BlurbGenreCollection_EN

Language Technology Group, Universität Hamburg

December 17, 2018

1 Introduction

The **BlurbGenreCollection-EN**[1] is a dataset consisting of advertising descriptions of books – so called blurbs – for the English language. Each blurb is categorized into one or multiple categories. The categories are structured in a hierarchy. This dataset follows the policies as described in the RCV1 dataset by Lewis et al. (2004). We adapt RCV1's properties, which have been explained by its authors in detail and refer to their description. The **minimum code policy** requires the assignment of at least one category to each document of the collection. The **hierarchy policy** ensures that every ancestor of a document's label is assigned as well.

The copyright to all blurbs belongs to **Penguin Random House (PRH)**, its licensors, vendors and/or its content providers since the blurbs were obtained through the Penguin Random House website[2]. The blurbs serve promotional/public purposes and permission has been granted by Penguin Random House to share this dataset. The dataset is shared under the CC BY-NC 4.0 license[3], which allows copying and redistributing the dataset as long as appropriate credit is given, especially to Penguin Random House and its content providers. In addition to indicating changes to the dataset, you may not use the dataset for commercial purposes.

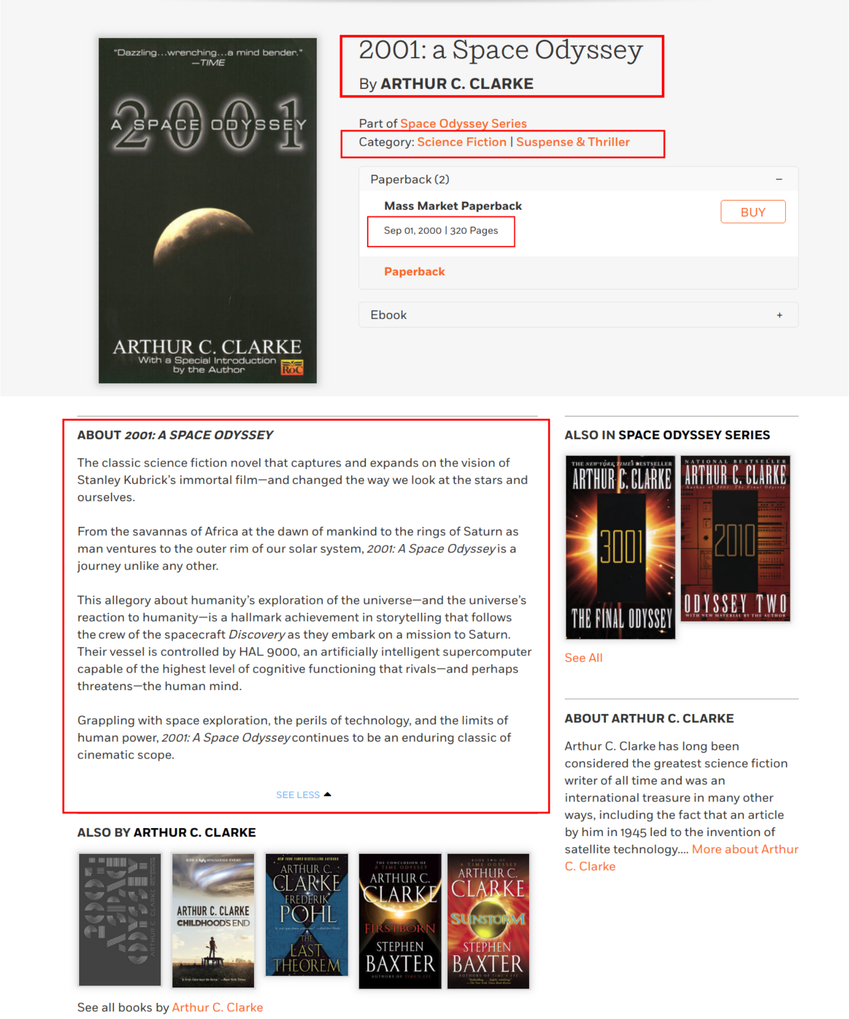
2 Contents

The Penguin Random House webpage lists every book with its respective blurb. We extracted the 'about sections' as well as the genres (category) of a book. We further extracted title, author, URL, ISBN, number of pages and the date of publication. The date of publication normally represents the publication date of the particular version of the book. The data we extracted is illustrated below.

[1] `https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/blurb-genre-collection.html`

[2] `https://www.penguinrandomhouse.com`

[3] `(https://creativecommons.org/licenses/by-nc/4.0/)`



2001: a Space Odyssey

By **ARTHUR C. CLARKE**

Part of Space Odyssey Series

Category: Science Fiction | Suspense & Thriller

Paperback (2) –

Mass Market Paperback

Sep 01, 2000 | 320 Pages

Paperback

Ebook +

ABOUT *2001: A SPACE ODYSSEY*

The classic science fiction novel that captures and expands on the vision of Stanley Kubrick's immortal film—and changed the way we look at the stars and ourselves.

