Topics Detection using Dowker Complex

American University
Researcher: Jaehee Lee
Advisor: Dr. Robinson

Introduction

Imagine going to a library where all the books are not organized like this...

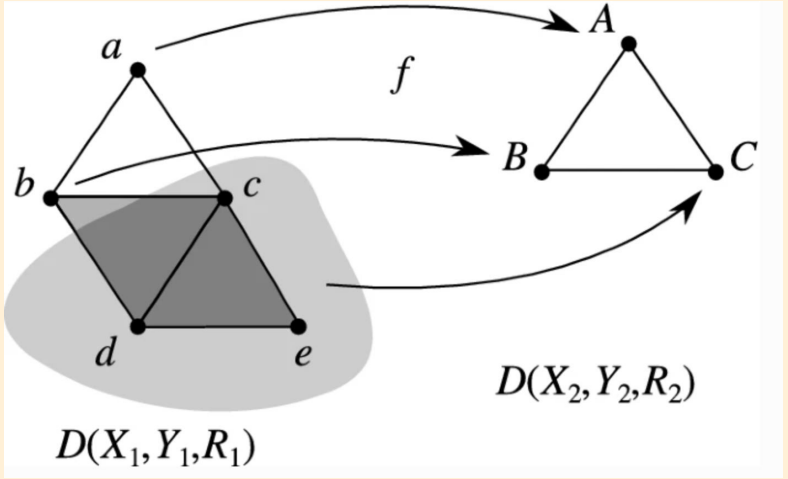
Research Goal

- Librarians traditionally classified books by hand using classification tables.
- ➔ Dewey Decimal Classification (DDC)
  Library of Congress Classification (LCC)

- Research Goal is to use the Dowker Complex method to classify documents into specific topic categories based on relevant/common terms.

Dowker Complex

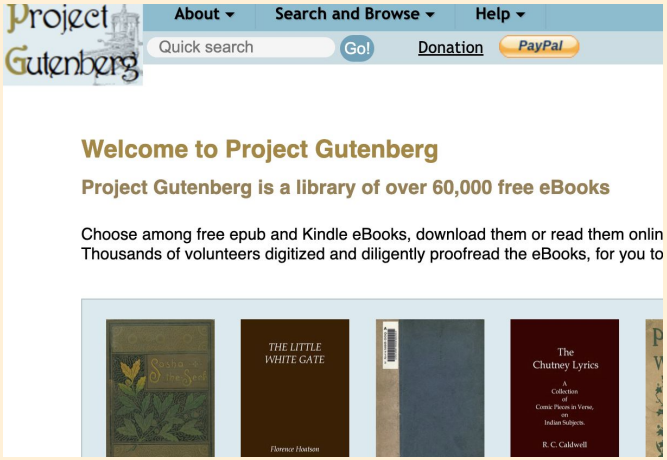
- In 1951, Dowker introduced
- Potentially used in many areas such as Mathematics and Data Science (Ghrist, 2014)
- The structure of an abstract simplicial complex
- Used the Dowker Complex based on word usage among documents
- In a matrix format to represent the relationship between terms and documents



[Cosheaf representations of relations and Dowker complexes, Dr. Robinson, 2022]

Data

- Data: Gutenbergr library in R
- Sample Size: 100 books
- Sampling Method: Stratified random sample
- Topics: Politics, Art, Biology, Cookery, Travel



Procedures

01 Download Data

Download 100 books from Gutenberg library

02 Convert

Convert text to a corpus

03 Clean Corpus

Remove stop words, numbers, etc.

04 Apply TDM

Apply Term Document Matrix (TDM)

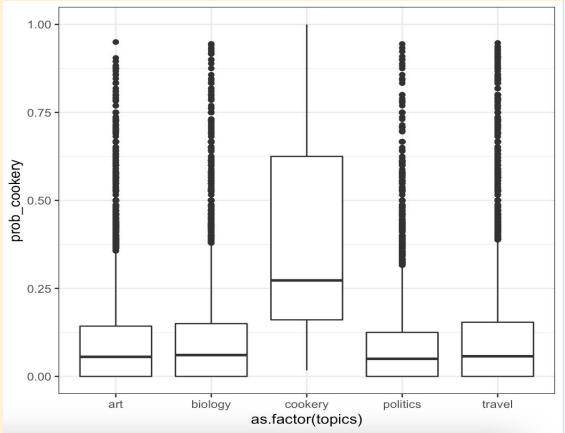
05 Identify

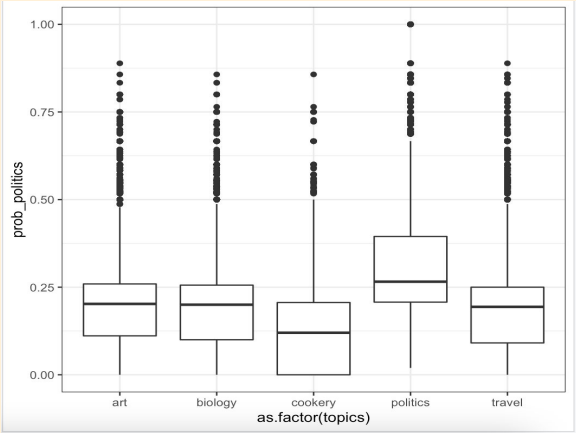
Identify Non-zero values

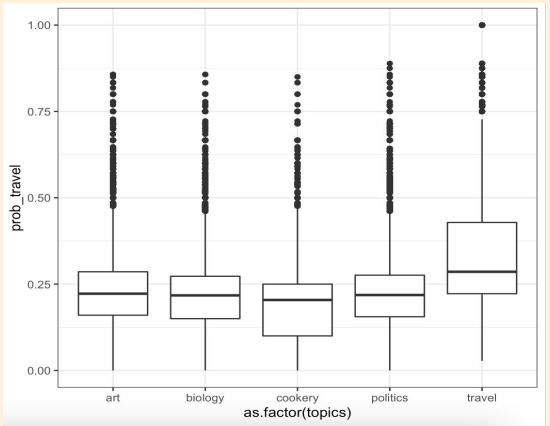
06 Dowker Complex

Apply the Dowker Complex

Results







The Dowker Complex separates the documents by their topics, as measured by the topic's probability

Comparison with Standard Methods

TF-IDF	K-Nearest Neighbors	Logistic Regression
TF-IDF computes documents' similarity **only based on its word count.** TF-IDF cannot group documents based on its relevant terms	By identifying the K closest neighbors to a new data point, it can generate forecasts for the the new datapoint. KNN **does not work well with textual data**	Logistic Regression assumes that between predictors and response variables is **linear**, but in the **complex data this may not always be the case**
Dowker Complex: • It is classifying documents by sets of relevant terms. • Conversely, we can find documents based on its relevant terms	Dowker Complex: • Dowker Complex can extract topological features from textual data by capturing the structure of the data • This project results show that applying this to textual data is useful.	Dowker Complex: • Can analyze complex data including those that may not be linear structure. • It can overcome non-linearities in the data

Conclusion

- This study can be useful to researchers and librarians who may want to classify big and complex textual data into specific categories or topics.
- This study also suggests that Dowker complex is useful in textual data analysis, also indicating that there are many things to discover in textual data using topological data analysis

References

Photos:

- https://www.istockphoto.com/photos/messy-library-book-stack
- Presentation format using Slide go

Research Paper:

- Robinson, M. Cosheaf Representations of Relations and Dowker Complexes.


Thank you