Asymmetric semantic search using document contextual embeddings on long documents

Nikita Fordui
University of Tartu
nikita.forduy@protonmail.com

Abstract

Asymmetric semantic search is a task of matching short prompts and long texts based on semantic meaning. This project describes approach to the task based on storing and search over sentence/text embeddings. As encoder for generating embeddings Sentence-BERT models family is used. For the task accuracy-base evaluation method is introduce, Different models with different retrieval mechanisms conditionings are compared.

1 Introduction

The task of asymmetric semantic search in NLP is a task of searching for corresponding long text based on a short prompt.

My idea for this project is to develop an effective asymmetric search pipeline that would work specifically for very long texts - books. For this I created dataset of 78 open access books in English language.

Pipeline for creating this search engine was pretty clear for me: create a dataset of books, turn them into embeddings, use the same approach to turn prompt into embedding and look for a nearest neighbor of prompt embedding in books embedding space.

Problems started when trying to find a way to have such a latent space so that similar short and long texts would be near each other. My first approach was to train doc2vec (Le and Mikolov, 2014) model on created dataset to generate embeddings and then use the same model for prompt embedding generation.

After initial research into this topic and try to do everything mentioned book embeddings were not representative of books meaning in this particular task. Even though prompts were similar to what the books are about embeddings of the prompts were far of from actual books with which prompt was constructed in mind. This meant that doc2vec (Le and Mikolov, 2014) model wasn't suitable for this approach.

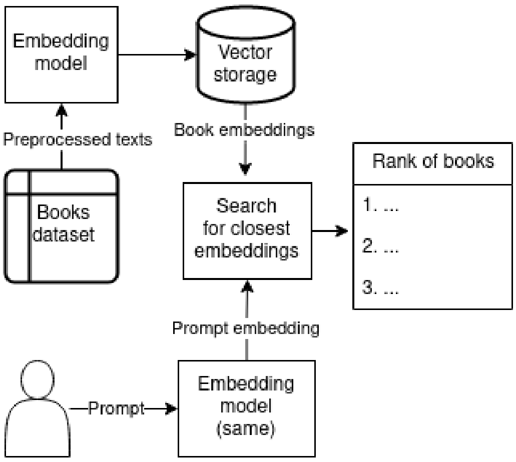


Figure 1: Assymetric semantic search system schema.

After testing out other different approaches based on aggregating of word embeddings I came to using Sentence-BERT (Reimers and Gurevych, 2019) as a base model.

2 Methods

2.1 Dataset

Dataset for creating this projects were 78 books (see Appendix B) collected by author from open-access library Project Gutenberg. For the preprocessing info about book name and author was extracted, books were cleaned of specific to data source artifacts, cleaned of all special characters and other unnecessary noise.

For further evaluation of our approach on the dataset each book metadata includes short description which would be used as ground truth prompt for evaluation (see 2.4).

2.2 Models for embeddings

For creating embeddings for the books as well as prompt encoding 6 different pretrained Sentence-BERT models were used available on Hugginface.

To run inference for the model sentence-transformers Python library was used.

Specific model choice was based on one main factor. These models was trained on MS Marco dataset (Bajaj et al., 2018) - dataset of pairs of short sentences and long paragraphs. Both prompts and long texts would be projected in same latent space. This way embeddings generated by the model are perfect for asymmetric semantic search.

2.3 Work with embeddings

For storing the embeddings `NumPy` library and it's saving capabilities were used. For the search process over the embedings `faiss` (Johnson et al., 2019) library was used.

2.4 Evaluation

Each book from the dataset contains its name, author and one-sentence description as metadata. These descriptions were used as ground truth prompts for quantitative system evaluation. Three metrics were chosen for evaluation: top-1, top-5 and top-10 accuracy.

Evaluation process looks like this:

1. Retrieve book description from metadata.
2. Use the book description as a prompt and run it through the system.
3. Get 10 closest books from the system.
4. If book which description we took is the first in the list - we count it as guessed to top-1 accuracy, if in top 5 of the list - to top-5 accuracy, if in list at all - to top 10 accuracy.
5. Repeat all the previous steps for all book descriptions.
6. Calculate final metrics.

