МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. І. Сікорського

Кафедра

Інформатики та програмної інженерії

(повна назва кафедри, циклової комісії)

**КУРСОВА РОБОТА**

з дисципліни «Аналіз даних в інформаційних системах»

(назва дисципліни)

на тему: «Прогнозування успішності фентезі літератури за кількістю різноманітних угрупувань. Порівняння логістичної регресії і дерева прийняття рішень для прогнозування типу фентезі»

Студента 2 курсу, групи ІП-01

Спеціальності 121 «Інженерія програмного забезпечення»

Храмченко Анатолія Сергійовича

«ПРИЙНЯВ» з оцінкою

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

доц. Ліхоузова Т.А. / доц. Олійник Ю.О.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Підпис Дата

Київ- 2022 рік

КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. І. Сікорського

(назва вищого навчального закладу)

Кафедра інформатики та програмної інженерії

Дисципліна Аналіз даних в інформаційних системах

Спеціальність 121 «Інженерія програмного забезпечення»

Курс 2 Група ІП-01 Семестр 4

**ЗАВДАННЯ**

на курсову роботу студента

Храмченко А.С.

(прізвище, ім’я, по батькові)

1. Тема роботи Прогнозування успішності фентезі літератури за кількістю різноманітних угрупувань. Порівняння логістичної регресії і дерева прийняття рішень для прогнозування типу фентезі

2. Строк здачі студентом закінченої роботи 19.06.2022

3. Вхідні дані до роботи методичні вказівки до курсової роботи, дані з сайтів

<https://en.wikipedia.org>

https://www.litres.ru

4. Зміст розрахунково-пояснювальної записки (перелік питань, які підлягають розробці)

Постановка задачі

Аналіз предметної області

Розробка сховища даних

Інтелектуальний аналіз даних

5. Перелік графічного матеріалу ( з точним зазначенням обов’язкових креслень )

6. Дата видачі завдання 16.04.2022

КАЛЕНДАРНИЙ ПЛАН

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Назва етапів курсової роботи | Термін виконання етапів роботи | Підписи керівника, студента |
| 1. | Отримання теми курсової роботи | 16.04.22 |  |
| 2. | Визначення зовнішніх джерел даних | 25.04.22 |  |
| 3. | Пошук та вивчення літератури з питань курсової роботи | 02.05.22 |  |
| 4. | Розробка моделі сховища даних | 19.05.22 |  |
| 6. | Обґрунтування методів аналізу даних | 30.05.22 |  |
| 5. | Застосування та порівняння ефективності методів інтелектуального аналізу | 12.06.22 |  |
| 6. | Підготовка пояснювальної записки | 19.06.22 |  |
| 7. | Здача курсової роботи на перевірку | 19.06.22 |  |
| 8. | Захист курсової роботи | 19.06.22 |  |

Студент Храмченко А.С.

(підпис) (прізвище, ім’я, по батькові)

Керівник доц. Ліхоузова Т.А.

(підпис) (прізвище, ім’я, по батькові)

Керівник доц. Олійник Ю.О.

(підпис) (прізвище, ім’я, по батькові)

"16" березня 2022 р.

АНОТАЦІЯ

Пояснювальна записка до курсової роботи: 121 сторінка, 10 рисунків, 28 таблиць, 4 посилання.

Об’єкт дослідження: інтелектуальний аналіз даних

Мета дослідження: створення програмного забезпечення для для аналізу даних подальшим прогнозуванням

Дана курсова робота включає в себе: опис підготовки даних для заповнення бази даних, опис створення програмного забезпечення для інтелектуального аналізу даних та прогнозування за допомогою різних моделей.

МОДЕЛЬ ПРОГНОЗУВАННЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ЛОГІСТИЧНА РЕГРЕСІЯ, ДЕРЕВО ПРИЙНЯТТЯ РІШЕНЬ, ГРАДІЄНТНИЙ СПУСК, АНАЛІЗ ТЕКСТУ

Зміст

[Вступ 5](#_Toc106540974)

[1 Постановка задачі 6](#_Toc106540975)

[2 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ 7](#_Toc106540976)

[3 ЗАПОВНЕННЯ СХОВИЩА ДАНИХ 8](#_Toc106540977)

[3.1. Збір даних 8](#_Toc106540978)

[3.2. Аналіз опису книги 8](#_Toc106540979)

[3.3. Занесення даних до сховища 9](#_Toc106540980)

[3.4. Отримання даних зі сховища 9](#_Toc106540981)

[4 Тестування програмного забезпечення 11](#_Toc106540982)

[4.1. План тестування 11](#_Toc106540983)

[4.2. Приклади тестування 12](#_Toc106540984)

[5 Інструкція користувача 28](#_Toc106540985)

[5.1. Робота з програмою 28](#_Toc106540986)

[5.2. Формат вхідних даних 33](#_Toc106540987)

[5.3. Системні вимоги 33](#_Toc106540988)

[Висновки 35](#_Toc106540989)

[Перелік посилань 36](#_Toc106540990)

[Додаток А Технічне завдання 60](#_Toc106540991)

[Додаток Б Тексти програмного коду 63](#_Toc106540992)

Вступ

Фентезі – один з найпопулярніших жанрів книг. Успішність книги залежить від багатьох факторів: кількості рас, кількості сюжетних поворотів, кількості персонажів та видів в’язків між ними. При написанні книги автору дуже важливо враховувати всі ці аспекти, оскільки від цього залежить кількість людей, що прочитають книгу, а як наслідок – прибутковість. Також, автору потрібно знати, чи відповідає книга обраному піджанру. Наприклад, якщо автор хоче, щоб книга була орієнтована на дітей, то він повинен прагнути жанру світлого фентезі. Для цього йому потрібно перевірити жанр, якому відповідає його книга «стороннім оком».

Отже, нами було вирішено створити програмне забезпечення, що допоможе вирішити ці проблеми. А саме, на основі даних про кількість сторінок, рік виходу та короткому опису сюжету книги прогнозувати кількість читачів та визначити жанр.

В рамках даної курсової роботи було розроблено сховище даних на основі фізичної моделі бази даних, функціонал якої було розроблено за допомогою SQL скриптів та програмне забезпечення, що заповнює базу даних, отримує дані з неї, проводить інтелектуальний аналіз даних та прогнозує результат.

В ролі системи керування сховищем даних для даної роботи буде виступати MySQL, а мова програмування для реалізації застосунку – Python3.

# Постановка задачі

Під час виконання курсової роботи потрібно виконати наступні завдання:

Створення застосунку, що на основі даних, що містяться у файлах, аналізує їх текст та заповнює базу даних.

Створення застосунку, що отримує дані з бази даних та проводить інтелектуальний аналіз способом логістичної регресії за допомогою градієнтного спуску та способом дерева рішень.

