RAG for school literature

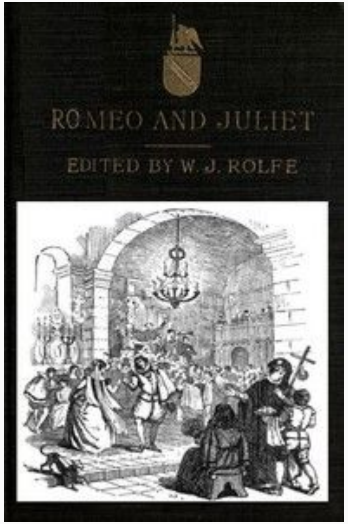
Maja Andrzejczuk, Piotr Bielecki, Paweł Gelar, Jakub Kasprzak, Maciej Orsłowski

Introduction to our project

Developing a RAG system for question answering.

Combining retrieval methods and generative models.

Focus on **famous classic books** often studied at schools (e.g., Romeo and Juliet, Hamlet, Around the World in 80 Days).



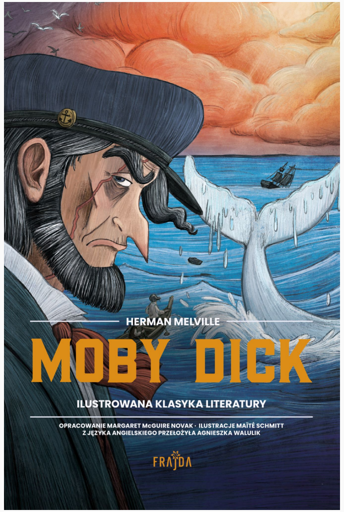
[Shakespeare's Tragedy of Romeo and Juliet by William Shakespeare | Project Gutenberg]

Datasets source: **Gutenberg Library**

Free access to a variety of classic books in raw text or PDF format

Example Books:

- *Ulysses* by James Joyce
- *Pride and Prejudice* by Jane Austen
- *Moby-Dick* by Herman Melville
- *The Adventures of Sherlock Holmes* by Arthur Conan Doyle
- *Dracula* by Bram Stoker


HERMAN MELVILLE
MOBY DICK
ILUSTROWANA KLASYKA LITERATURY
OPRACOWANIE MARGARET McGUIRE NOVAK · ILUSTRACJE MAÏTÉ SCHMITT
Z JĘZYKA ANGIELSKIEGO PRZEŁOŻYŁA AGNIESZKA WALULIK
FRAJDA

Moby Dick Herman Melville książka | TaniaKsiazka.pl

RAG (Retrieval-Augmented Generation)

Hybrid model combining document retrieval and text generation.

- **Document retrieval:** Retrieve relevant documents or passages from a large corpus.
- **Text generation:** Use a language model to generate answers based on the retrieved information.

Methodology and Framework Selection

1. Data Loading and Preprocessing:

   We use **Langchain & Unstructured** libraries to load PDF and TXT files and split their contents into chunks with overlap.

2. Vector Embeddings and Search:
   - Embedding model **all-MiniLM-L6-v2** converts passages into vector embeddings;
   - For a vector database, we use **FAISS**, which facilitates indexing and retrieval of document passages.

Methodology and Framework Selection

3. Cross-Encoder Reranker Model:

   bge-reranker-large: Re-ranks passages to enhance relevance.

4. LLM (Large Language Model):

   Qwen2.5-7B: Generates natural language answers from ranked passages.

Final pipeline

- After PoC, we parametrised and compacted all system components into **one reusable class**;
- We also loaded all books once into vectorstore and preserved its state to allow for faster initialisation;
- The class provides a static *quickstart()* method that initialises all system components with default parameters;
- Users can enhance the system with extra books using the *add_books()* method;
- Once initialised, the system can be queried by calling the *answer()* method;
- All methods are parametrised, and the whole class contains exhaustive documentation.

