Books for the Young (1882) by Caroline Hewins was a guide for children’s literature that contained a list of 1005 titles recommended for developing children. It was assembled before the idea of “children’s literature” being culturally codified as an accepted genre. Hewins and her work is largely considered to be one of the pioneers of children’s literature and continues to stand as an influential figure in the American Library Association (ALA).

The goal of the study is to analyze Hewins’ classification of the books and ultimately compare the underlying social standards in 19th-century literature with today’s. Her list included not only the titles she considered suitable as “children’s literature” but also other metadata. One additional piece of information that she included was the gender she imagined as appropriate: “Boys and Girls”, “Boys” or “Girls”. This poster will thus primarily focus on analyzing gender readership with regards to text topic and type.

Data

Our dataset consists of 1092 individual book text files along with a spreadsheet of 1) characteristics such as gender of author(s), year of publication, etc. 2) Hewins’ classification of each text file, such as readership regarding children of different genders, genre, etc.

Preprocessing Pipeline

1) Text Cleaning or pre-processing of corpus, including omitting front- and back-matter, and correcting OCR errors;
2) DocuScope Tagging to investigate text-type variation;
3) Topic Modeling, which followed Jockers’ procedure of part-of-speech tagging texts, splitting them into 1000-word chunks, and applying Latent Dirichlet Allocation (LDA) to each;
4) PCA for dimension reduction on both extracted topics (by means for each text) and DocuScope frequencies (normalized per 100 tokens); and
5) Agglomerative Hierarchical Clustering (AHC) to further explore how texts group by DS frequencies.

Analysis and Results

Given that metadata our team set out to explore the relationships among those categories and variables extracted using 2 techniques: Topic Modeling and DocuScope (a tagger that assigns words and phrases into rhetorically oriented categories).

I. Topical Modeling

Notable Findings:
- Books imagined for girls are all at the negative end of PC2, which is comprised of topics related to domesticity, and none of the positive end of PC2 which is made up of topics related to exploration and conquest.
- Books imagined for boys contain more variation in the topics.

II. DocuScope Tagging

Notable Findings:
- Principal Component Analysis (PCA)
  - While articles for boys (green) were spread out among different clusters, articles solely for girls (yellow) centered around the 5th cluster.
- Cluster 5 - the majority of files with Readership as Girls (~63% high in Interactive, Description, and Positive)
- InformationChangePositive
- It shows that in Hewins’ corpus, for texts imagined for an audience of exclusively girls, the text-types have less variation in their cluster locations.

Citations: