Slide 1:

Good morning everyone. It is my pleasure to present today as a graduating student of Data Science. My name is Jaehee Lee, and I would like to take this opportunity to present my research on "Topics Detection Using Dowker Complex," under the guidance of Dr. Robinson. While the research is still in progress, I am thrilled to share with you the progress we have made so far.

Slide 2:

Imagine walking into a library where books are scattered around with no proper organization. It would take an incredibly long time to locate a specific book or you might end-up leaving empty-handed. However, if the library categorizes their books by topics and clearly labels them, it would be much more efficient to find the book you desire.

Slide 3:

In the United States, the Dewey Decimal Classification and Library of Congress Classification are the widely used classification systems in libraries. Back in the day, libraries used to manually classify books by referring to classification tables. It involves significant manual labor and librarians deep understanding of the subjects. With the rise of vast amounts of text data, organizing them becomes crucial to efficiently explore and find the required information. You may have come across TF-IDF, which stands for Term Frequency-Inverse Document Frequency, a widely-used method to determine the importance of a word in a document. However, in some cases, TF-IDF may not be sufficient. In my research, I have used the Dowker Complex method to classify books into specific topics based on relevant and common terms. This method proves to be effective in cases where traditional methods may not be sufficient.

Slide 4:

Before discussing the research, I would like to give you a brief description of the Dowker Complex, it was introduced in 1951 by Yael Dowker. Dr. Robinson and I , like others, believe that the Dowker Complex can be potentially used in many areas such as Mathematics and Data Science. The Dowker Complex is an abstract simplicial complex and in this research we used it based on word usage among documents. When I apply the Dowker Complex, it is modeled in a matrix format to represent the relationship between terms and documents.

Slide 5:

Before talking about my research, I would like to compare the Dowker Complex method with other standard methods. The results of our findings reveal that the Dowker Complex can be more effective than TF-IDF. Even though TF-IDF is widely used, it has limitations because it computes documents' similarity only based on its word count. Indicating that TF-IDF cannot group documents based on its relevant terms. However, the dowker complex can classify documents by sets of relevant terms. Conversely, we can find documents based on terms that occur together.

Another popular method is the K-Nearest Neighbour method. This is used by identifying the K closest neighbors to a new data point, it can generate forecasts for the new datapoint. KNN does not work well with textual data. If KNN does detect relevant geometric features then so will the Dowker Complex. This project results show that applying Dowker Complex to textual data is useful.

Logistic Regression is used a lot but the assumptions behind is that between predictors and response variables is linear, but in the complex data this may not always be the case. Can analyze complex data including those that may not be linear structure. It can overcome non-linearities in the data

Slide 6:

When conducting data science research, one of the crucial steps is to provide a clear description of the data being used. For my research, I selected the Gutenberg data which comprises over 60,000 books, from which I randomly sampled 100 books. To ensure that I had a balanced representation of different topics, I chose to explore five main areas: politics, art, biology, cookery, and travel, and randomly selected 20 books from each of these topics. To obtain a representative sample, I used the stratified random sampling method.

Slide 7:

For those interested in replicating this research, here are the steps that I took before applying the Dowker Complex. First, I downloaded 100 books from gutenberg library. Then I combined all of these books. Second, I converted this text data into a corpus. Then I cleaned the corpus which involves removing stop words and numbers, white spaces, as well as, changing upper case characters into lower case characters in order to reveal the plain text that describes the contents of the books. Then I applied the Terms Document Matrix which describes the frequency of terms that occur in a collection of documents. Then I identified non zero values. Lastly, I applied the Dowker Complex Function in R.

Slide 8: As this table shows, the dowker\_nest function organizes books according to the behavior

Slide 9: The result that I have got is very interesting. As the boxplot for each subject shows, the fact that the probability of each subject is higher than that of other subjects shows that the dowker complex is effective. This result of the topic detection using the dowker complex shows that it separates the documents by their topics as measured by topics’ probability.

Slide 10:

For the future direction of this research is that this can be applied to help search engines rank documents by relevant terms. Same technique is used right now as part of a file behavior based search engine that Dr. Robinson is working on and some of his same code is being used in that project. Along with this, this research project can be useful to researchers and librarians who may want to classify big and complex textual data into specific categories or topics.