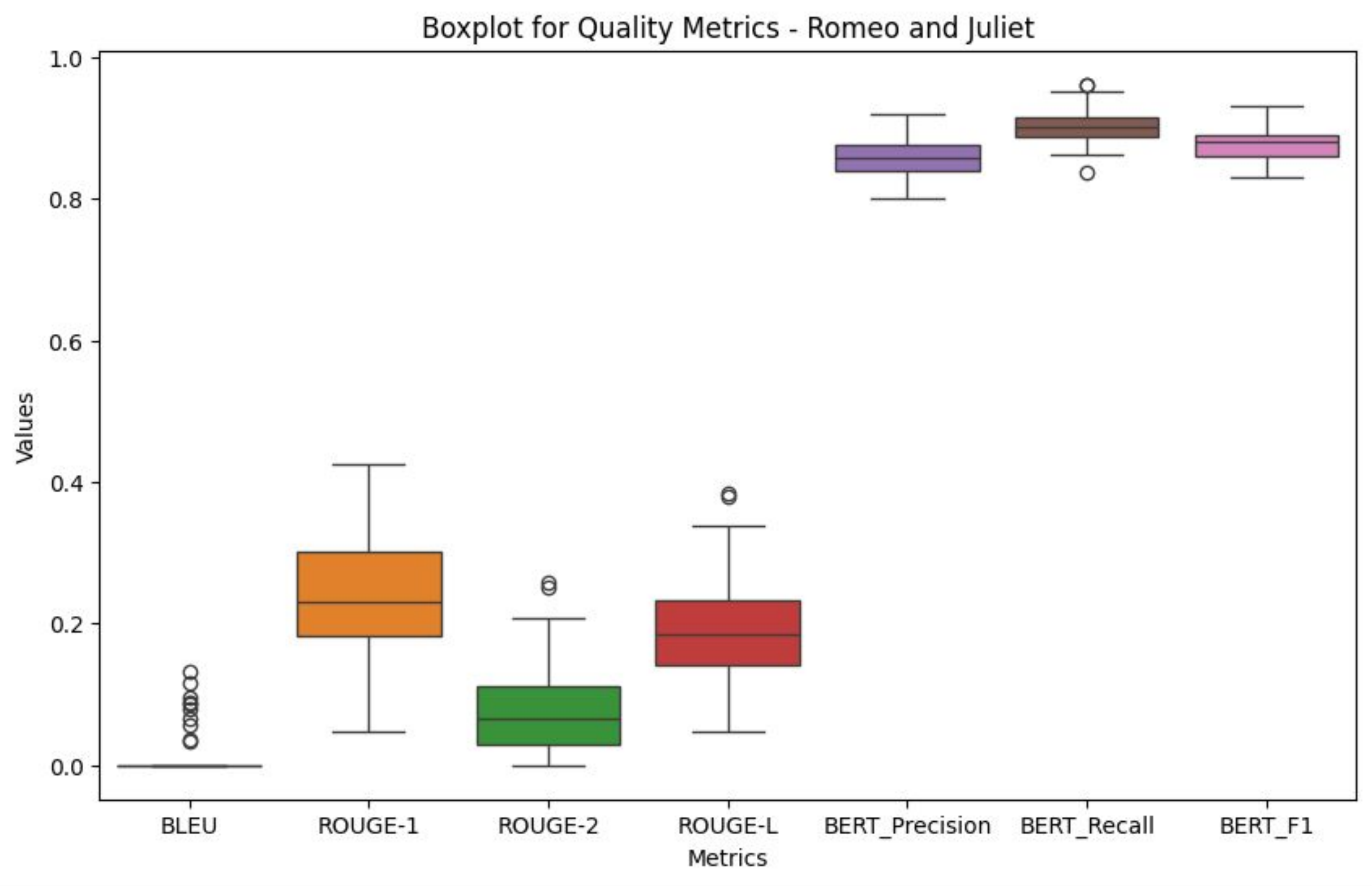
Evaluation

- Goal: assess the quality of both RAG's answers and retrieval;
- Focus on facts retrieval and not open-ended questions;
- Test questions and answers generated by an LLM: 48 questions for *Romeo and Juliet*, and 44 for *Odyssey*;
- RAG's vector populated with all books