2.5 Demo

Demo for this paper is available by the link `https://huggingface.co/spaces/nikiandr/assym_sem_search`. It was written using Gradio, Python library for creating ML applications.

2.6 Code

All codebase for the project is available by the link `https://github.com/nikiandr/nlp_project`.

3 Results

Let's start with results produced with evaluation schema described in 2.4. Six models with different retrieval mechanisms which provide best results (Reimers, 2022) on MS Marco dataset (Bajaj et al., 2018) retrieval task were used.

	Accuracies		
Models	Top-1	Top-5	Top-10
Cosine similarity models			
`msmarco-distilbert-cos-v5`	**0.64**	**0.86**	**0.92**
`msmarco-MiniLM-L6-cos-v5`	0.47	0.74	0.87
`msmarco-MiniLM-L12-cos-v5`	0.49	0.71	0.79
Dot product models			
`msmarco-distilbert-base-tas-b`	0.74	0.92	**0.97**
`msmarco-distilbert-dot-v5`	0.73	**0.95**	0.96
`msmarco-bert-base-dot-v5`	0.74	0.91	0.96

Table 1: Top-1, top-5, and top-10 accuracies for ground truth prompts on models used.

As we can see in Table 1 cosine similarity conditioned models generally perform worse on our task then dot product conditioned models. We can even compare models with similar setups (`msmarco-distilbert-cos-v5` and `msmarco-distilbert-dot-v5`): having one underlying model but conditioned on different retrieval mechanisms, these two models' differences in performance demonstrate that dot product based models are better in this particular case.

This gives us empirical evidence for usage of dot product conditioned models for this specific task in this setup as well as which particular models may be used.

More qualitative results on different prompts can be found in Appendix A.

4 Discussion

My main idea for this project was to come up with functional pipeline for the task described as well as figure out which approaches work better for this very specific task. I feel like this was accomplished in the project but there is a lot more that can be done.

Idea for this project came from trying to solve the same task for Ukrainian language but in process of figuring out the direction there were a lot of problems which would take much more time to figure out then given for this project. It feels that it would be nice future direction to transfer this approach to other low resource languages.

Another discussion which can be generated from this project is how to come up with a way of more effectively deal with documents of different sizes e.g. short stories vs large novels.

5 Conclusions

To conclude, in this project author came up and implemented pipeline for asymmetric semantic segmentation, studied impact of difference between different models for the pipeline and came up with eddective evaluation schema for the task.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. ArXiv:1405.4053 [cs].

Nils Reimers. 2022. Msmarco models.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

A Result examples

Here are couple examples of prompts results compared between cosine similarity and dot product conditioned models.

Prompt: Book about captain swimming through the sea on a submarine.	
`msmarco-distilbert-cos-v5`	`msmarco-distilbert-base-tas-b`
1. Gulliver's Travels	1. Twenty Thousand Leagues Under the Seas
by Jonathan Swift: 0.37	by Jules Verne: 95.12
2. The Life and Adventures of Robinson Crusoe	2. Treasure Island
by Daniel Defoe: 0.35	by Robert Louis Stevenson: 92.34
3. Three Men in a Boat	3. Adventures of Huckleberry Finn
by Jerome K. Jerome: 0.35	by Mark Twain: 91.49
4. Treasure Island	4. Three Men in a Boat
by Robert Louis Stevenson: 0.32	by Jerome K. Jerome: 91.33
5. The Hound of the Baskervilles	5. The Life and Adventures of Robinson Crusoe
by Arthur Conan Doyle: 0.32	by Daniel Defoe: 91.26

Table 2: Results for prompt *Book about captain swimming through the sea on a submarine.*