Створення застосунку, що тестує моделі прогнозування.

Проаналізувати параметри на значемість та вирівняти дані за допомогою SMOTE.

()

Використати мову програмування Python 3 для реалізації застосунку.

Курсовий проект здати до дедлайну (початок сесії) та виконати у єдиному стилі написання коду (coding style).

# АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Сьогодні фентезі є актуальним жанром книжок. Жанр фентезі почав активно розвиватися лише у XX ст., а отже багато письменників бажають його доповнити своїми творами. Написання такої книги потребує багато зусиль, а отже автору потрібно знати, на яку кількість аудиторії його книга може бути розрахована. А також важливо на який вік орієнтована книга. Якщо автор хоче орієнтуватися на більш дорослу аудиторію, то йому потрібно прагнути до жанру темного фентезі. Отже, автору потрібно знати, якому жанру його книга більше відповідає для подальшого виправлення.

У програмному забезпеченні буде реалізовано наступну функціональність:

* Аналіз опису сюжету книг на різноманітні характеристики
* Заповнення бази даних
* Інтелектуальний аналіз даних
* Використання декількох моделей прогнозування даних
* Прогнозування жанру книги
* Прогнозування успішності книги

# ЗАПОВНЕННЯ СХОВИЩА ДАНИХ

## Збір даних

Для виконання курсової роботи ми вирішили зібрати про книгу наступні параметри: Назва, автор, кількість сторінок, рік виходу, оцінка на ресурсі LiveLib, оцінка на ресурсі LitRes, реальність світу (1 – напівреальний, 0 – повністю вигаданий), піджанр (темне чи світле фентезі) та кількість оцінок. А також короткий опис сюжету книги.

Ці дані було отримано з ресурсів:

* <https://en.wikipedia.org> (назва, автор, кількість сторінок, рік виходу)
* <https://www.litres.ru> (оцінки та їх кількість)

Ці дані було розміщено у файлах такої структури: перший рядок – назва, другий рядок – автор, третій рядок – кількість сторінок, четвертий рядок – рік виходу, п’ятий та шостий рядок – оцінки, сьомий рядок – реальність світу, восьмий рядок – кількість оцінок, дев’ятий рядок – піджанр, інші рядки – опис сюжету. Приклад заповненого файлу представлено на рисунку 3.1.

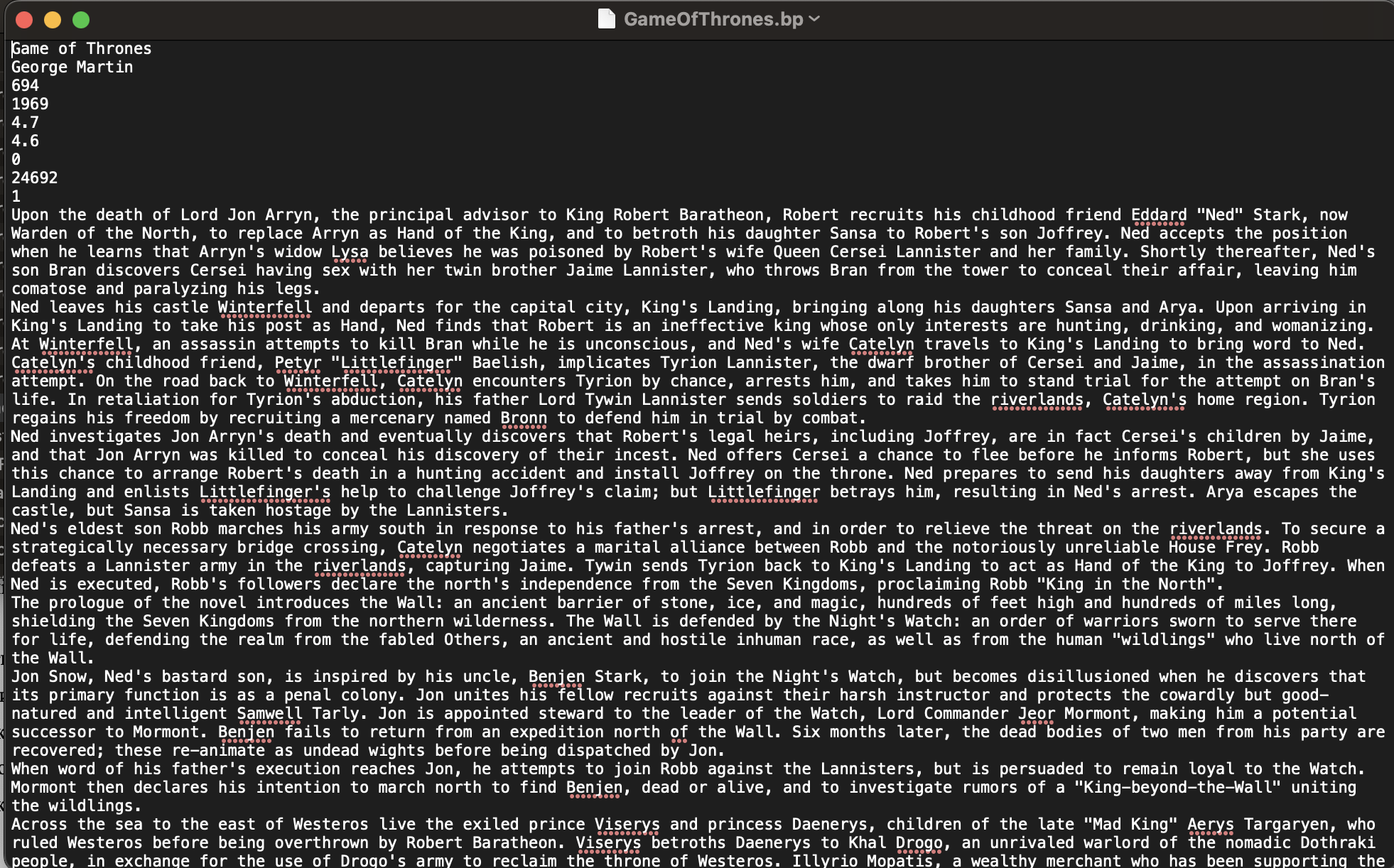


Рисунок 3.1 – Приклад заповненого файлу

Таким чином було створено 125 файлів. 20% даних було відведено для подальшого тестування. Їх назви записані у файл “test.bp”. Інші файли, призначенні для навчання, були записані до файлу “books.bp”.

## Аналіз опису книги

З короткого опису сюжету нами було вирішено виокремити такі характеристики:

* Кількість персонажів
* Відносна кількість чоловічих та жіночих персонажів
* Кількість сюжетних поворотів
* Кількість локацій
* Кількість рас
* Відносна кількість любовних, сімейних, дружніх та ворожих стосунків

Кількість персонажів рахується як кількість імен, що були у тексті. Список всіх імен було отримано з відкритого джерела даних: <https://parse-dashboard.back4app.com/apps/42b8ca5e-4c7a-4fae-bec7-b937694499d6/browser/NamesList>

При зустрічі чоловічого імені збільшувалася кількість чоловіків, при зустрічі жіночого – кількість жінок.