From the savannas of Africa at the dawn of mankind to the rings of Saturn as man ventures to the outer rim of our solar system, *2001: A Space Odyssey* is a journey unlike any other.

This allegory about humanity's exploration of the universe—and the universe's reaction to humanity—is a hallmark achievement in storytelling that follows the crew of the spacecraft *Discovery* as they embark on a mission to Saturn. Their vessel is controlled by HAL 9000, an artificially intelligent supercomputer capable of the highest level of cognitive functioning that rivals—and perhaps threatens—the human mind.

Grappling with space exploration, the perils of technology, and the limits of human power, *2001: A Space Odyssey* continues to be an enduring classic of cinematic scope.

SEE LESS ▲

ALSO BY ARTHUR C. CLARKE

See all books by Arthur C. Clarke

ALSO IN SPACE ODYSSEY SERIES

See All

ABOUT ARTHUR C. CLARKE

Arthur C. Clarke has long been considered the greatest science fiction writer of all time and was an international treasure in many other ways, including the fact that an article by him in 1945 led to the invention of satellite technology.... More about Arthur C. Clarke

Figure 1: Website snippet the data was collected from. The specific parts are highlighted in red boxes. Snippet taken in October 2018.

Since the webpages only lists the most specific genre, we manually created the hierarchy for writing genres and add all parent labels during post-processing. Further pruning was done to remove genres that capture properties which do not rely on content but on the shape or form of a book. From the Childrens' Book category, three out of ten sub-genres were kept as most of them specify the presentation and shape of a book and not the content. For example 'Picture Book', 'Boxed Sets', 'Board Books' and so on. The genre Audiobooks and all its descendants were removed as well for obvious reasons. As the last preprocessing step, we remove every book that has assigned genres (combinations) that appear less than five times in the complete dataset.

An exemplary entry of the resulting dataset is shown below:

```
<book date="2018-08-18" xml:lang="en">
  <title>2001: a Space Odyssey</title>
  <body>
  The classic science fiction novel that captures and expands on the
      vision of Stanley Kubrick's immortal film-and changed the way we look
      at the stars and ourselves. From the savannas of Africa at the dawn
      of mankind to the rings of Saturn as man ventures to the outer rim of
      our solar system, 2001: A Space Odyssey is a journey unlike any
      other. This allegory about humanity's exploration of the universe-and
      the universe's reaction to humanity-is a hallmark achievement in
      storytelling that follows the crew of the spacecraft Discovery as
      they embark on a mission to Saturn. Their vessel is controlled by HAL
      9000, an artificially intelligent supercomputer capable of the
      highest level of cognitive functioning that rivals-and perhaps
      threatens-the human mind. Grappling with space exploration, the
      perils of technology, and the limits of human power, 2001: A Space
      Odyssey continues to be an enduring classic of cinematic scope.
  </body>
  <copyright>(c) Penguin Random House</copyright>
  <metadata>
  <topics>
    <d0>Fiction</d0>
    <d1>Science Fiction</d1>
    <d1>Mystery & Suspense</d1>
    <d2>Suspense & Thriller</d2>
  </topics>
  <author>Arthur C. Clarke</author>
  <published>Sep 01, 2000 </published>
  <page_num> 320 Pages</page_num>
  <isbn>9780451457998</isbn>
  <url>https://www.penguinrandomhouse.com/books/325356/2001-
  a-space-odyssey-by-arthur-c-clarke/</url>
  </metadata>
</book>
```

Listing 1: Example entry of a book in the dataset BlurbGenreCollection-EN

Additionally, a file that contains only the hiararchy in form of parent-child relationships is provided.

3 Quantitative characteristics

The datset is split into three subsets for training, validation and testing by appling stratified sampling to ensure that splits do not disfigure the distribution of labels. The total data is split in the ratio of 64%, 16% and 20% for train, dev and test respectively.

Dataset	BlurbGenreCollection-EN
Number of samples	91,892
Average length of blurb	157.51
Total number of classes $\lvert\mathcal{L}\rvert$	146
Classes on level 1;2;3;4	7; 46; 77; 16
Average number of genres per book	3.01

Table 1: Quantitative characteristics of both introduced datasets. The average length of blurbs is measured in number of words. l = 1;2;3;4 lists the number of genres on each level l of the hierarchy.

