Topological Data Analysis
Group project:
Text 2

Marko Prelevikj
63130345
mp2638@student.uni-lj.si

Bozhidar Stevanovski
63190410
bs9682@student.uni-lj.si

January 15, 2020

1 Introduction

One of the earliest known literary works is *Epic of Gilgamesh* dating from cca. 2150 BC. Today, after many centuries, the literature evolved and expanded, and now includes even a variety of literary genres.

Although the Sumerians/Babylonians did not live in a time when computers are available in order to write about the Gilgamesh's tale on one, today we have a lot of literary works in a digital format, enabling whole new fields which work with them, such as Natural Language Processing (NLP). One of the recent advances in the NLP field was due to the text embedding methods, which have shown to be able to effectively capture underlying semantic and syntactic information.

In this project we use persistence homology on sentence-level text embedding to explore differences in literary genres.

2 Methodology

Our method follows the basic principles of topological data analysis: picking a dataset, preprocessing it, building a complex, calculating homology, and finally evaluating the model. In the following subsections we explained each step a bit more granularly. We performed our experiment on a corpus of 1941 distinct text documents. We repeated our experiment twice: first including all words from the texts, and repeated with omitting all stop words.

2.1 Dataset

There have been written a lot of literary works so far, and with the rise of the Internet, a lot of them have been made publicly available for anyone to access for free. One such online source is Project Gutenberg [1], where each book has been classified in at least one category, or in their terms, a bookshelf.

For our purposes, we have used a preprocessed dataset by Lahiri [2], which consists of more than 3000 text documents, each representing a single literary work. Each document is identified by the name of its author and its title. The downside of this dataset is that, unfortunately, it does not contain any information about the category to which the books belongs to.

Finally, we categorized 37 authors into 8 distinct categories: *mysticism*, *historical*, *children*, *adventure*, *psychological*, *social*, *detective* and *plays*. The categorization was performed by determining the category of the majority of an author's books, for example majority of Sir Arthur Conan Doyle's books are detective because he was mostly writing about the adventures of Sherlock Holmes. The detailed categorization of each author and the amount of his/her documents used is presented in Appendix A.

2.2 Data processing

The pipeline for processing the data is rather simple: we split each document into its sentences and calculated the sentence embeddings using LASER [3]. Each of the sentence embeddings contains 1024 components, so the result from LASER was a $1024 \times N$, where N is the number of sentences in the document. We squashed this matrix to a single column by averaging each component of the sentence embedding over all sentences, yielding a 1024×1 vector. We stack the resulting vectors into $1024 \times M$ matrix, where M is the number of documents for each category. Finally, we use SVD decomposition to extract the first $r = 5$ components and reduce the final output to $5 \times M$, which we use for building our model of the data.

For the second part of our experiment, we exclude all the stop words from the sentences. We perform the exclusion after dividing the document into sentences. To remove the stop words, we use `nltk`'s corpus [4] for English stop words.

2.3 Model

The reduced-dimension matrices are taken as input point clouds to produce an α-shape complex. We have experimented with 2 different approaches of selecting the increasing sequence of points for the filtration of our complex.

The first (the naive and unsuccessful) one was selecting R to be the largest distance between any two points in the point cloud, and taking the values $0 = r_0 < r_1 < \cdots < r_9 < r_{10} = R$ that partition the interval $[0, R]$ into 10 equal parts to be the filtration values. However, since the α-complex is a subcomplex of the Čech and the Delaunay, the value of R turned out to be too large, and even the first sublevel complex using the value that one tenth as large as R, i.e. r_1, included all simplices of the α_∞ complex. Hence, this resulted in a trivial filtration: $S = K_{r_0} \leq K_{r_1} = \ldots K_{r_{10}} = \alpha(S)$, where the point cloud is denoted with S. Therefore, this approach was discarded.