Для визначення інших параметрів було створено файли з переліком усіх ключових слів, що стосуються цих тем.

Для того, щоб усі книги були у рівних умовах усі абсолютні показники було поділено на кількість слів у описі. Таким чином довжина опису ніяк не буде впливати на параметри.

Код аналізу текстів представлено у додатку А, файли Book.py та TextWorker.py.

## Занесення даних до сховища

Код створення сховища та занесення даних представлено у додатку А, файл DBWorkre.py та детально описано у роботі Берлінського Я.В.

Результат заповнення сховища даних зображено на рисунку 3.2

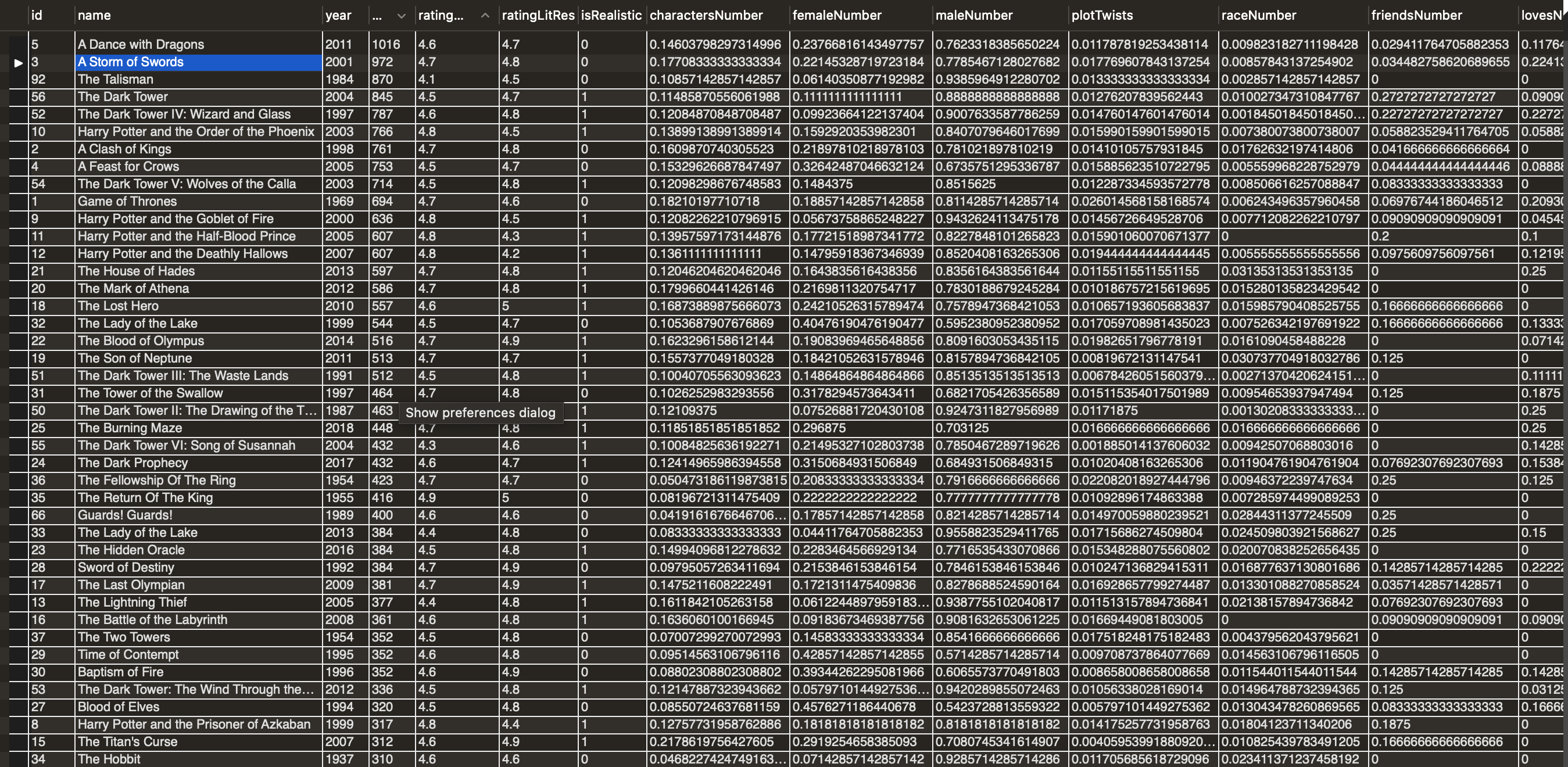


Рисунок 3.2-Скріншот заповненого сховища даних

## Отримання даних зі сховища

При отримані даних зі сховища було вирішено звести кожен параметр до проміжку [0;1]. Це потрібно для пришвидшення роботи алгоритмів навчання.

Для досягнення цієї мети була використана така формула (Формула 3.1):

* + - * 1. Формула для зведення параметрів у проміжок [0;1], де x – значення параметру, xmin – мінімальне значення параметру, xmax – максимальне значення параметру

Таким чином, мінімальне значення будь-якого параметру набуде 0, а максимальне – 1. Код зміни даних представлено у додатку А, файл DataWorker.py

# ІНТЕЛЕКтуальний аналіз даних

## Обґрунтування вибору методів інтелектуального аналізу даних

Для вирішення задачі прогнозування жанру книги (задача класифікації) потрібно використовувати інтелектуальний аналіз даних. Спираючись на дані про відомі книги, ми зможемо створити математичну модель, що допоможе прогнозувати належність книги до темного чи світлого фентезі. Для алгоритму інтелектуального аналізу важливі 2 характеристики: швидкість та якість. Тому було обрано 2 методи інтелектуального аналізу для порівняння:

* Логістична регресія з методом градієнтного спуску. (якість)
* Дерево прийняття рішень. (швидкість)

## Теоретичне обґрунтування логістичної регресії та методу градієнтного спуску

Логістична регресія використовується для того, щоб прогнозувати дані, що можуть набувати значень 0 чи 1. Для цього використовується сигмоїда. (Формула 4.1)

* + - * 1. Формула сигмоїди

Ця функція не може набувати значень більших за 1 та менших за 0. Можемо побачити це на груфіку. Рисунок 4.1

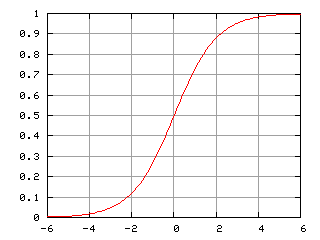


Рисунок 4.1 – Графік сигмоїди

Ця функція залежить від z. Оскільки ми маємо багато параметрів, то виразимо z певною лінійною функцією. (Формула 4.2)

* + - * 1. Лінійна функція z, де xi – jй параметр, – jй коефіцієнт, n – кількість параметрів

Нам потрібно знайти найбільш підходящі . Таким чином, нам потрібно мінімізувати різницю між y та f(z). В результаті математичних перетворень отримаємо, що нам потрібно мінімізувати функцію у формулі 4.3.