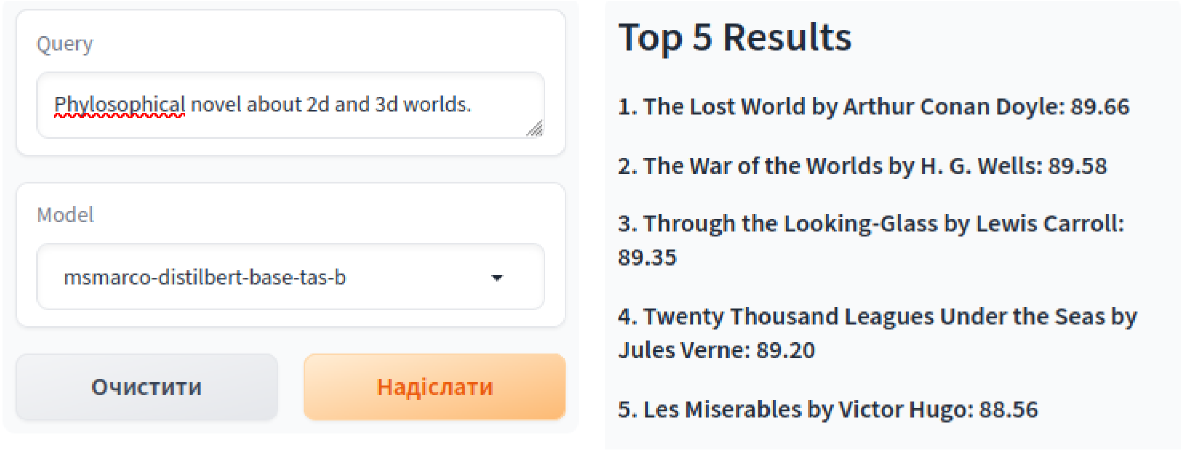
Prompt: Book about love and pain.	
`msmarco-distilbert-cos-v5`	`msmarco-distilbert-base-tas-b`
1. Three Men in a Boat	1. Middlemarch
by Jerome K. Jerome: 0.32	by George Eliot: 90.76
2. Winnie-the-Pooh	2. A Tale of Two Cities
by A. A. Milne: 0.30	by Charles Dickens: 90.25
3. Middlemarch	3. The Picture of Dorian Gray
by George Eliot: 0.30	by Oscar Wilde: 89.88
4. Pride and Prejudice	4. Pride and Prejudice
by Jane Austin: 0.29	by Jane Austin: 89.77
5. Ivanhoe: A Romance	5. Three Men in a Boat
by Walter Scott: 0.28	by Jerome K. Jerome: 88.94

Table 3: Results for prompt *Book about love and pain.*

Prompt: Book about love and pain.	
`msmarco-distilbert-cos-v5`	`msmarco-distilbert-base-tas-b`
1. Flatland: A Romance of Many Dimensions	1. A Tale of Two Cities
by Edwin Abbott Abbott: 0.33	by Charles Dickens: 90.03
2. Twenty Thousand Leagues Under the Seas	2. The Picture of Dorian Gray
by Jules Verne: 0.27	by Oscar Wilde: 88.89
3. The Time Machine	3. The Lost World
by H. G. Wells: 0.27	by Arthur Conan Doyle: 88.24
4. The Call of Cthulhu	4. The Time Machine
by H. P. Lovecraft: 0.26	by H. G. Wells: 87.61
5. The War of the Worlds	5. The War of the Worlds
by H. G. Wells: 0.25	by H. G. Wells: 87.50

Table 4: Results for prompt *Book about captain swimming through the sea on a submarine.*

Here are different examples retrieved from demo which could be found here: https://huggingface.co/spaces/nikiandr/assym_sem_search.



Query

Phylosophical novel about 2d and 3d worlds.