In the second approach, after constructing the Delaunay complex (α_∞ complex), R was chosen to be the largest distance between points such that there exists a 1-simplex including both of them as its vertices. The filtration values $0 = r_0 < r_1 < \cdots < r_9 < r_{10} = R$ were again chosen by partitioning the interval $[0, R]$ into 10 equal parts.

Having the filtration of the α complex, we proceed to build the persistence diagrams for the first three dimensions of the data, i.e. for dimension $d \in \{0, 1, 2\}$. This procedure is identically repeated for each of the 8 document type matrices, resulting in 8 separate filtrations and their corresponding persistence diagrams.

For the purposes of implementation we utilized the Gudhi library [5]. However, since it works on the set of reals, we discretized the birth and death values using our r_i values. More specifically, each finite real value v provided by Gudhi, was transformed as the smallest $r_i, i \in \{0, \ldots, 10\}$ such that $r_i \geq r$. We should note that the case of such r_i not to exist is impossible due to the way we chose $R = r_{10}$.

Based on the persistence diagrams, we were able to calculate the pairwise bottleneck distances of each category, for all dimensions separately. We built a distance matrix D with the obtained distances and we applied an agglomerative clustering algorithm in order to construct dendrograms.

We chose to build dendrograms because they illustrate the hierarchical relationship of the underlying data. In our case we show the similarity of distinct literary categories, i.e. the most similar ones are connected first and then the most similar to the newly born category is merged, etc. until we have a single group consisting of all categories.

The procedure was repeated two times, first taking all the words into consideration and including them in the embeddings, and a second repetition for the case when the stop-words are excluded.

3 Results

As a result of the experiment we obtained two distinct sets of results, based on inclusion/exclusion of the stop words. In both cases the model generated persistence diagrams for each category separately, which are visually presented in Appendix B. They include the dimensions $0, 1$ and 2 on same image, and the point multiplicity is handled by point-transparency, by including inversely proportional relationship between the multiplicity of a point and its transparency. The dendrograms we obtained from the bottleneck distances of the persistence diagrams are presented in Appendix C.

The obtained dendrograms were closely examined by an experienced professional [1] and it was concluded that the dendrogram shown in Figure 1, which is obtained by omitting the stop words and for $d = 0$ was an almost perfect classification from a semantic point of view of the categories. Even though the distances among the persistence diagrams for all categories are very small, the distinction between the categories and how the clusters are merged hierarchically together is correct.

The rest of the dendrograms are performing rather worse. There are cases where the dendrograms do not make any sense, such a cases are presented in Figure 4e and Figure 4f. But then again, there are also cases, such as in Figure 4c, where we observe that there is almost no distinction between the categories *psychological* and *detectives*, whilst *historical*, *children*, *social* and *adventure* are perfectly distinguished, just like in Figure 1. This led us to the conclusion that there is a negative correlation between the dimension d and the discriminatory performance of the persistence diagrams, i.e. with the increasing of the dimension, the performance is decreased. And thus, we concluded that we get the best performance in each case at $d = 0$.

To achieve peak performance of our model we had to minimize the amount of noise we include in the input data, i.e. remove the stop words from the corpus which we used for analysis. Another indicator are the distances observed in both the persistence diagrams and the dendrograms, which are in the rank of 10^{-3}, and when working with such small distances the probability of making an error is increased, since there is not much room left for error. Which led us to the conclusion that our model is very susceptible to noise.