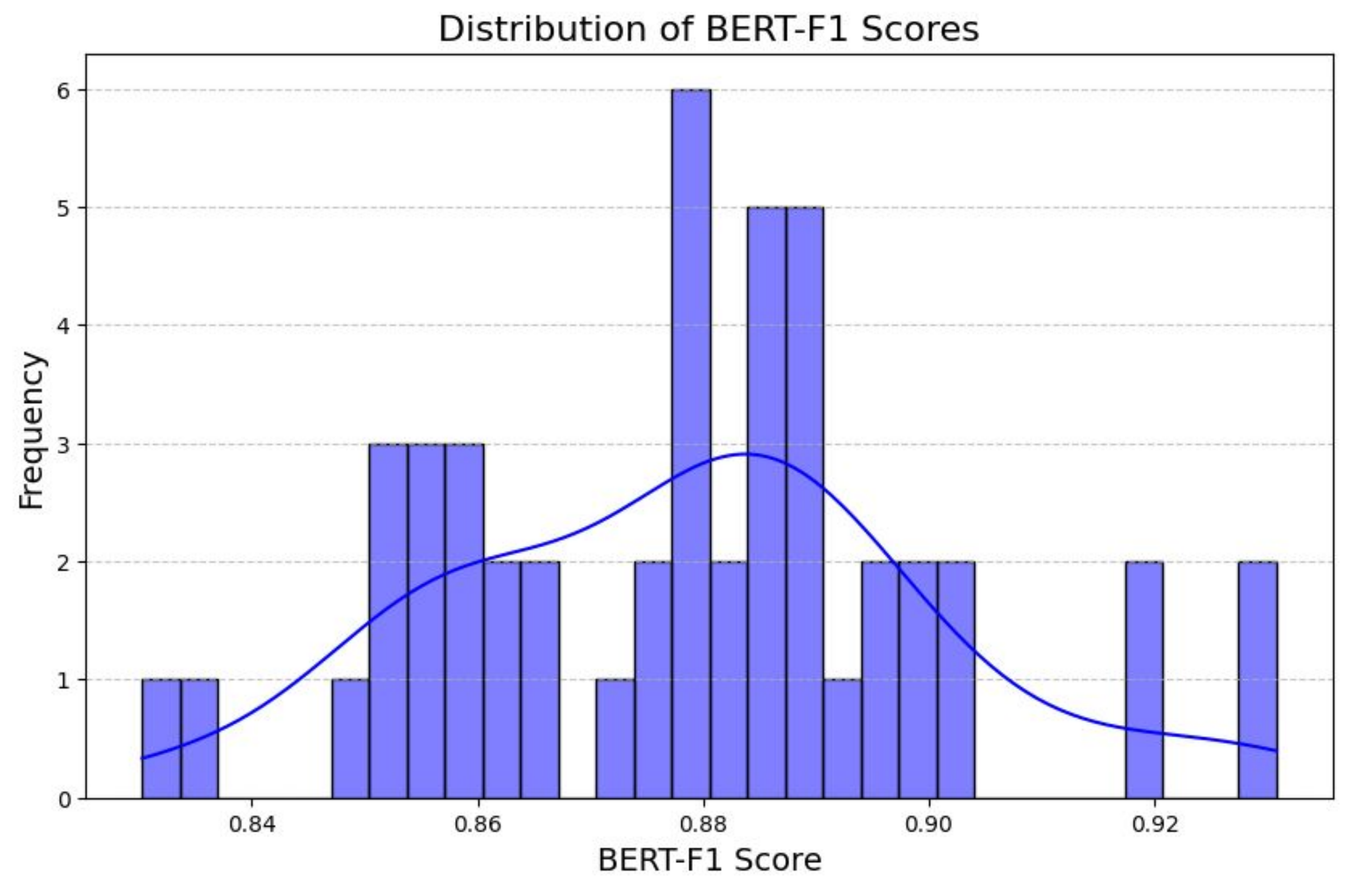
Evaluation metrics

- **BLEU:** Measures the overlap of n-grams between the generated and reference text;
- **ROUGE:** A family of metrics (ROUGE-1, ROUGE-2, ROUGE-L) that evaluate the recall of n-grams and longest common subsequences;
- **BERT-Based Scores:** Metrics such as BERT Precision, BERT Recall, and BERT F1, which use deep contextual embeddings to evaluate semantic similarity between sentences.

Results for Romeo and Juliet



Results for Romeo and Juliet



Example Q&A with high scores

1. **How does Benvolio try to help Romeo?**

Answer: Benvolio advises Romeo to forget Rosaline and look at other women to ease his heartache.

RAG Output (BERT F1 = 0.9308): Benvolio tries to help Romeo by suggesting he forget about Rosaline and examine other women, and by encouraging him to join in the party to distract him from his love for Rosaline.

2. **What is Romeo's reaction to his banishment?**

Answer: Romeo is devastated by his banishment, feeling it is worse than death because it separates him from Juliet.

RAG Output (BERT F1 = 0.9284): Romeo reacts with extreme distress and despair to his banishment, viewing it as worse than death due to his deep love for Juliet and the separation it imposes between them. He believes he has been "banish'd from the world" and that his life has ended.

Example Q&A with low scores

1. **What is Romeo's mood at the beginning of the play?**

Answer: Romeo is melancholic and lovesick over Rosaline, who does not return his affections.

RAG Output (BERT F1 = 0.8303): Romeo seems to be in a state of confusion and anxiety at the beginning of the play, as he questions his feelings towards his recent dance partner and worries about the consequences of his love for a girl from his family's enemy. He also expresses a sense of urgency to attend the party despite his misgivings, indicating a mix of excitement and apprehension. (approx. 35 words)

2. **What is the significance of the poison?**

Answer: The poison represents Romeo's desperate choice, believing Juliet is truly dead.

RAG Output (BERT F1 = 0.8521): The poison, particularly arsenic, serves as a powerful tool for murder that leaves no physical trace, making it difficult to detect and prosecute the perpetrator, highlighting the contrast between Eastern and Western approaches to crime and punishment. Its use reflects the sophistication of chemical knowledge and the potential for both malicious intent and hidden motives in the characters' actions.

Test for hallucinations

Test Process:

- A random sample of 10 outputs was reviewed.
- All outputs correctly matched the retrieved passages.

Example:

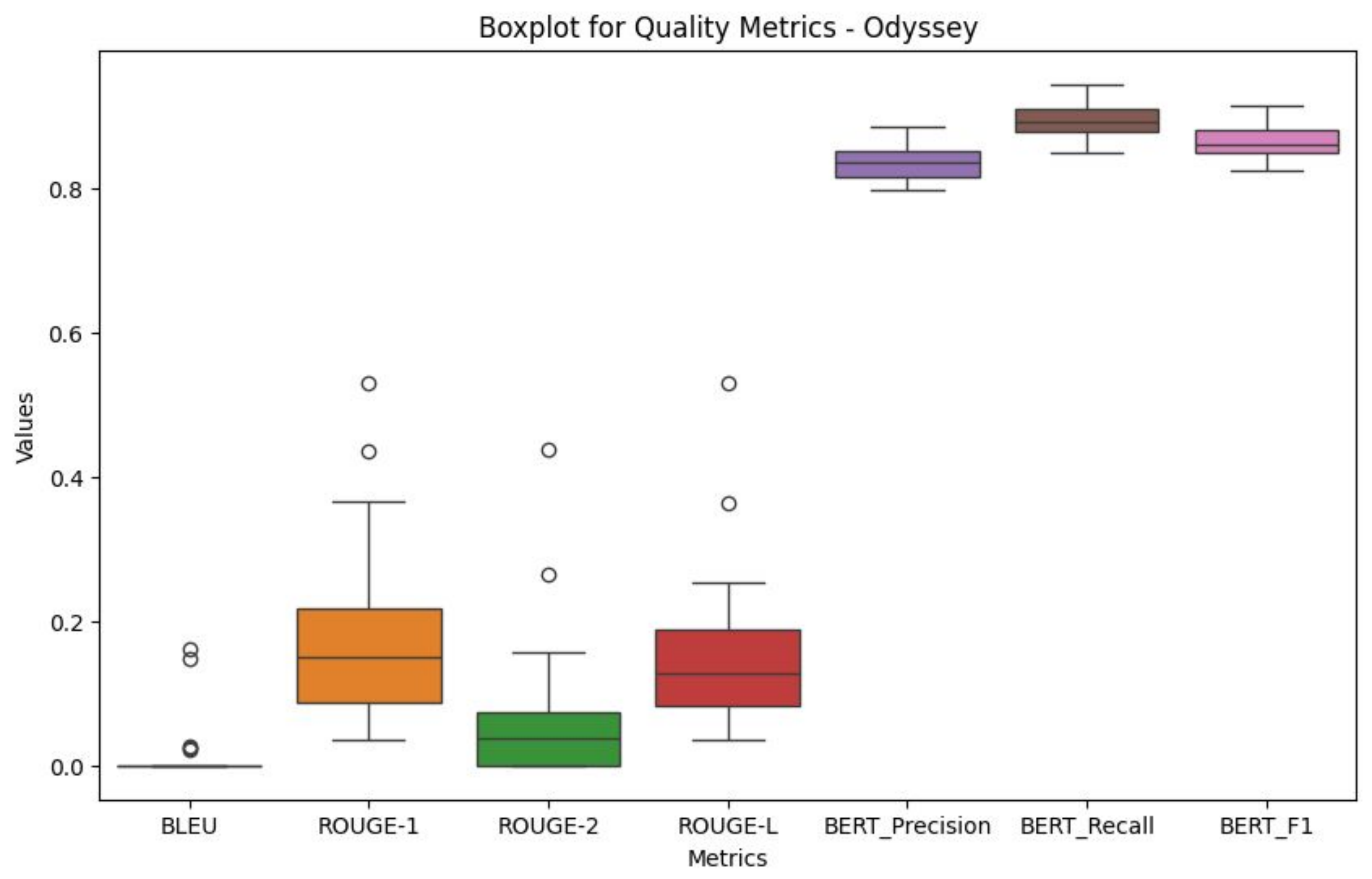
- Question:

  How does Juliet react when she discovers Romeo's identity?

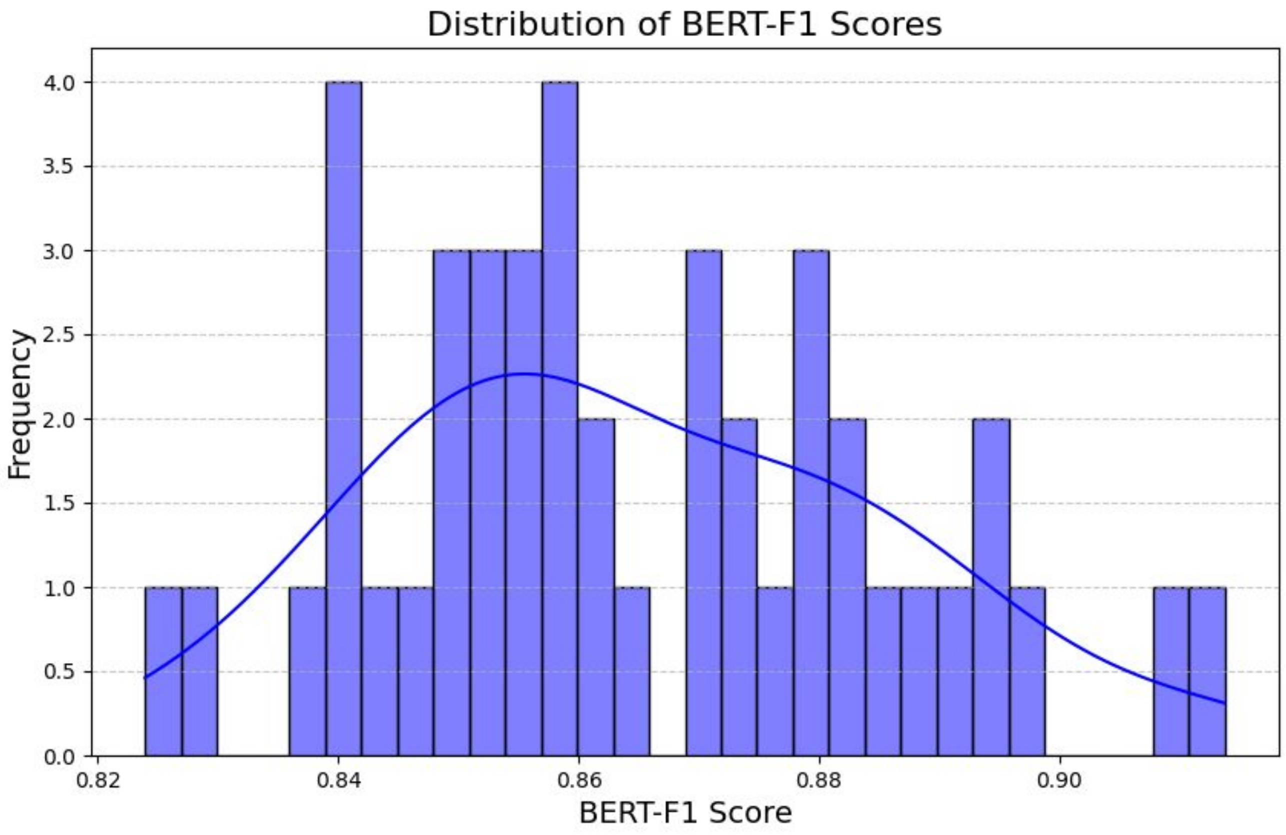
- RAG Output:

