

Московский авиационный институт  
(национальный исследовательский университет)

Факультет информационных технологий и прикладной  
математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Информационный поиск»

Студент: А. Р. Белушкин  
Преподаватель: А. А. Кухтичев  
Группа: М8О-407Б-22  
Дата:  
Оценка:  
Подпись:

Москва, 2025

# 1 Источники данных

Для лабораторной работы был выбран корпус текстов из двух источников:

## 1. Project Gutenberg

- **Сайт:** <https://www.gutenberg.org>
- **Содержит:** тексты классической литературы в формате Plain Text UTF-8.
- **Примеры документов:**
  - The Adventures of Sherlock Holmes
  - Pride and Prejudice

## 2. Standard Ebooks

- **Сайт:** <https://standardebooks.org>
- **Содержит:** тексты классической литературы, аккуратно отформатированные для электронных книг.
- **Примеры документов:**
  - The Red Thumbmark
  - The Call of the Wild
- Для анализа использовались тексты со страниц, заканчивающихся на `text/single-page`

# 2 Характеристика корпусов

Корпус	Формат документов	Доп. мета-информация	Разметка текста
Gutenberg	Plain Text UTF-8	Название, автор, год публикации	Заголовки, шапка и футер
Standard Ebooks	HTML → извлечённый TXT	Название, автор, год публикации	HTML-разметка, абзацы

### Комментарии:

- Gutenberg содержит длинные художественные тексты.
- Standard Ebooks предоставляет тексты классической литературы, аккуратно структурированные и очищенные.

# 3 Выделение текста

## 3.1 Gutenberg

- Удалены первые и последние 50 строк (стандартные шапка и футер Project Gutenberg).

## 3.2 Standard Ebooks

- Из HTML извлечены только страницы `text/single-page`.
- Извлечены:
  - заголовки и подзаголовки
  - текстовые абзацы
- Игнорированы:
  - навигация по сайту
  - ссылки на внешние ресурсы
  - служебные блоки и изображения

## 4 Поиск по корпусу

### 4.1 Gutenberg

- Встроенный поиск: <https://www.gutenberg.org/ebooks/search>
- Примеры запросов: “Sherlock Holmes”, “Alice in Wonderland”
- Недостаток: полнотекстовый поиск недоступен.

### 4.2 Standard Ebooks

- Встроенный поиск: <https://standardebooks.org/ebooks>
- Примеры запросов: “The Red Thumbmark”, “The Call of the Wild”
- Недостатки:
  - поиск ограничен по метаданным и заголовкам
  - возможны промежуточные страницы без текста `single-page`

## 5 Статистическая информация о корпусах

Корпус	Кол-во документов	Размер сырых файлов (кбайт)	Средний размер
Gutenberg	75000	12172500	162.3
Standard Ebooks	100000	20180000	201.8

### Примечания:

- В Standard Ebooks «сырой» размер включает HTML-разметку.
- Размер текста учитывает только очищенный контент со страниц `text/single-page`.
- Разница в длине документов объясняется природой контента: книги разного объема.

## 6 Выводы

1. Оба корпуса подходят для анализа, так как представляют тексты художественной литературы разной природы и разной обработки.
2. Gutenberg содержит большие классические произведения, Standard Ebooks — аккуратно отформатированные тексты классики, готовые для электронных книг.
3. Поиск в Gutenberg и Standard Ebooks работает по разным принципам, что влияет на релевантность.
4. Основные недостатки:
  - Gutenberg: отсутствие полнотекстового поиска.
  - Standard Ebooks: необходимо фильтровать только страницы `text/single-page` для анализа.