3.1 Genres

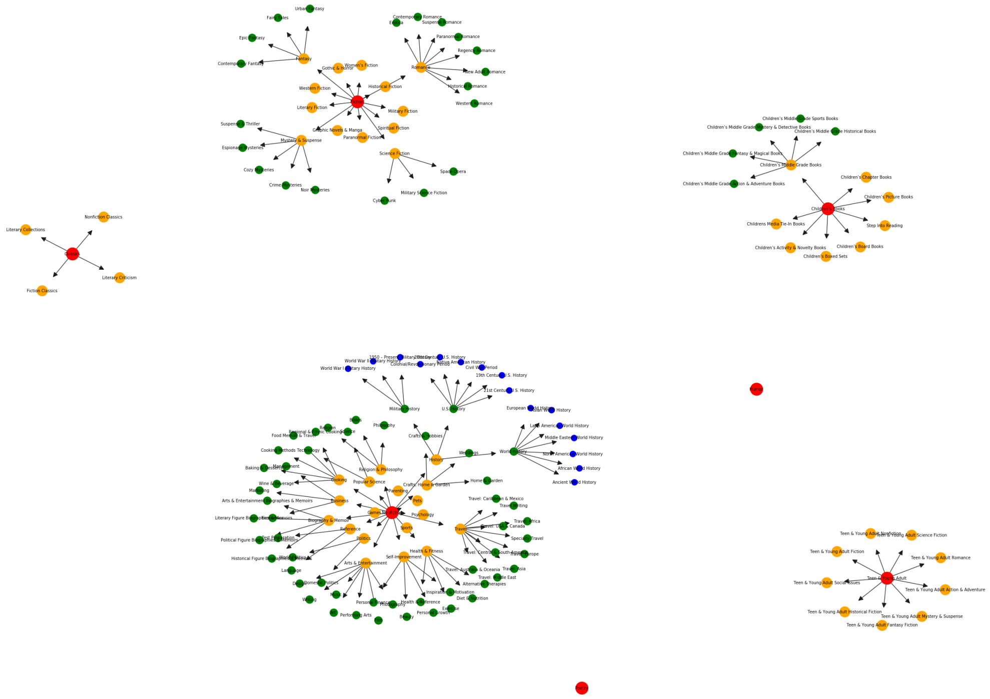


Figure 2: Illustration of the underlying genre hierarchy

The hierarchy of the dataset (Figure 2) consists of four levels and is organized as a forest. It is important to note that the most specific genre of a book does not have to be a leaf. For instance, the most specific class of a book could be Children's Books, although Children's Book has further children genres, such as Middle Grade books. However, a great number of books are simply not classified into more special classes on the website. Analyzing the occurrence of each genre combination on a log scale shows that the English dataset has a distribution in which some labels have disproportionately few or many examples as shown in the Figure 3.

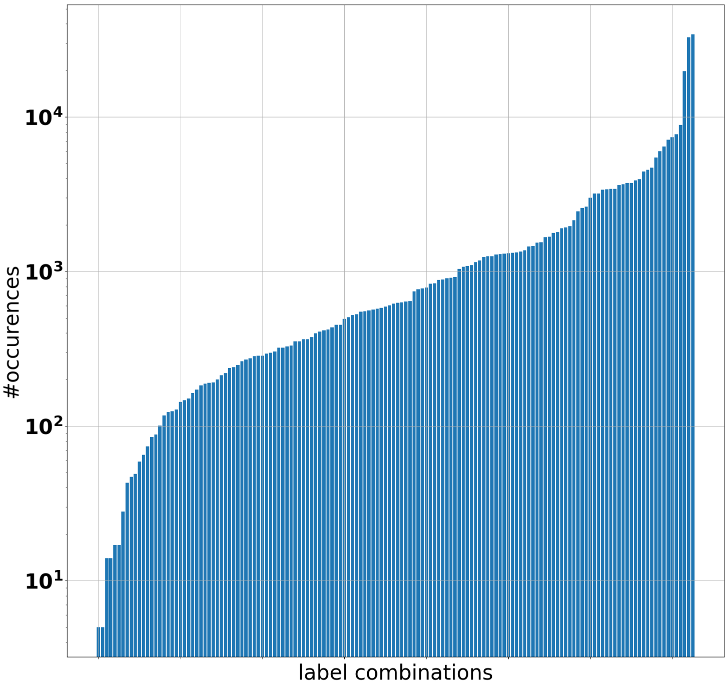


Figure 3: Occurence of label combinations in ascending order in log scale.

4 Classification

The dataset is used for a hierarchical multi-label classification task, where each label is part of a hierarchy. The baseline model that has been employed to classify books into their respective writing genre is the Support-Vector Machine (Cortes and Vapnik, 1995). The following results are created exclusively on basis of a book's blurb.

Classifier	SVM
Precision	85.37
Recall	61.11
F1	71.23
Subset accuracy	35.79

Table 2: Classifier results for micro recall, precision, F1 and subset accuracy

References

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine learning 20*(3), 273–297.

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research 5*(Apr), 361–397.