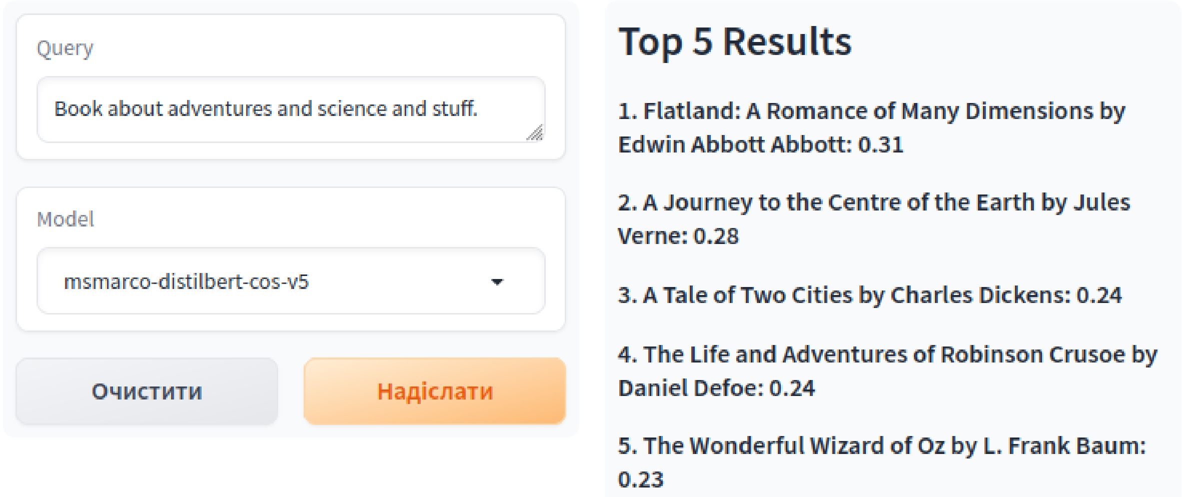
Model

msmarco-distilbert-base-tas-b

Top 5 Results

1. The Lost World by Arthur Conan Doyle: 89.66
2. The War of the Worlds by H. G. Wells: 89.58
3. Through the Looking-Glass by Lewis Carroll: 89.35
4. Twenty Thousand Leagues Under the Seas by Jules Verne: 89.20
5. Les Miserables by Victor Hugo: 88.56

Figure 2: Demo run for prompt *Philosophical novel about 2d and 3d worlds.*



Query

Book about adventures and science and stuff.

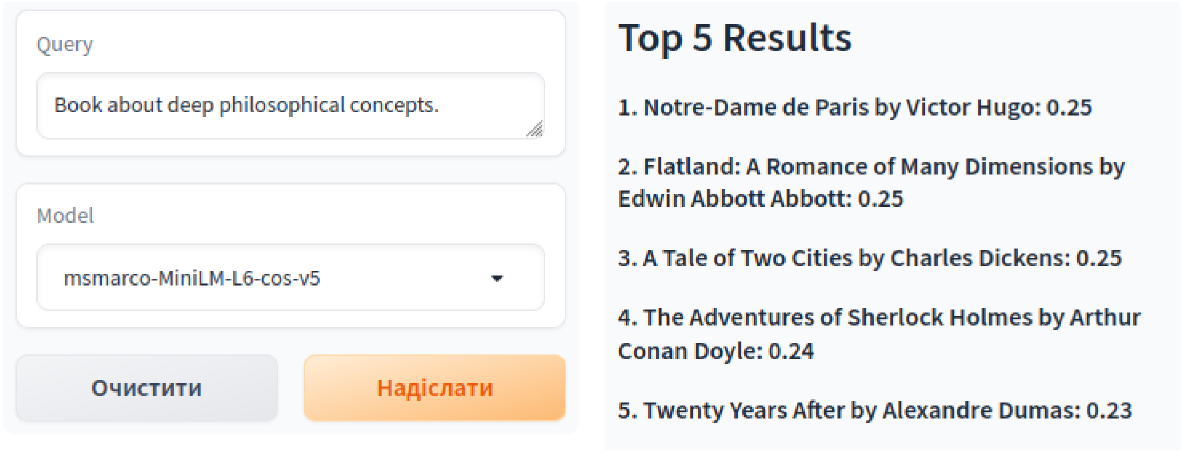
Model

msmarco-distilbert-cos-v5

Top 5 Results

1. Flatland: A Romance of Many Dimensions by Edwin Abbott Abbott: 0.31
2. A Journey to the Centre of the Earth by Jules Verne: 0.28
3. A Tale of Two Cities by Charles Dickens: 0.24
4. The Life and Adventures of Robinson Crusoe by Daniel Defoe: 0.24
5. The Wonderful Wizard of Oz by L. Frank Baum: 0.23