4 Conclusion

In this project we presented an approach to capture the underlying semantic difference between literature categories, using persistence homology on sentence-level document embedding. We em-

[1] Marko's mother is a high-school literature professor :)

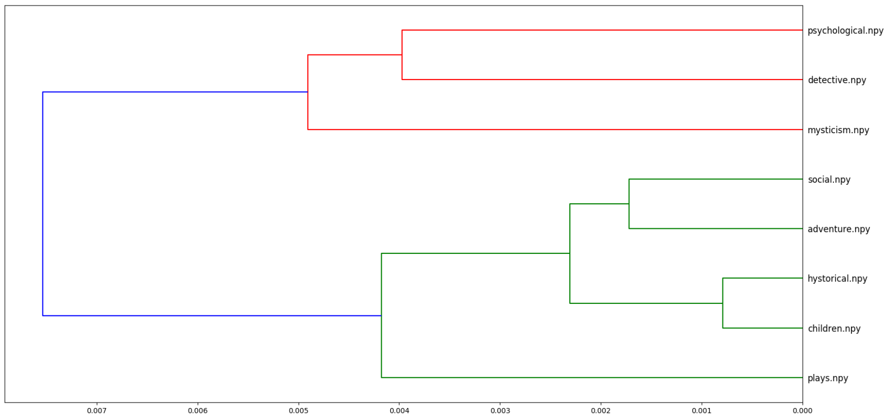


Figure 1: We obtained our experiment's best result by omitting the stop words at $d = 0$. The colours represent the discovered clusters, which merge together the most similar categories.

ploy this model on eight categories (mysticism, historical, children, adventure, psychological, social, detective and plays). The evaluation showed that the persistence homology in dimension 0 outperforms dimensions 1 and 2, and at this dimension, excluding the stop words from the documents yields perfect classification.

Appendix

A Dataset categorization

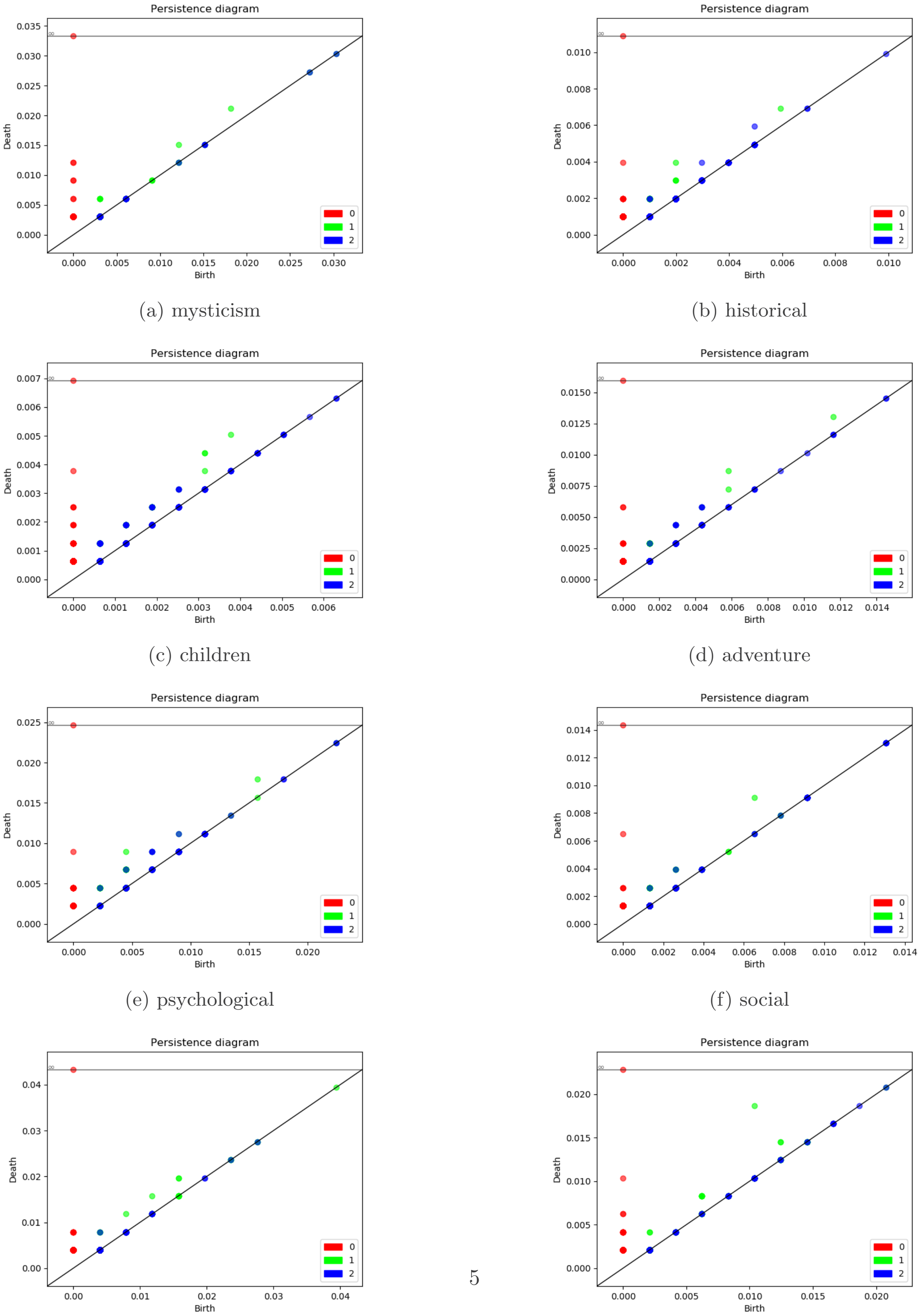
The dataset categorization is presented in Table 1.

