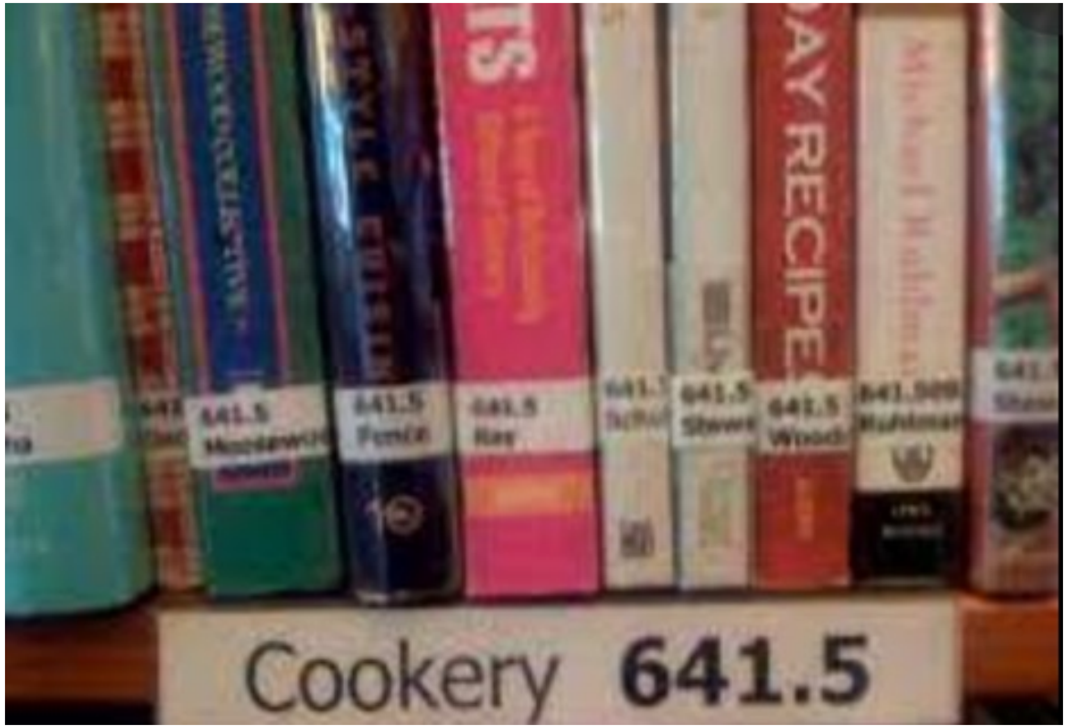
CLASSIFYING TEXT DATA - DOWKER COMPLEX

RESEARCHER: JAEHEE LEE

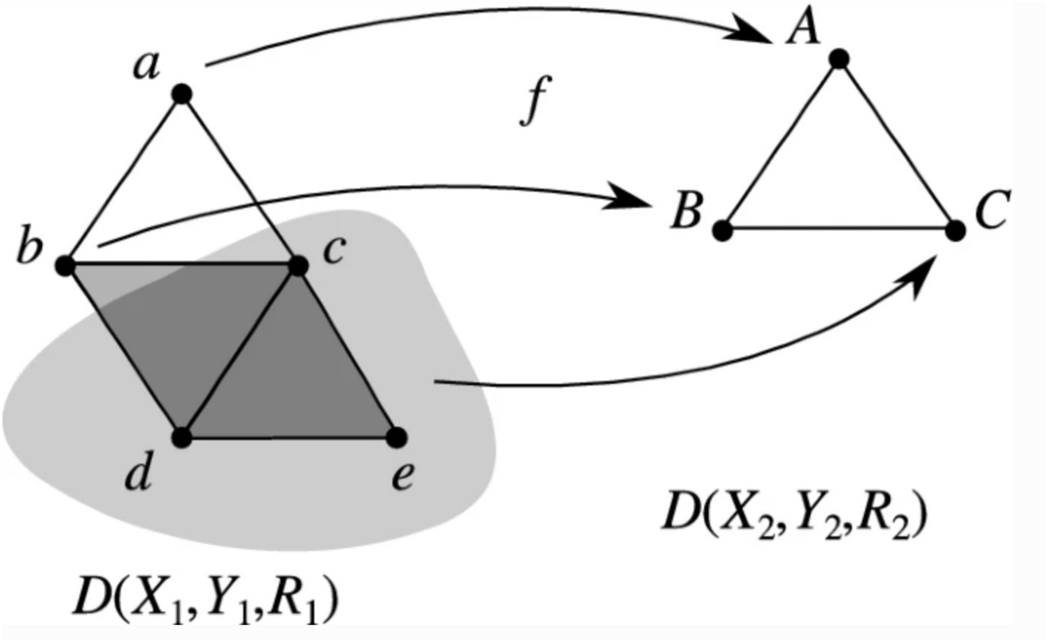
ADVISOR: DR. ROBINSON



INTRODUCTION

- Dewey Decimal Classification (DDC)
- **Term Frequency-Inverse Document Frequency (TF-IDF)**: Calculate the importance of a word in a document
- In some cases, TF-IDF is not effective
- The **Dowker Complex** -> Explore how relevant the terms are between the documents.