Figure 3: Demo run for prompt *Book about adventures and science and stuff.*



Query

Book about deep philosophical concepts.

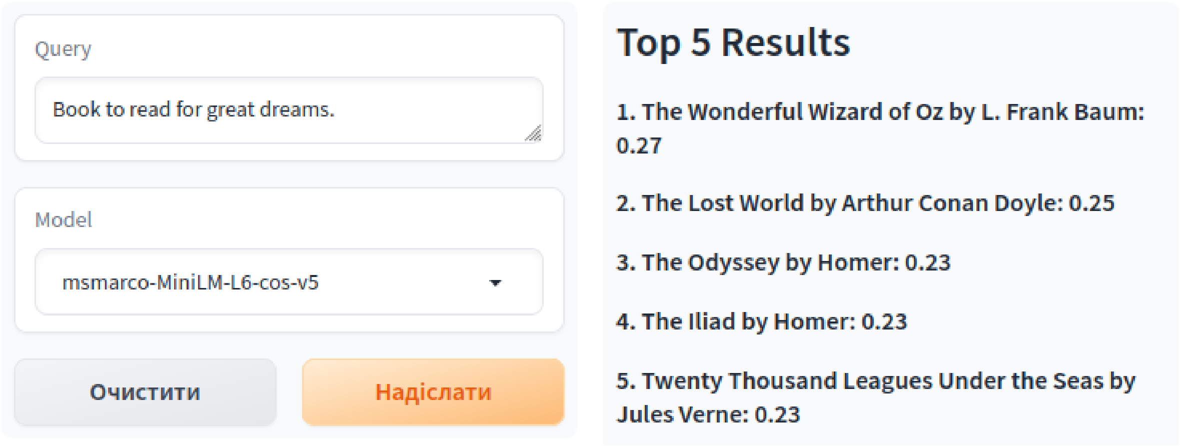
Model

msmarco-MiniLM-L6-cos-v5

Top 5 Results

1. Notre-Dame de Paris by Victor Hugo: 0.25
2. Flatland: A Romance of Many Dimensions by Edwin Abbott Abbott: 0.25
3. A Tale of Two Cities by Charles Dickens: 0.25
4. The Adventures of Sherlock Holmes by Arthur Conan Doyle: 0.24
5. Twenty Years After by Alexandre Dumas: 0.23

Figure 4: Demo run for prompt *Book about deep philsophical concepts.*



Query

Book to read for great dreams.

