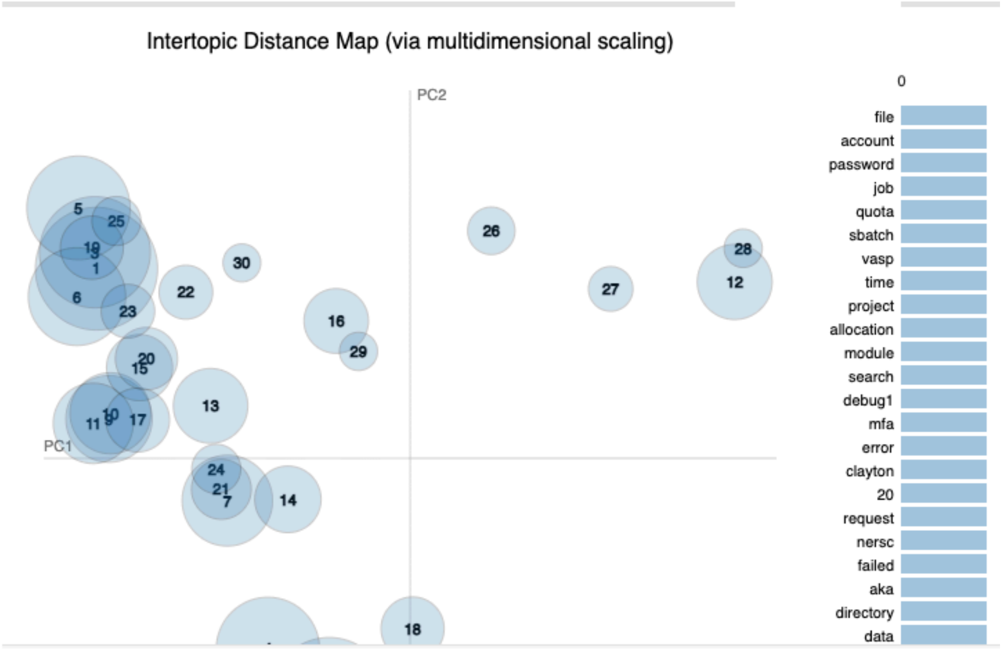


Figure 8: An Intertopic Distance Map generated with pyLDA and number of topics set to 30.

Dendrograms were used to visualize clustering, providing a clear view of the relationship between tickets (see fig.8). The clusters were analyzed to understand common issues or topics within each group. The size of each circle represents the percentage of tickets assigned to that topic group. The position of the center of each circle represents its location along principal components determined by pyLDA and is useful for interpreting how similar or distinct two topics may be from one another. Circle centers with larger distances from others, indicate a more distinct topic. This helped in identifying recurring problems that could be addressed through targeted interventions by the NERSC staff.

Conclusion

This summer 2024 study demonstrates the potential of ***NLP and ML*** techniques to improve the management of user tickets at NERSC. By identifying common issues and patterns in user behavior, the NERSC can make informed decisions to enhance its support services. The insights gained from this analysis not only help to address immediate user concerns but also contribute to the overall improvement of the research environment at the NERSC.

Key Findings

- Word frequency analysis revealed key terms related to applications and issues that frequently appeared together in tickets. By examining these correlations, this study generates a two-dimensional matrix representing the types of issues associated with specific applications. This matrix serves as a valuable tool for understanding user challenges linked to a particular software or package.
- Sentence embedding and clustering approaches have demonstrated the potential to automatically group tickets based on their content. The similarity matrix and topic clusters generated from the LDA model offered a structured way to analyze a large volume of tickets, providing actionable insights for improving user support. The use of the *"all-MiniLM-L6-v2"* language model was particularly effective in encoding tickets into 384-dimensional vectors, which were then used to visualize the relationships between different topics.

Future Work

There are several areas for future work to further refine the analysis and enhance the effectiveness of NLP and ML techniques. These areas includes:

1. **Improving Text Data Cleaning:** There need to enhance the preprocessing steps to remove personal and system-specific information that may not contribute to the analysis.
2. **Exploring Advanced Clustering Methods:** Future research could explore different clustering and topic modeling techniques, especially those suited to high-dimensional datasets, to improve the accuracy and interpretability of the results.
3. **Training ML Algorithms:** This study suggests training Machine Learning algorithms specifically for sentence transformation and clustering tasks to better capture the nuances of user tickets.

4. **Automated Labeling with Large Language Models (LLM):** Implementing a large language model to automatically label each topic cluster can significantly reduce the manual effort involved in categorizing tickets, making the system more efficient.

Acknowledgements

I would like to thank Sustainable Research Pathways (SRP) and the Lawrence Berkeley Lab's Computational Research Division for the Summer Student Program. Special thanks to Lisa Bruzdzinski and her team for meticulously planning and executing an extensive and inclusive summer program. The sense of community fostered among the summer visitors, especially through the weekly slack updates, greatly enriched my experience and allowed me to connect with fellow students.

Heartfelt thanks Dr. Rebecca Hartman-Baker and the Computer Science Mentoring Program Committees for organizing an inspiring speed-mentoring program. It was invaluable to hear staff members share their career journeys and motivate us to pursue our own paths.

I am also grateful to the NERSC Administration Team. Every person I met at NERSC was not only exceptionally talented with a wealth of knowledge and skills but was also warm and approachable. Dr. Brandon Cook, the Programming Environments and Models Group Lead, created a welcoming environment that valued everyone. During my final week, I was honored to share my summer project at a team meeting, a moment that made me feel truly heard and appreciated.

I owe special thanks to my outstanding mentors, Dr. Erik Palmer and Dr. Lipi Gupta. Their patience and dedication were instrumental in guiding me through the various coding processes. They were always there to tackle my questions and explain complex concepts, ensuring that I understood every step. Their advice to "always ask for help" was a constant reminder that support was just a Slack message away. They welcomed me as part of the team from day one, and their mentorship pushed me to grow in ways I had never imagined. Dr. Erik Palmer kept me engaged with a range of tasks that challenged me to expand my skills. Thanks to their faith in me, I accomplished more than I thought possible. Their kindness, belief in my potential, and ongoing encouragement will remain with me as I move forward in my career.

References

[4] National Energy Research Scientific Computing Center. (n.d.). NUG Annual Meetings. https://www.nersc.gov/users/NUG/annual-meetings/nug-2/

[1] Sentence-Transformers. (n.d.). *all-MiniLM-L6-v2*. Hugging Face. Retrieved July 24, 2024, from https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] Grootendorst, M. (n.d.). *BERTopic. GitHub*. Retrieved July 24, 2024, from https://github.com/MaartenGr/BERTopic/blob/master/README.md

[3] Sievert, C., & Shirley, K. (2014, June). LDAvis: *A method for visualizing and interpreting topics.* In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).