* + - * 1. Формула cost-функції для логістичної регресії

Для того, щоб підібрати параметри для F при яких F набуває найменшого значення потрібно часткову похідну від кожної змінної прирівняти до 0 та розв’язати систему рівнянь. Оскільки рівняння не є лінійними, то зробити це програмно важко. Тому використовують метод градієнтного спуску, що дозволяє приблизитись до оптимальних значень параметрів. Метод градієнтного спуску полягає у тому, що параметри функції змінюються короткими кроками, в результаті чого значення функції поступово зміщується до мінімуму. Для градієнтного спуску у логістичній регресії використовується формула 4.4

* + - * 1. Формула для градієнтного спуску

## Теоретичне обґрунтування дерева прийняття рішень

Метод дерева прийняття рішень полягає у тому, що на основі відомих даних формується бінарне дерево. Кожен вузол дерева оцінює певний параметр та в результаті оцінки відправляє його до нащадка. Результатом такого проходу буде потрапляння до листа дерева, що містить значення 1 або 0 – результат прогнозу.

## Підготовка даних до інтелектуального аналізу

Розглянемо кожний з зібраних факторів для того, щоб визначити їх вплив на результат. Знайдемо середнє значення кожного із параметрів при світлому фентезі, темному фентезі та взагалі. Якщо різниця між середнім та середнім при умові буде більшою за 0.05, то фактор можна вважати значимим. Графіки залежностей представлено на рисунках 4.2-4.16. Фактори розглянуті у таблиці 4.1.

