Classification of Children’s Literature

By: Shenyi Xie, Chang Liu, Yixin Pan, William Huang Project Advisor: David Brown

Background & Introduction

Caroline Hewins is a notable 19th-century American librarian who made substantial contributions to children’s library services. She is considered the founding mother of children’s literature categorization; she provided children’s literature categories to thousands of books.

The goal of our research is to:

- **Analyze if the genre categories identified by Hewins align with the unsupervised identification of document groupings.**
- **Find the main differences between Hewin’s classification and unsupervised classification and understand what might have lead to these differences.**

Data Overview

The data used in our research, provided to us by Rebekah Fitzsimmons of CMU’s Heinz College, consists of cleaned text files derived from 1075 books, and metadata detailing each book, including the author, publication year, and the genre assigned by Hewins. Specifically, the books are classified into 16 genres:

Genre name	Genre Descriptions	Texts
History	History, Historical Biography, Tales, and Novels	267
Home	Home and School Life	235
Travel	Travel and Adventure, Imaginary Voyages and Stories of Various Countries	235
Science	Science	111
Poetry	Poetry and Selections for Reading and Speaking	56
Myths	Myths, Legends, and Traditional Fairy Tales	49
Reference	Reference Books and Literary Miscellany	33
FairyTales	Modern Fairy Tales	27
Example	Counsel and Example	11
Art	Arts and Manufactures Books	10
Farm	Farming, Gardening, Plants, and Trees	9
Draw	Drawing and Painting	9
Outdoor	Out-door Sports	8
Amuse	Household Arts and Amusements	7
Health	Health and Strength	6
Music	Music	2

Methods

- Extract 25 topics from all documents using LDA.
- Generate a network with documents as nodes and the number of shared topics among nodes as edges
- Cluster network based on edge betweenness: edge betweenness of an edge (u,v) measures the proportion of shortest paths between any pair or nodes that pass through (u,v). The edges with higher edge betweenness are more likely to connect separate clusters.
- Affinity Propagation Clustering: Clustering based on the ‘message passing’ between data points without predetermined the number of clusters. Utilized word embeddings extracted from a purpose-built model trained on 19th century literature.