  When Juliet discovers Romeo's identity, she is shocked and conflicted, as he is the son of her family's long-standing enemy, prompting her to exclaim, ""My only love sprung from my only hate!"" and expressing the depth of her newfound love despite their families' feud. <u>She is also dismayed by his banishment, leading her to seek a solution through Friar Lawrence.</u>

Results for Odyssey



Results for Odyssey



Example Q&A with high scores

1. **How long does Odysseus stay on Calypso’s island?**

Answer: Odysseus stays on Calypso’s island for seven years.

RAG Output (BERT F1 = 0.9102): According to Odysseus’ account, he stayed on Calypso’s island for seven years before she sent him on his way of her own free will.

2. **What effect does the lotus plant have on Odysseus’s men?**

Answer: The lotus plant makes them forget their desire to return home.

RAG Output (BERT F1 = 0.8930): The lotus plant causes Odysseus's men to become indifferent to home and desire to stay and consume the plant without wanting to return, despite weeping bitterly. This shows the powerful and intoxicating effect the lotus has on those who consume it.

Example Q&A with low scores

1. **What is the fate of the Phaeacian ship that brings Odysseus home?**

Answer: Poseidon turns the ship into stone as punishment.

RAG Output (BERT F1 = 0.8365): The Phaeacian ship that brings Odysseus home is rooted in the sea just as it is getting into port by the earth-encircling Neptune, who then drives it down with the flat of his hand so as to root it in the ground

2. **What is the omen that confirms Odysseus's return to Ithaca?**

Answer: An eagle carrying a dead goose appears as an omen.

RAG Output (BERT F1 = 0.8521): Based on the passage, the omen confirming Odysseus's return to Ithaca is the sight of a hawk, Apollo's messenger, carrying a dove, which Theoclymenus interprets as a sign that Telemachus will remain powerful and his house will be the most royal in Ithaca. However, this seems to be an interpretation of the omen for Telemachus rather than a direct confirmation of Odysseus's return. The actual omen for Odysseus's return is not explicitly stated in the given passages. The passage focuses more on the gathering of the gods and Telemachus's departure. The mention of the hawk and dove seems to be a separate event not directly tied to Odysseus's return. If an omen for Odysseus's return is needed, it would likely be inferred from the overall context of the passage, which suggests that the gods are favoring Odysseus's journey home despite (the answer was too long to be shown in full).

Errors and mix-ups

- Excerpts from “Iliad” instead of “Odyssey”
- Footnotes, page numbers
- Information from beyond the passage (innate knowledge of the LLM)
- Too long answers

Thank you for your attention!