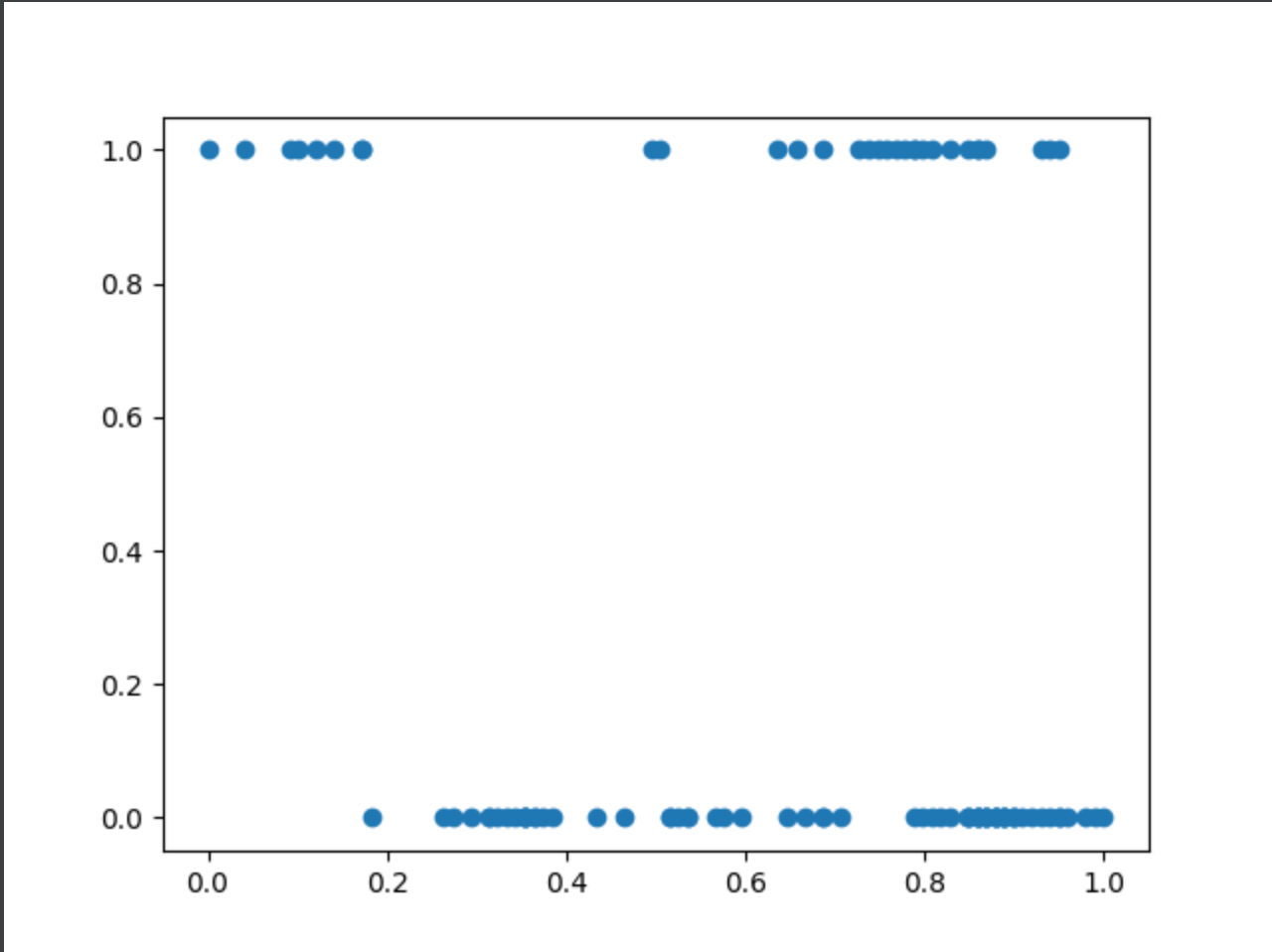


Рисунок 4.2 - Графік залежності жанру фентезі від року виходу

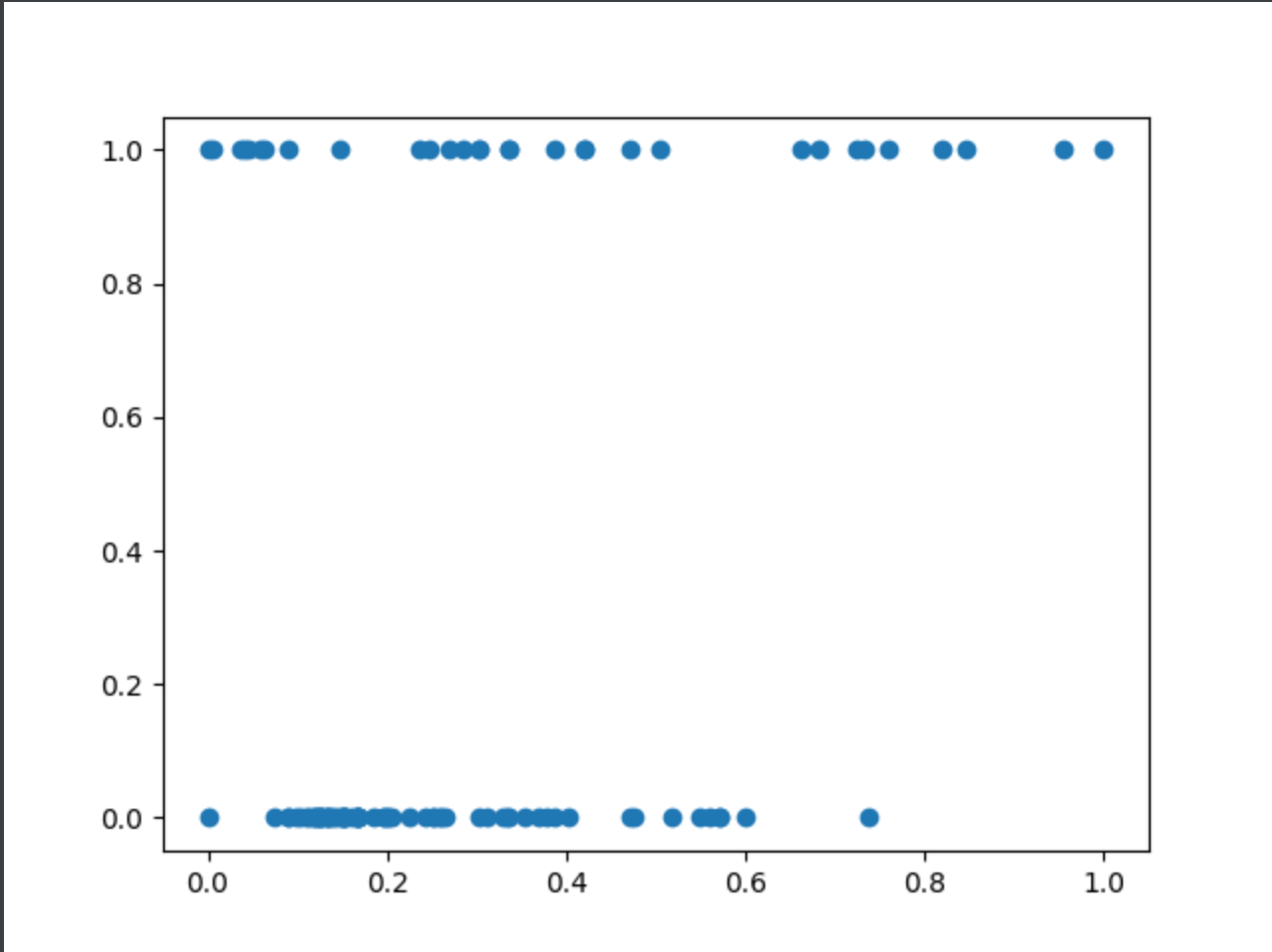


Рисунок 4.3 - Графік залежності жанру фентезі від кількості сторінок