Analysis & Results

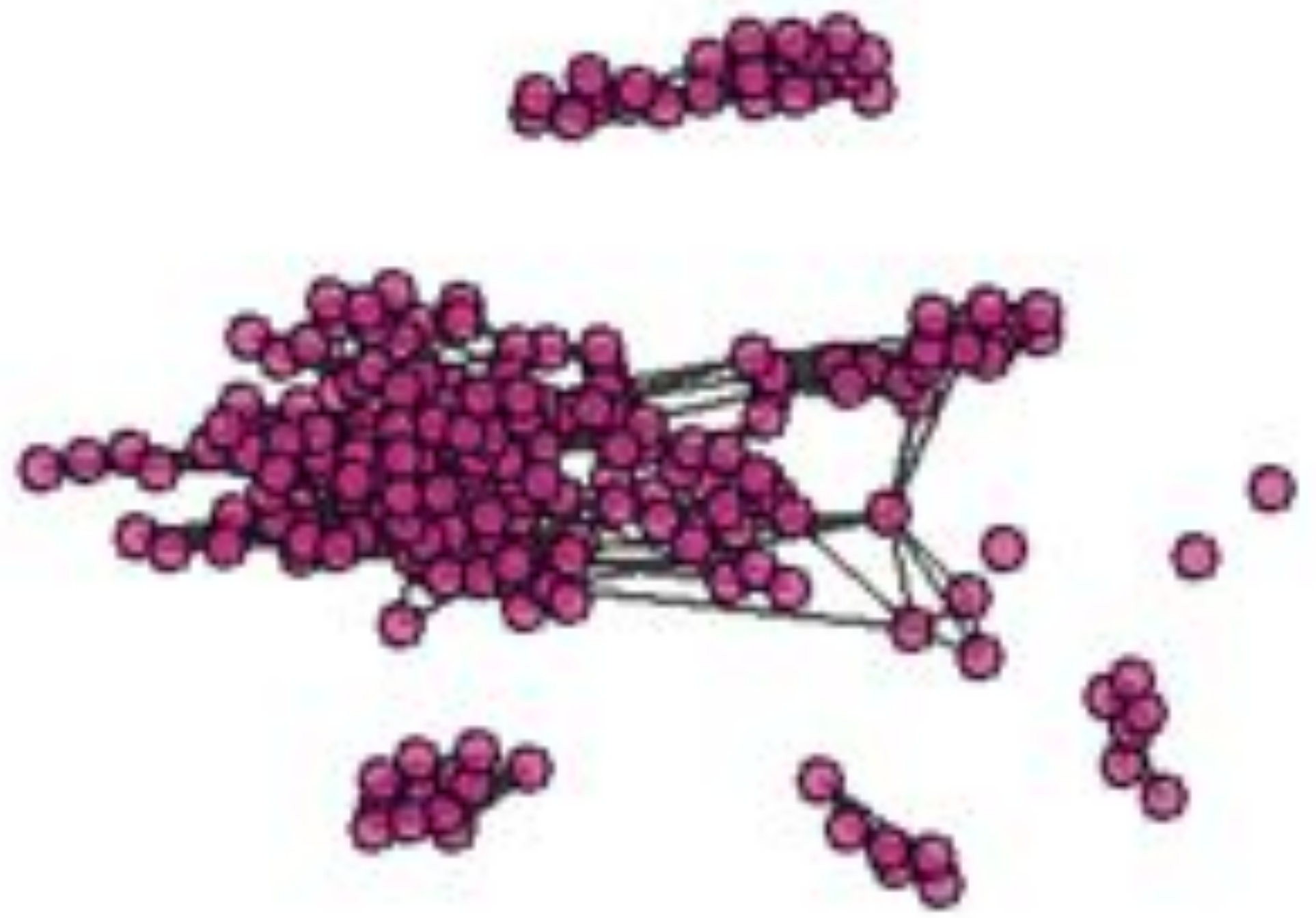

Figure 1: HOME genre network

- As an example, the generated network for the HOME genre produces a scatter of nodes (texts) with edges (shared topics) between them.
- With regards to outliers, we found that a group of 22 books, all written by Abbott, stood out from the rest. The oddity of the texts signifies the author’s distinct style of writing. A potential explanation is Abbott’s intention of a younger audience when writing.

- Edge betweenness clustering on the network found 29 clusters. Figure 2 takes cluster 12 as an example.
  - Most documents in Cluster 12 are from the HOME Category.
  - There are documents from other categories. The works of Dickens, even though all categorized into TRAVEL by Hewins, are identified to have closer relationships with texts in the HOME category.