B Persistence diagrams

The persistence diagrams of each categories are presented in Figure 2 and Figure 3 for all words and with the omission of stop words, respectively.

C Dendrograms

The resulting dendrograms from the distances among each persistence diagrams pairwise are presented in Figure 4.



(a) mysticism

(b) historical

(c) children

(d) adventure

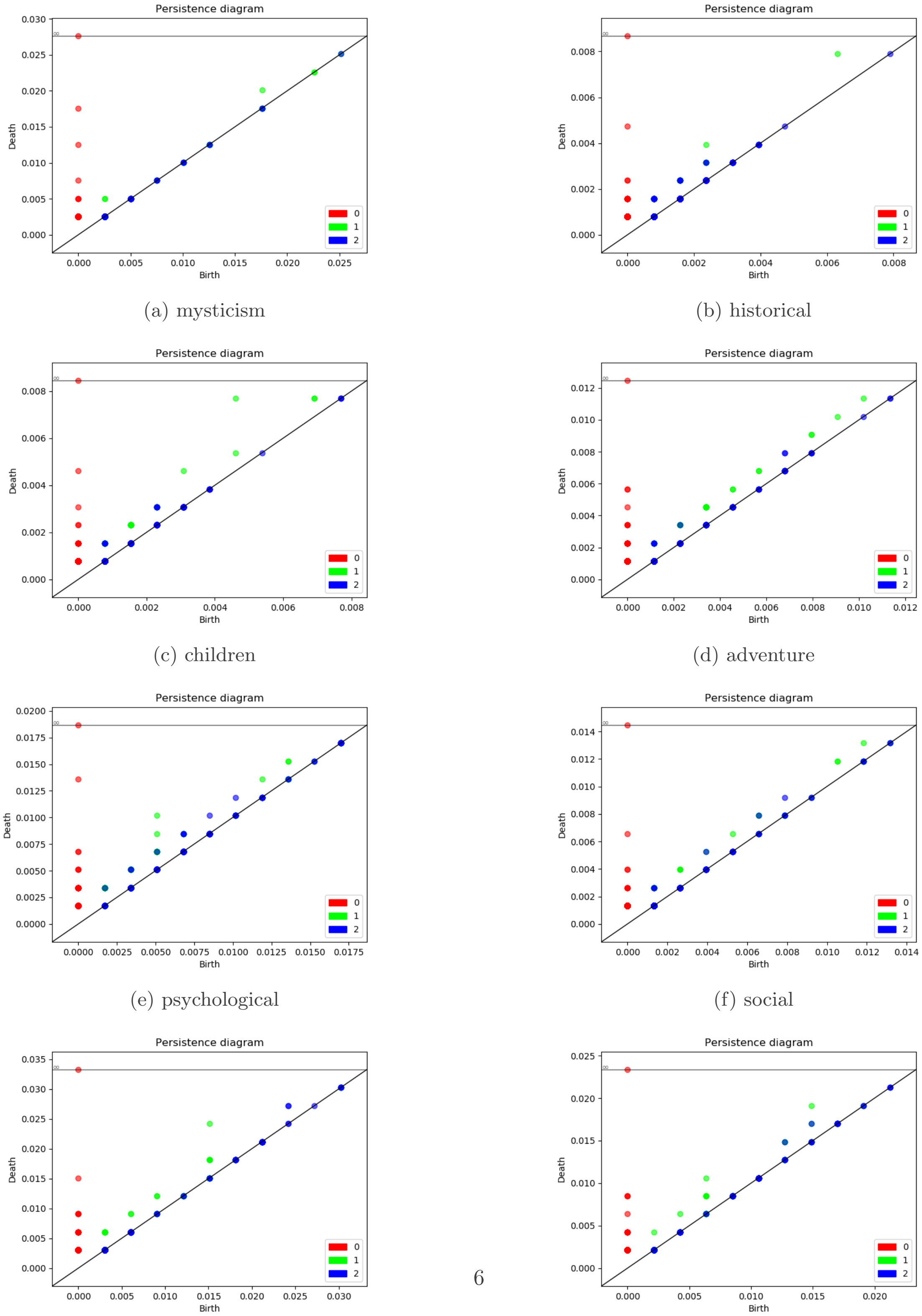
(e) psychological

(f) social

(g) detective

(h) plays

Figure 2: Persistence diagrams of documents per category including all words.