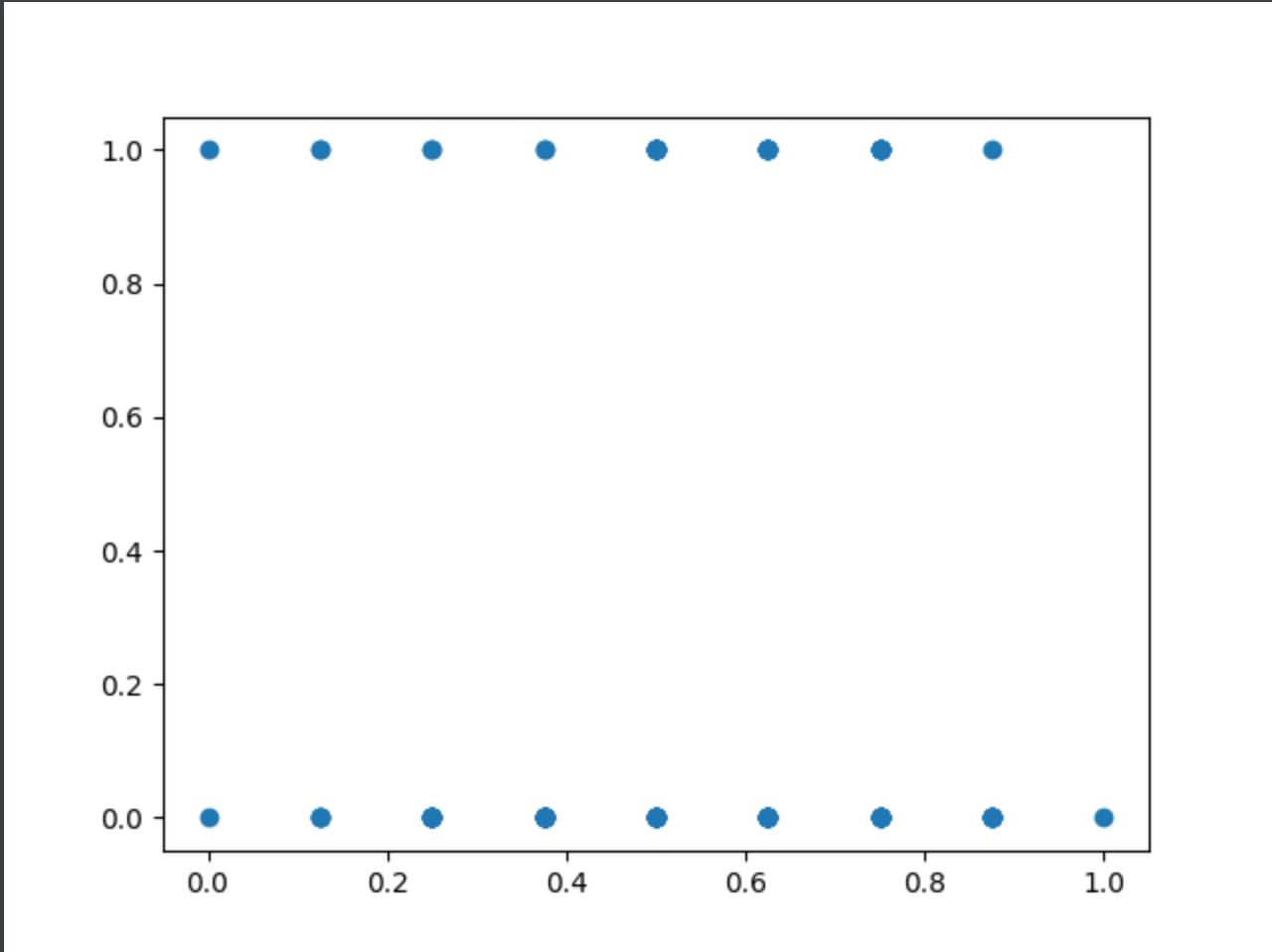


Рисунок 4.4 - Графік залежності жанру фентезі від рейтингу на LiveLib

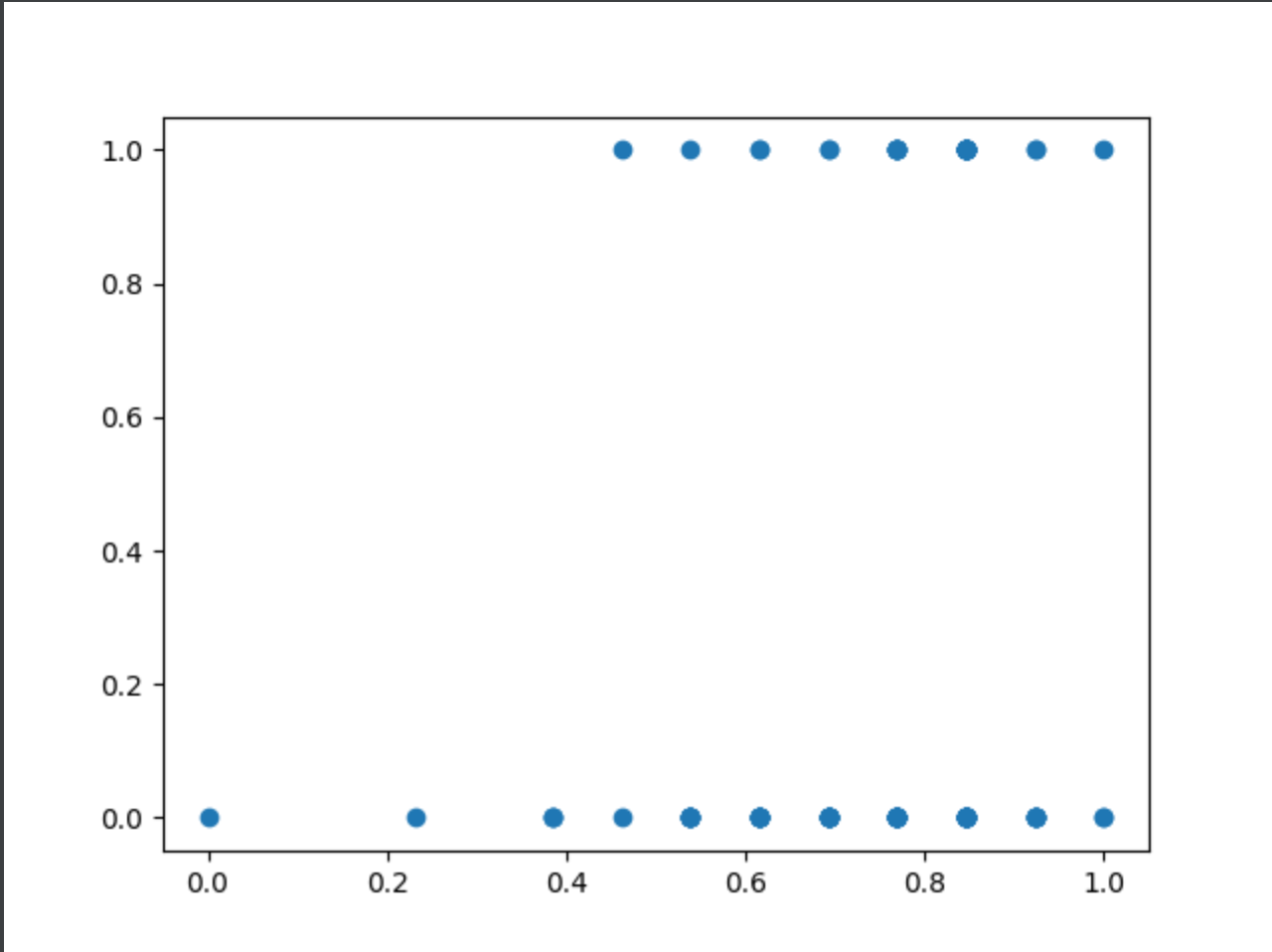


Рисунок 4.5 - Графік залежності жанру фентезі від рейтингу на LitRes