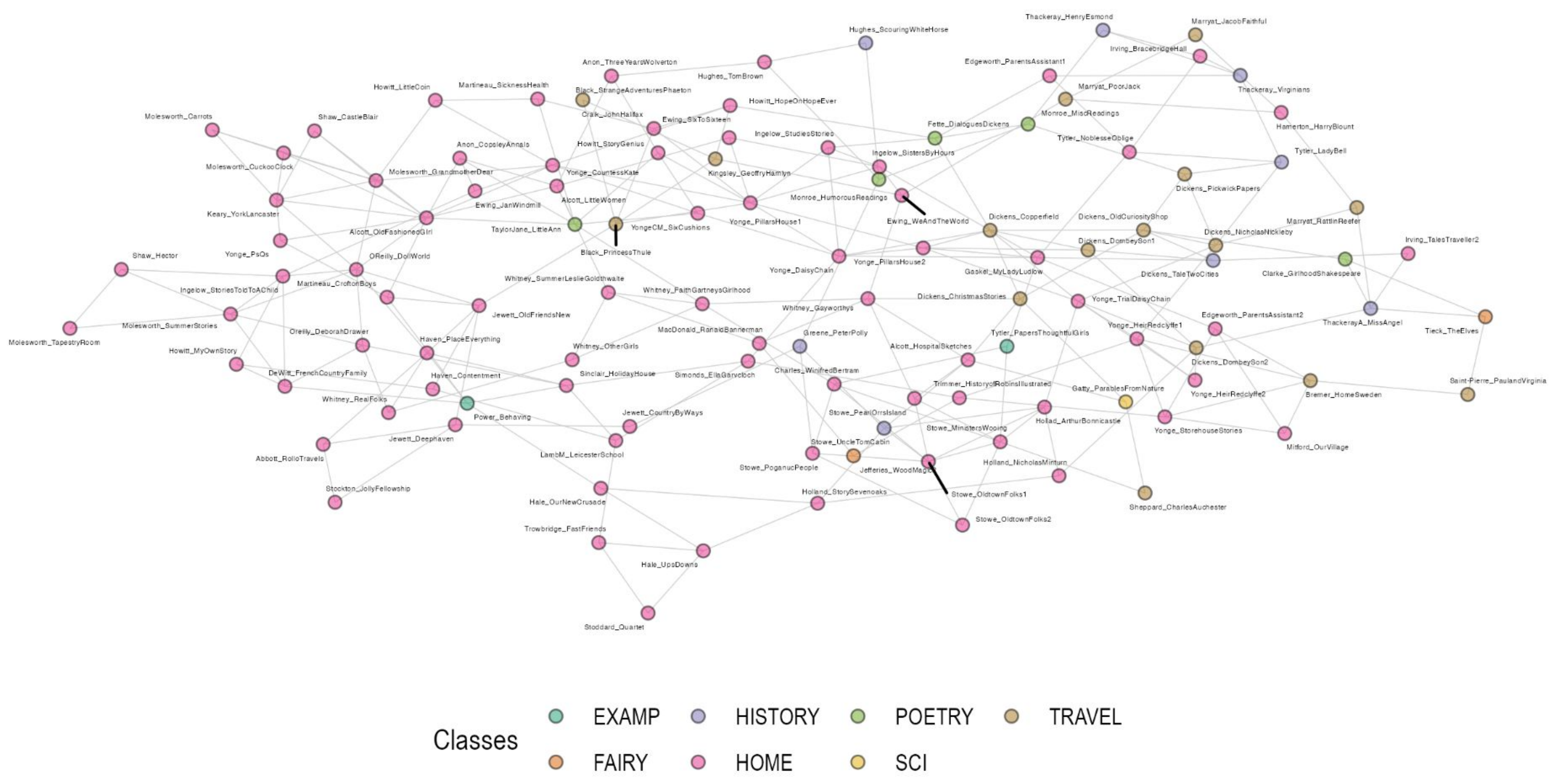

Figure 2: Cluster 12 from edge betweenness clustering

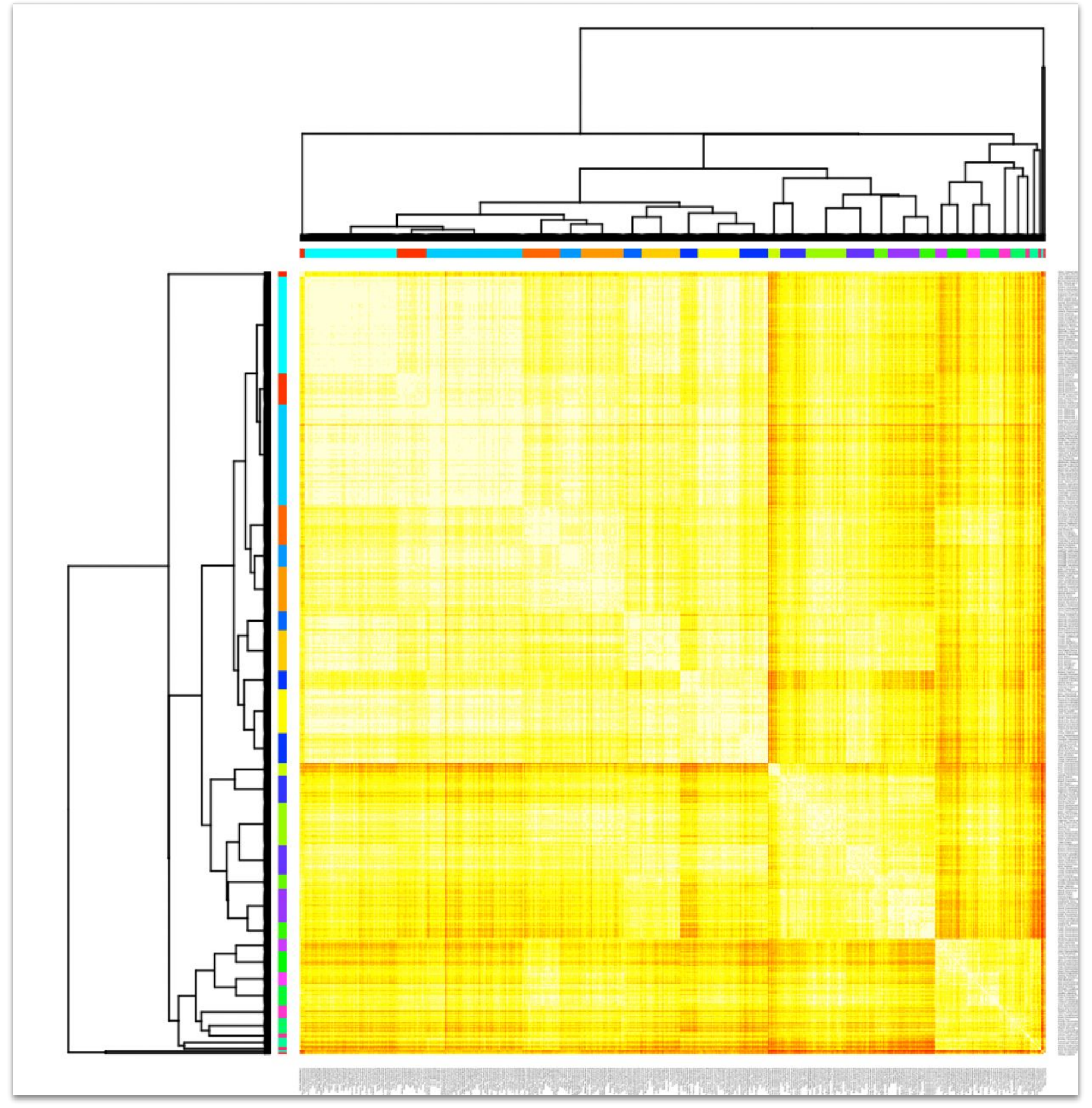

Figure 3: Heatmap and dendrogram with 30 clusters

- With affinity propagation clustering, the similarities between clusters are found from the ‘ground up’ using the word embeddings and negative squared distances as the standard similarity measure.
  - Between clusters: Most clusters are similar with each other as different clusters have texts belonging to the same genres.
  - Within clusters: There are interactions between genres. Not a single genre is purely dominant within one cluster.
    - Drawing and Poetry have little to no shared overlap with other genres.
    - 30% of clusters have History as the dominant genre, followed by Home, Travel, and then Science.

Conclusions

Genres identified by Hewins generally align with the results from the unsupervised classification, but some groups of outliers appear. Specifically, works of Dickens and Abbott are identified differently using unsupervised classification.

References

- Cluster_edge_betweenness:https://ieeexplore.ieee.org/abstract/document/6019678
- B. J. Frey and D. Dueck (2007). Clustering by passing messages between data points. *Science*, 315:972-976. DOI: 10.1126/science.1136800.