(a) mysticism

(b) historical

(c) children

(d) adventure

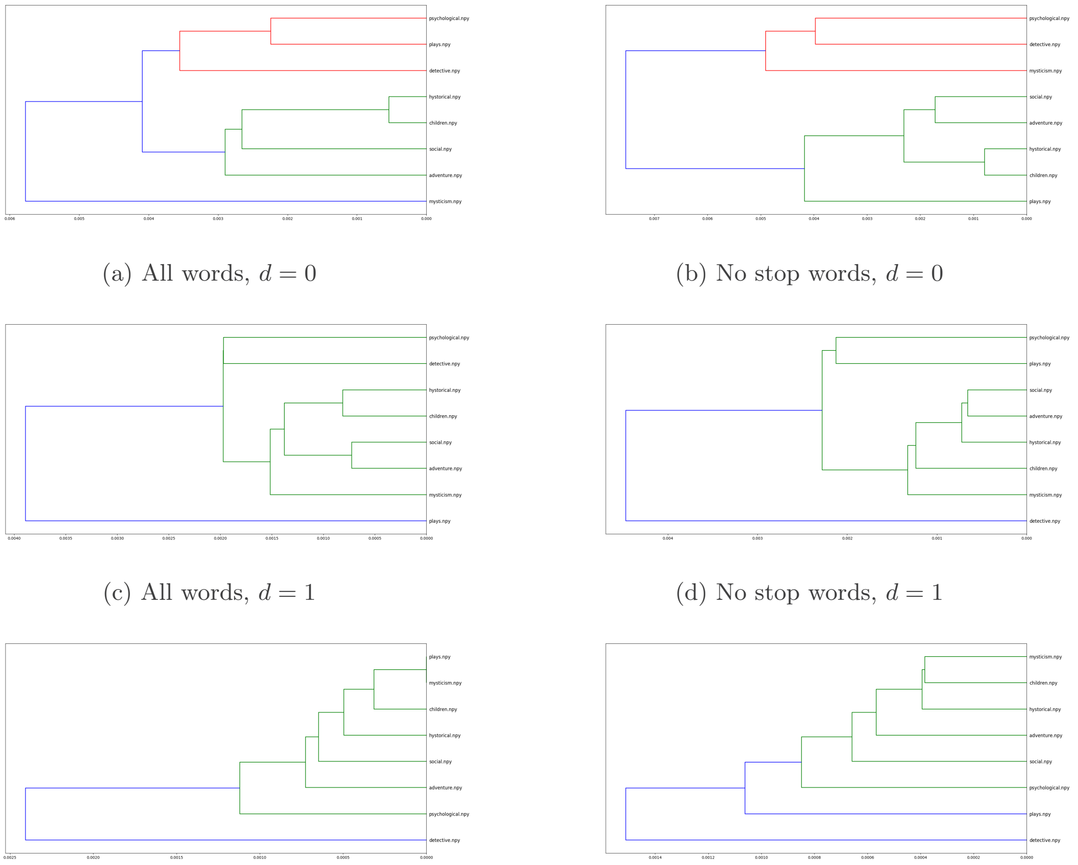
(e) psychological

(f) social

(g) detective

(h) plays

Figure 3: Persistence diagrams of documents per category omitting all stop words.



(a) All words, $d = 0$

(b) No stop words, $d = 0$

(c) All words, $d = 1$

(d) No stop words, $d = 1$

(e) All words, $d = 2$

(f) No stop words, $d = 2$

Figure 4: Agglomerative clustering based on the bottleneck distance between all categories. On the left we have included all words, whilst on the right we omitted the stop words. Each row presents a single dimension.

References

[1] Project gutenberg. (n.d.).
URL https://www.gutenberg.org/

[2] S. Lahiri, Complexity of Word Collocation Networks: A Preliminary Structural Analysis, in: Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 96–105 (April 2014).
URL http://www.aclweb.org/anthology/E14-3011

[3] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond (2018). arXiv:1812.10464.

[4] E. Loper, S. Bird, Nltk: The natural language toolkit, in: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002 (2002).

[5] The GUDHI Project, GUDHI User and Reference Manual, GUDHI Editorial Board, 2015 (2015).
URL http://gudhi.gforge.inria.fr/doc/latest/