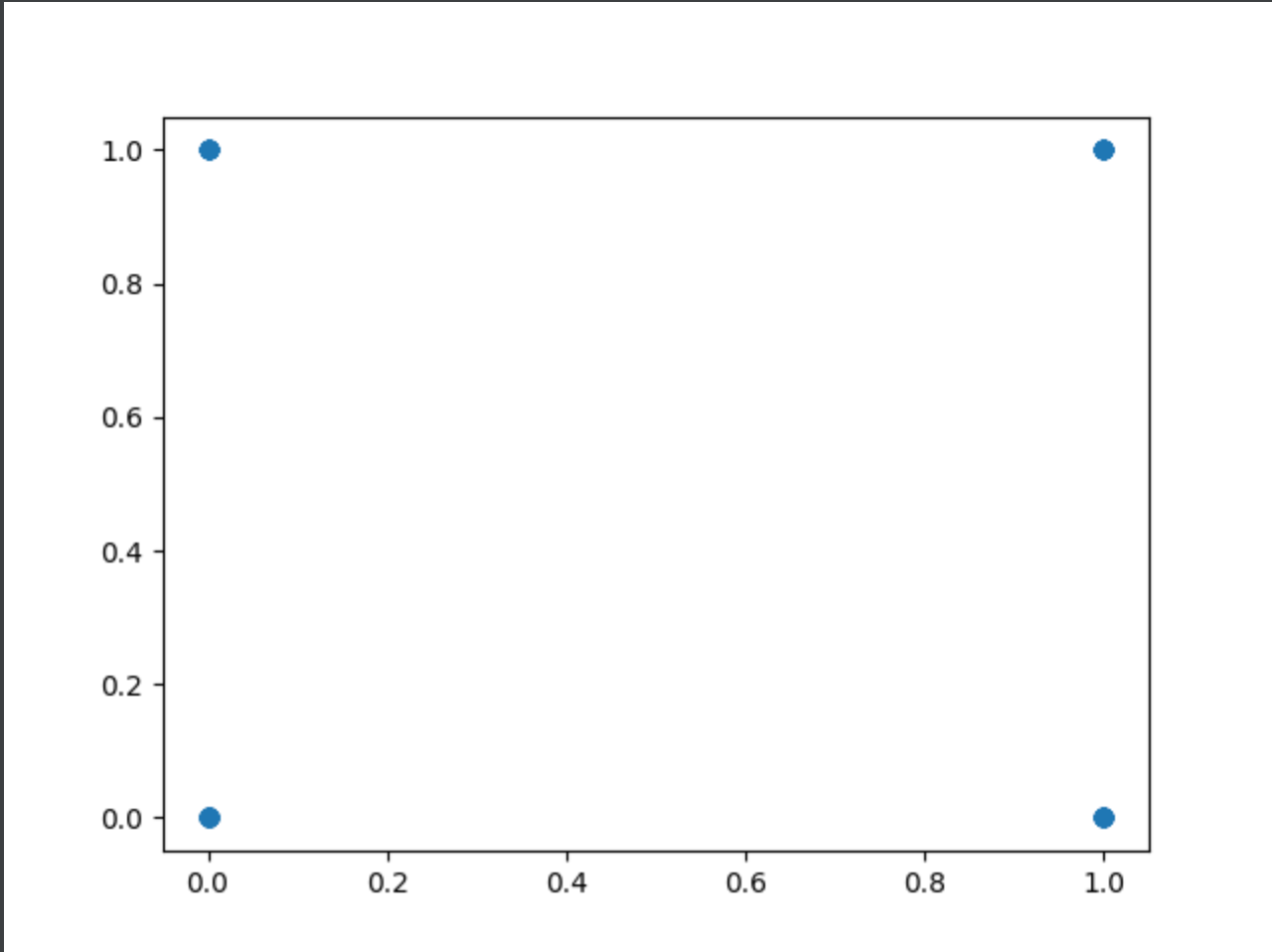


Рисунок 4.6 – Графік залежності жанру фентезі від реалістичності світу

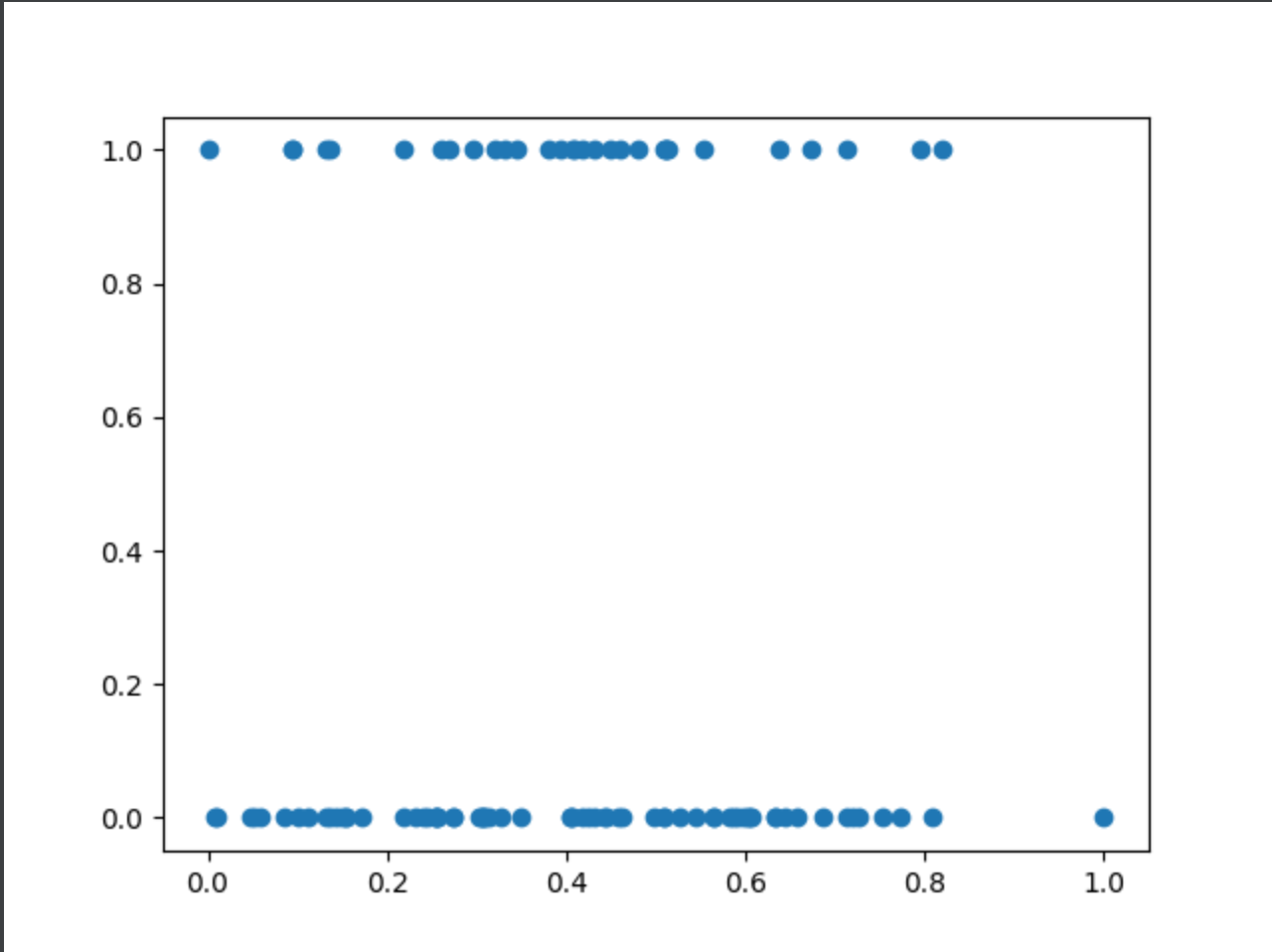


Рисунок 4.7 – Графік залежності жанру фентезі від кількості персонажів