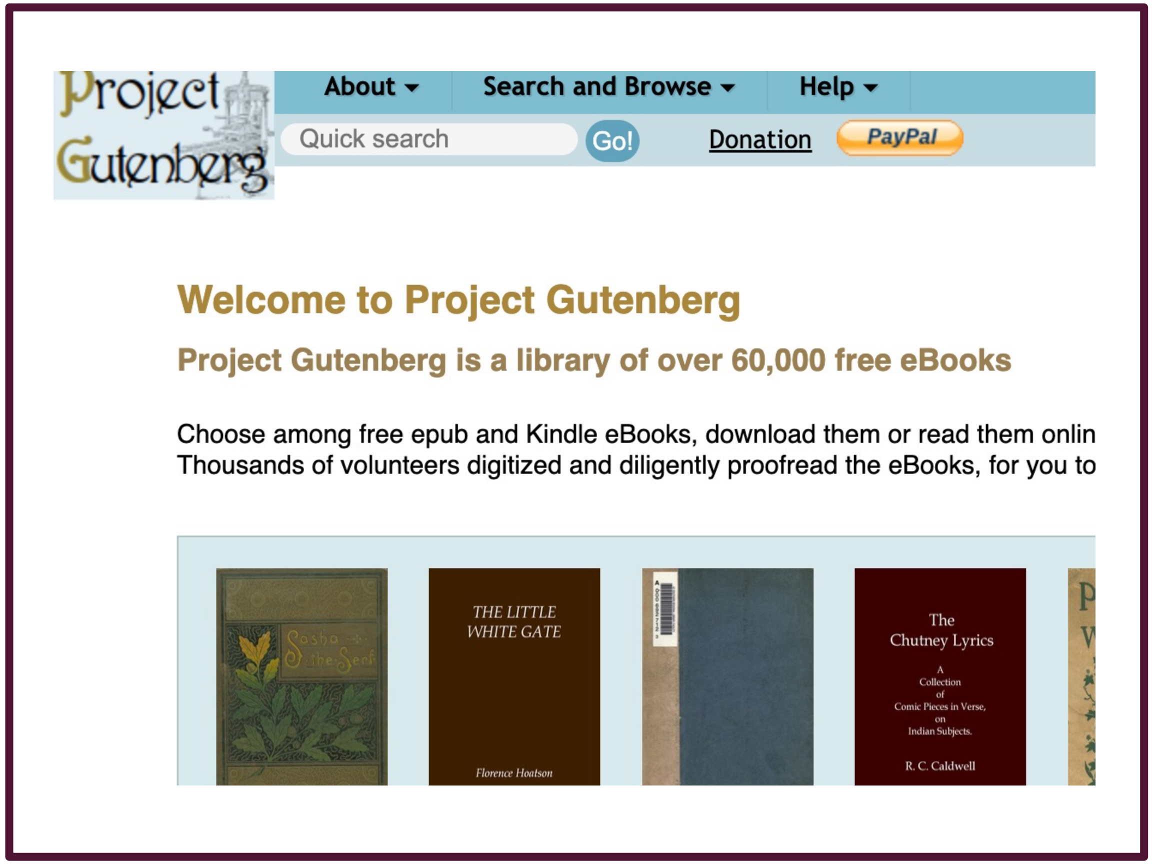
DOWKER COMPLEX



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

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

RESEARCH MATERIALS



Welcome to Project Gutenberg

Project Gutenberg is a library of over 60,000 free eBooks

Choose among free epub and Kindle eBooks, download them or read them onlin
Thousands of volunteers digitized and diligently proofread the eBooks, for you to