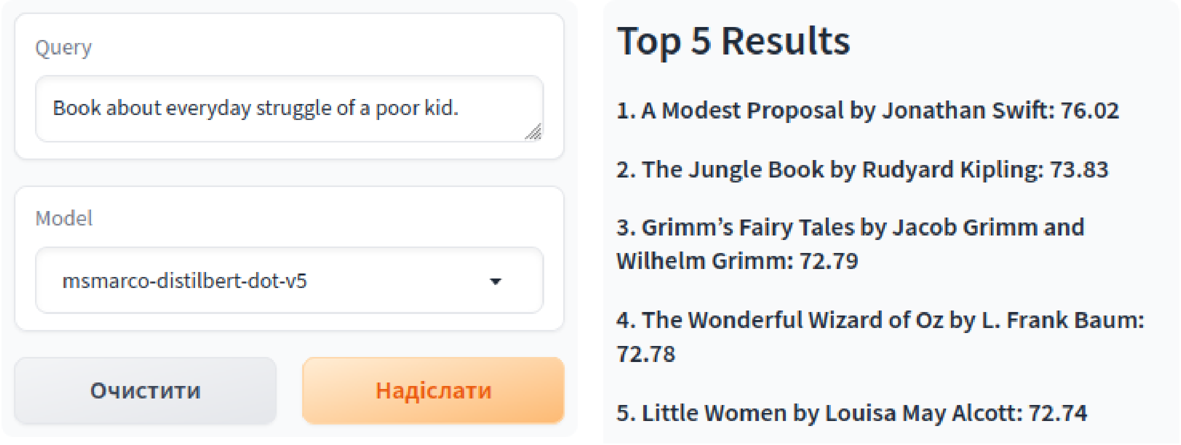
Model

msmarco-MiniLM-L6-cos-v5

Top 5 Results

1. The Wonderful Wizard of Oz by L. Frank Baum: 0.27
2. The Lost World by Arthur Conan Doyle: 0.25
3. The Odyssey by Homer: 0.23
4. The Iliad by Homer: 0.23
5. Twenty Thousand Leagues Under the Seas by Jules Verne: 0.23

Figure 5: Demo run for prompt *Book to read for great dreams.*



Query

Book about everyday struggle of a poor kid.

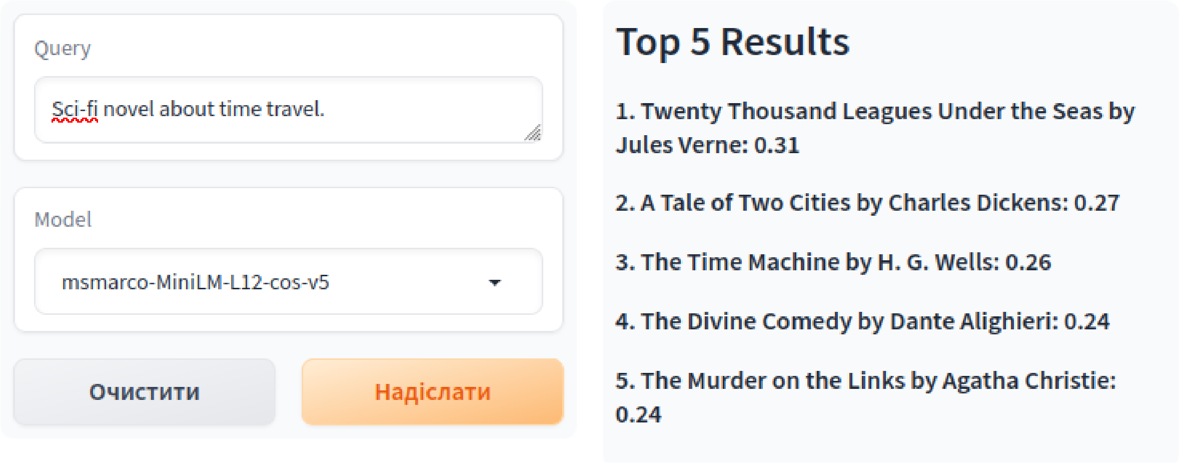
Model

msmarco-distilbert-dot-v5

Top 5 Results

1. A Modest Proposal by Jonathan Swift: 76.02
2. The Jungle Book by Rudyard Kipling: 73.83
3. Grimm's Fairy Tales by Jacob Grimm and Wilhelm Grimm: 72.79
4. The Wonderful Wizard of Oz by L. Frank Baum: 72.78
5. Little Women by Louisa May Alcott: 72.74

Figure 6: Demo run for prompt *Book about everyday struggle of a poor kid.*



Query

Sci-fi novel about time travel.

Model

msmarco-MiniLM-L12-cos-v5

Top 5 Results

1. Twenty Thousand Leagues Under the Seas by Jules Verne: 0.31
2. A Tale of Two Cities by Charles Dickens: 0.27
3. The Time Machine by H. G. Wells: 0.26
4. The Divine Comedy by Dante Alighieri: 0.24
5. The Murder on the Links by Agatha Christie: 0.24

Figure 7: Demo run for prompt *Sci-fi novel about time travel.*

B Dataset structure

Here is a list of all books collected and preprocessed into the final dataset.