Category	Author	Book count
mysticism	William Wymark Jacobs	97
	G K Chesterton	39
		$\sum = 136$
historical	George Alfred Henty	89
	William Dean Howells	84
	Henry James	72
	James Fenimore Cooper	36
	Sir Walter Scott	35
	William Makepeace Thackeray	30
		$\sum = 346$
children	R M Ballantyne	88
	Andrew Lang	60
	Charlotte Mary Yonge	60
	Jacob Abbott	47
	Mark Twain	47
	Charles Kingsley	44
	Rudyard Kipling	43
	Frank Richard Stockton	32
	Thornton Waldo Burgess	31
		$\sum = 452$
adventure	Robert Louis Stevenson	79
	Bret Harte	58
	Edward Stratemeyer	58
	Henry Rider Haggard	52
	Jack London	48
	Harold Bindloss	43
	Daniel Defoe	40
		$\sum = 378$
psychological	Nathaniel Hawthorne	86
	Joseph Conrad	34
		$\sum = 120$
social	Anthony Trollope	71
	Charles Dickens	61
	John Galsworthy	40
	P G Wodehouse	35
	Louisa May Alcott	34
		$\sum = 241$
detective	Sir Arthur Conan Doyle	57
	Edward Phillips Oppenheim	53
	Wilkie Collins	32
		$\sum = 142$
plays	George Bernard Shaw	42
	John Ruskin	42
	Lyman Frank Baum	42
		$\sum = 126$

Table 1: Categorization of each author included in our experiment. The table lists how many of his/her works we included in our calculations and the sum of all documents per category. In total, we used 1941 distinct text documents.