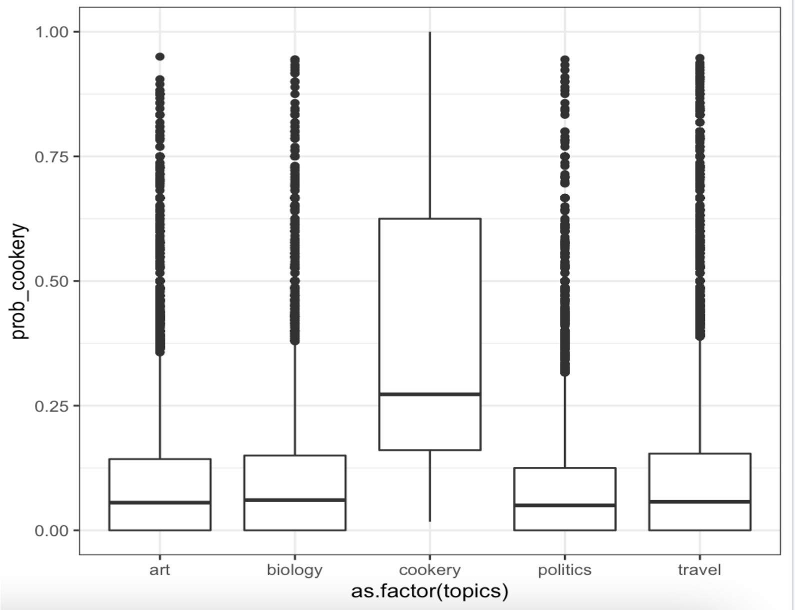
- Research Goal: Classify documents into specific topic categories based on relevant/common terms
- Tool: R
- Method: Dowker Complex
- Data: gutenbergr library
- Sample Size: 100 books
- Topics: Politics, Art, Biology, Cookery, Travel