Name	Author
The Iliad	Homer
The War of the Worlds	H. G. Wells
Cranford	Elizabeth Cleghorn Gaskell
The Great Gatsby	F. Scott Fitzgerald
Heidi	Johanna Spyri
The Prince	Niccolo Machiavelli
Ivanhoe: A Romance	Walter Scott
The Importance of Being Earnest	Oscar Wilde
Around the World in Eighty Days	Jules Verne
A Doll's House	Henrik Ibsen
Kim	Rudyard Kipling
Grimm's Fairy Tales	Jacob Grimm and Wilhelm Grimm
The Blue Castle	L. M. Montgomery
The Trial	Franz Kafka
The Picture of Dorian Gray	Oscar Wilde
Oliver Twist	Charles Dickens
Hamlet	William Shakespeare
The Tempest	William Shakespeare
Moby Dick; Or, The Whale	Herman Melville
The Strange Case of Dr. Jekyll and Mr. Hyde	Robert Louis Stevenson
The Hound of the Baskervilles	Arthur Conan Doyle
Dracula	Bram Stoker
Pollyanna	Eleanor H. Porter
Great Expectations	Charles Dickens
Tarzan and the Lost Empire	Edgar Rice Burroughs
Frankenstein	Mary Shelley
Winnie-the-Pooh	A. A. Milne
The Murder on the Links	Agatha Christie
Through the Looking-Glass	Lewis Carroll
Beyond Good and Evil	Friedrich Nietzsche
The Life and Adventures of Robinson Crusoe	Daniel Defoe
The Time Machine	H. G. Wells
A Journey to the Centre of the Earth	Jules Verne
Death in Venice	Thomas Mann
Pride and Prejudice	Jane Austin
The Call of Cthulhu	H. P. Lovecraft
Les Miserables	Victor Hugo
Ulysses	James Joyce
The Odyssey	Homer
Peter Pan	James Barrie
A Tale of Two Cities	Charles Dickens
Flatland: A Romance of Many Dimensions	Edwin Abbott Abbott
The Enchanted April	Elizabeth Von Arnim
The Divine Comedy	Dante Alighieri
The Adventures of Sherlock Holmes	Arthur Conan Doyle
Treasure Island	Robert Louis Stevenson
Little Women	Louisa May Alcott

Name	**Author**
The Adventure of Tom Sawyer	Mark Twain
Three Men in a Boat	Jerome K. Jerome
Romeo and Juliet	William Shakespeare
Alice’s Adventures in Wonderland	Lewis Carroll
Life on the Mississippi	Mark Twain
Don Quixote	Miguel de Cervantes Saavedra
Metamorphosis	Franz Kafka
Jane Eyre: An Autobiography	Charlotte Brontë
Utopia	Saint Thomas More
Pygmalion	Bernard Shaw
The Voyage of the Beagle	Charles Darwin
Twenty Thousand Leagues Under the Seas	Jules Verne
Gulliver’s Travels	Jonathan Swift
A Christmas Carol in Prose	Charles Dickens
Siddhartha	Herman Hesse
A Midsummer Night’s Dream	William Shakespeare
A Study in Scarlet	Arthur Conan Doyle
Notre-Dame de Paris	Victor Hugo
A Room with a View	E.M. Foster
Twenty Years After	Alexandre Dumas
The Jungle Book	Rudyard Kipling
The Three Musketeers	Alexandre Dumas
The Wonderful Wizard of Oz	L. Frank Baum
Martin Eden	Jack London
The Lost World	Arthur Conan Doyle
The Sea-Wolf	Jack London
Adventures of Huckleberry Finn	Mark Twain
The Gun	Philip K. Dick
David Copperfield	Charles Dickens
Middlemarch	George Eliot
A Modest Proposal	Jonathan Swift