**EXECUTIVE SUMMARY**

**Natural Language Analysis of Build a Better Grinnell Survey Data**

Shreyas Agrawal, Jiayi Chen, David Dai, Hong Doan

Spring 2023

"Build A Better Grinnell" is a community-based initiative focused on enhancing the quality of life in Grinnell, Iowa. Our project aims to gather valuable insights from Grinnell residents, collect innovative ideas from the community, and establish a shared understanding of strengths, visions, and priorities. Our ultimate objective is to gain a deeper understanding of the current needs of Grinnell residents, identify what they value about their community, and pinpoint the key issues that require attention to make Grinnell an even better place to live.

Two datasets were provided: a raw survey dataset with responses to six strength-based questions, five needs-based questions, and optional demographic data; and a dataset comprising frequency counts for issues identified from the needs-based questions.

Since the demographic question, which asks respondents to share information about themselves or their groups, is optional and open-ended, we encountered challenges in extracting useful demographic information from this question. Instead, we employed alternative approaches by analyzing the remaining 11 questions to uncover patterns such as "People who enjoy XXX in Grinnell are also frustrated about XXX in Grinnell."

We started the analysis by creating word clouds and word frequency plots for each individual question in the entire survey. However, after removing stop words and conjunctions, we found limited valuable information. The most frequently appearing terms were related to the community, Grinnell, and the town itself. To gain a clearer understanding, we examined the words occurring before and after the most frequently appearing terms. Nonetheless, manually reading the output was necessary to obtain a comprehensive understanding of the Grinnell residents' perspectives. To automate this process, we employed various clustering methods. Specifically, we utilized K-means clustering and PAM clustering to identify clusters or main themes based on respondent answers for each question, including both strengths and needs questions. Our analysis revealed that K-means clustering provided more distinct clusters compared to PAM clustering. We used the elbow method and advice from our mentor and decided to set our clustering number as 5.

We also performed sentiment analysis. By utilizing packages from the takeover team, we established a positive response threshold of 0.95. This threshold indicates that responses with a score above 0.95 are considered positive, scores below -0.95 are considered negative, and scores falling outside these ranges are considered neutral. We conducted sentiment analysis for each question, generating a matrix with all the responses and their respective scores. We then combined two questions to uncover any interesting insights. For instance, exploring negative responses to both positive-oriented questions might yield valuable findings. However, we acknowledge that this grouping and stacking method is not tailored to our client's specific needs. Thus, the methods tried that do not meet clients’ requirements are not listed in the final deliverables folder. While for the sentiment analysis, the clients found the overall sentiment scores might be useful since they provided them with an overarching view of the respondents' perspectives.

Finally, we received an updated frequency count from our clients which highlighted 35 prioritized issues identified by Grinnell respondents and the project coordinator. To address this update, we created a keyword list based on the identified issues. Employing k-means clustering with specific strength-based questions, we sought to identify the most representative groups. This approach effectively demonstrates which groups of people who appreciate certain aspects of Grinnell have also identified particular issues.

However, we note certain limitations. Sentiment analysis can be skewed by strongly negative words, regardless of their context. For example, the outcomes of sentiment analysis are exemplified by situations where a response with a highly negative word like "disaster" can lead to the algorithm categorizing the entire response as negative, disregarding the positive context it may have. The reason is that the strong negativity of the word overrides the overall positive context of the response.

Another limitation is the unconventional way we used to choose the optimal number of clusters in k-means clustering. Due to the nature of our dataset, using the number of clusters determined by the elbow method would result in indistinguishable overlapping clusters. Therefore, after consulting Professor Miller, we opted to manually select five clusters, a strategy that led to more distinctive results, even though it deviates from the conventional optimal point typically indicated by the elbow method.

Future research may include social network analysis or creation of a deep learning model predicting respondents' needs based on their perceptions of Grinnell's strengths and their vision for Grinnell. Additionally, future surveys should be conducted with more categorical questions instead of open ended optional questions.

In conclusion, our analysis offers a holistic understanding of the Grinnell community's needs and sentiments, using various data analysis techniques, such as word analysis, clustering, and sentiment analysis. We hope our findings can contribute significantly to improving the quality of life in Grinnell.