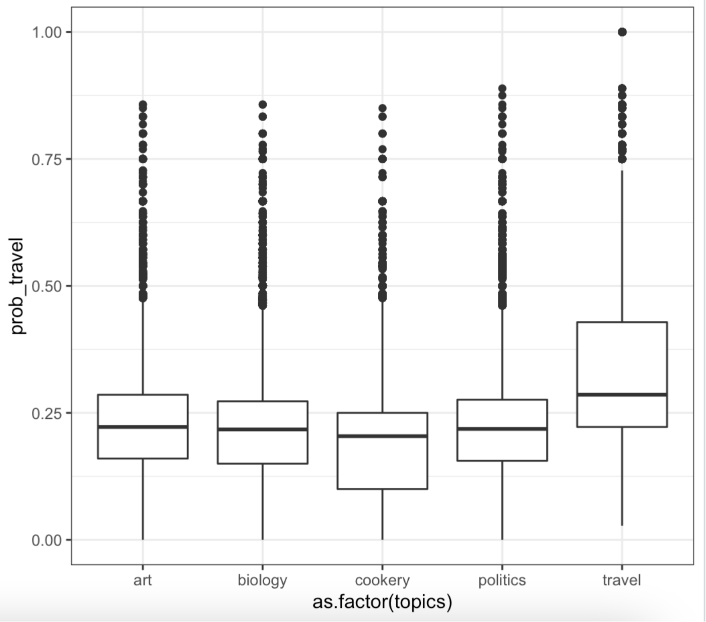
PROCEDURES

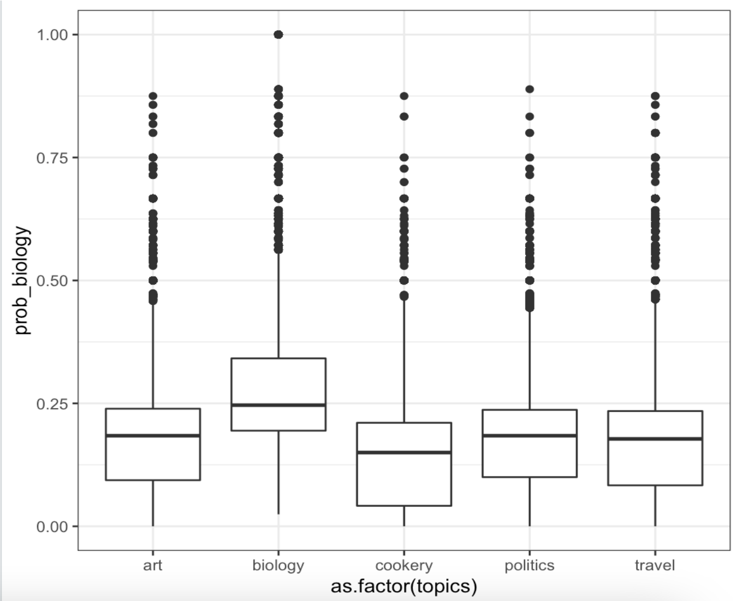
1. Gutenberg library download 100 books (politics 20 , art 20, biology 20, cookery 20, travel 20)
2. Convert text to a corpus
3. Clean corpus
4. Apply TDM (Terms Document Matrix)
5. Identify non-zero values
6. Apply Dowker Complex Function designed by Dr. Robinson

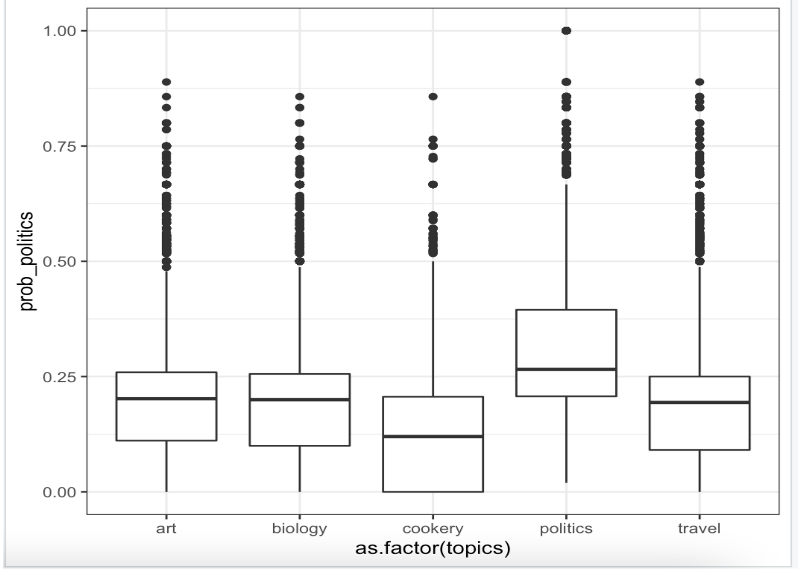
RESULTS

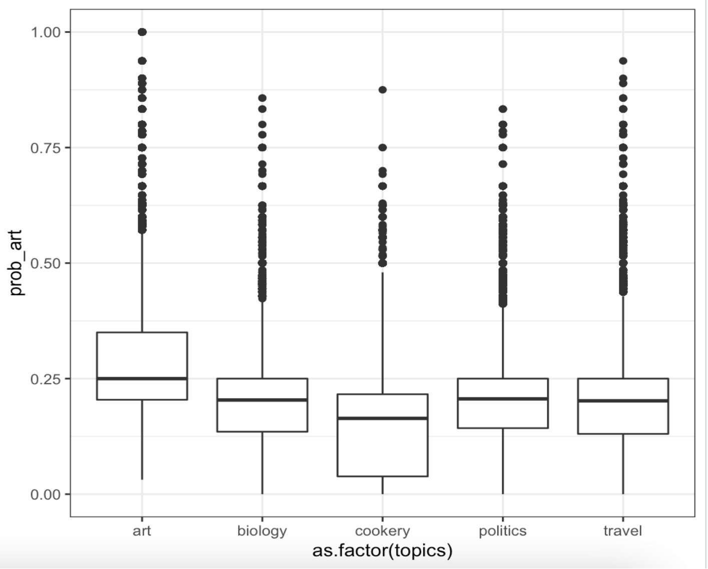
The current results show that the Dowker Complex separates the documents by their topics, as measured by the topic's probability, more efficiently than TF-IDF.











FUTURE DIRECTION!

- This research can be applied to help search engines rank documents by relevant terms.

REFERENCES

1. **Robinson, M. Cosheaf Representations of Relations and Dowker Complexes.**
2. https://www.istockphoto.com/photos/messy-library-book-stack

Q & A

THANK YOU