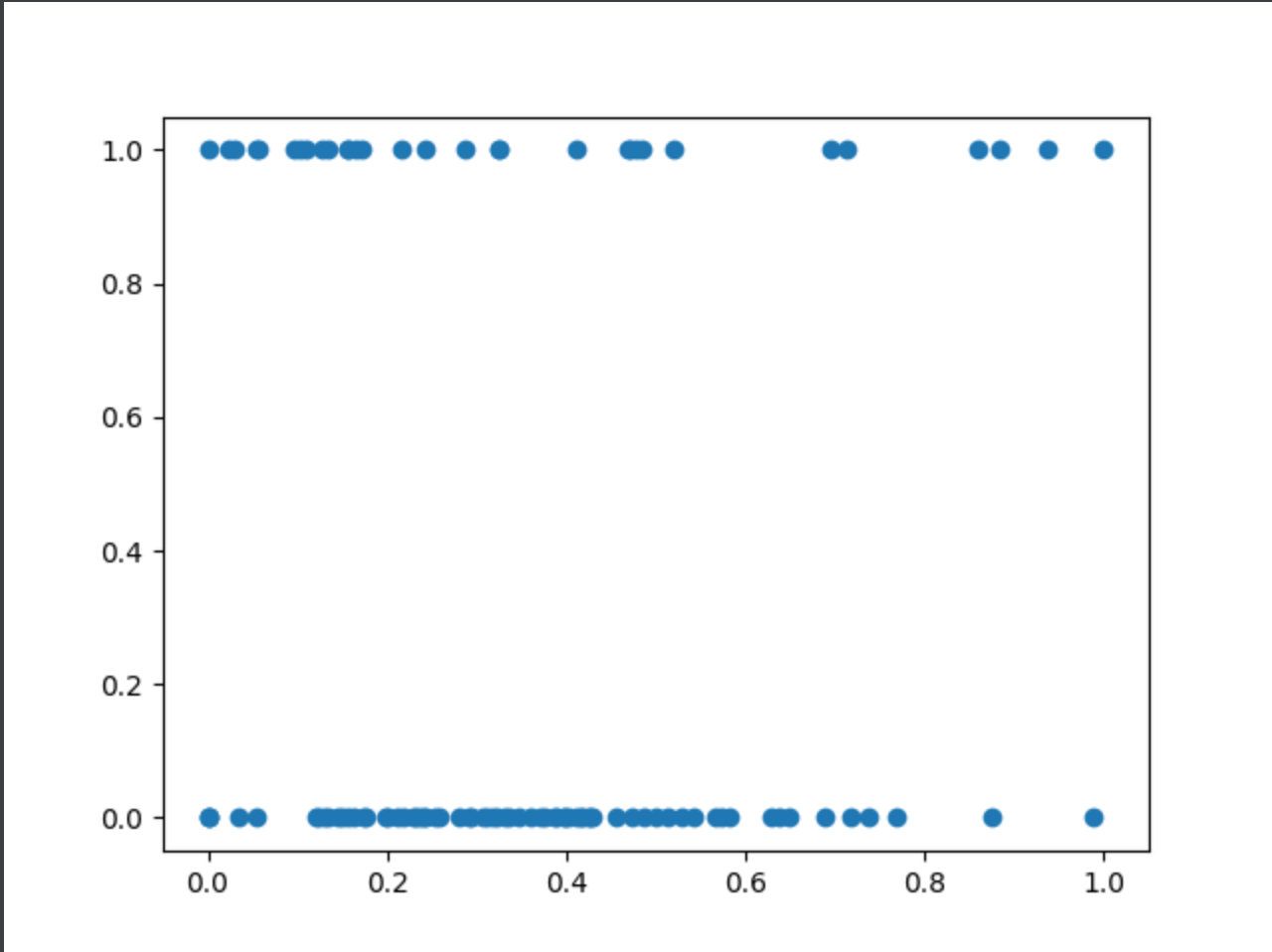


Рисунок 4.8 – Графік залежності жанру фентезі від відсотку жінок

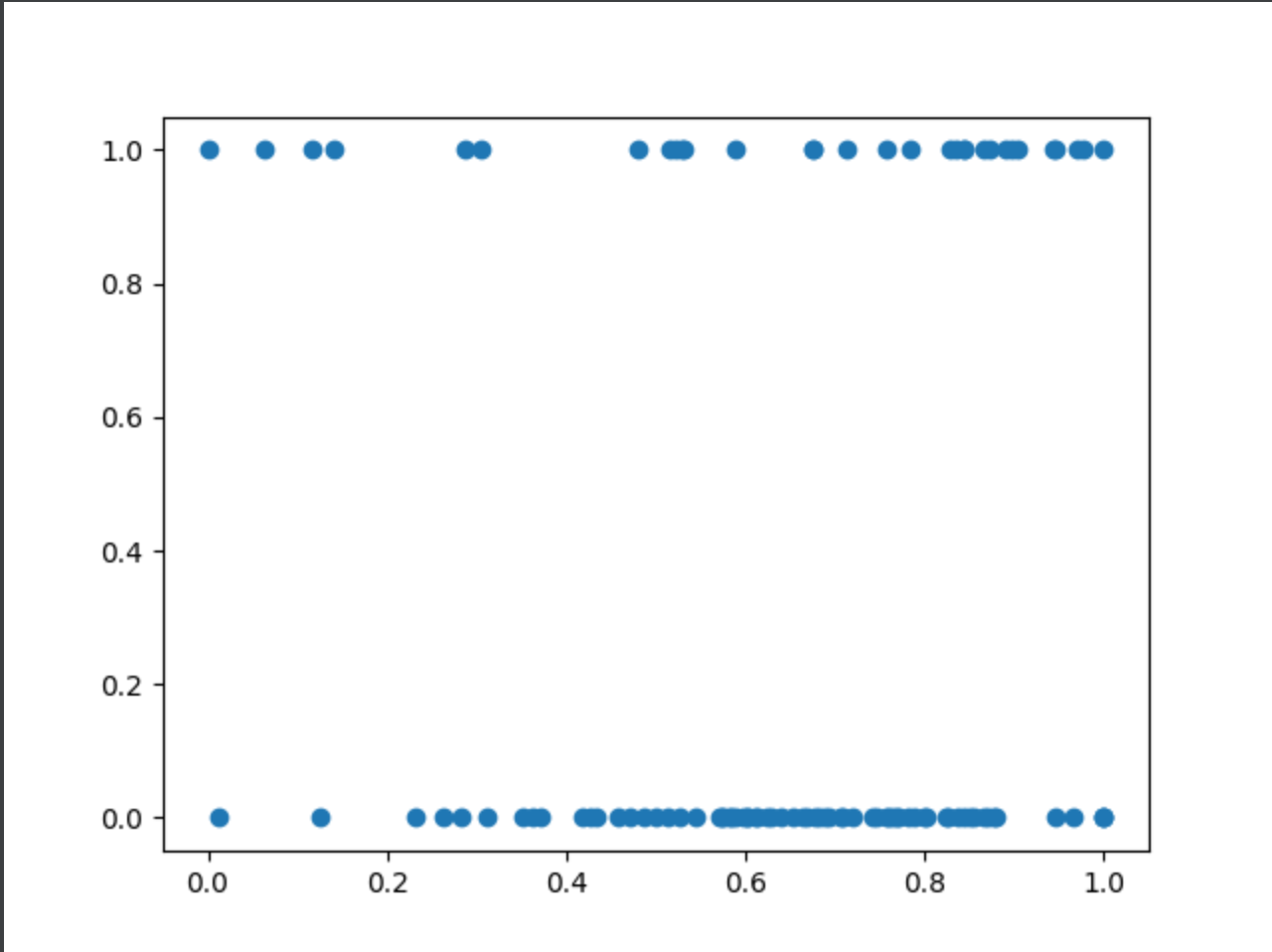


Рисунок 4.9 – Графік залежності жанру фентезі від відсотку чоловіків

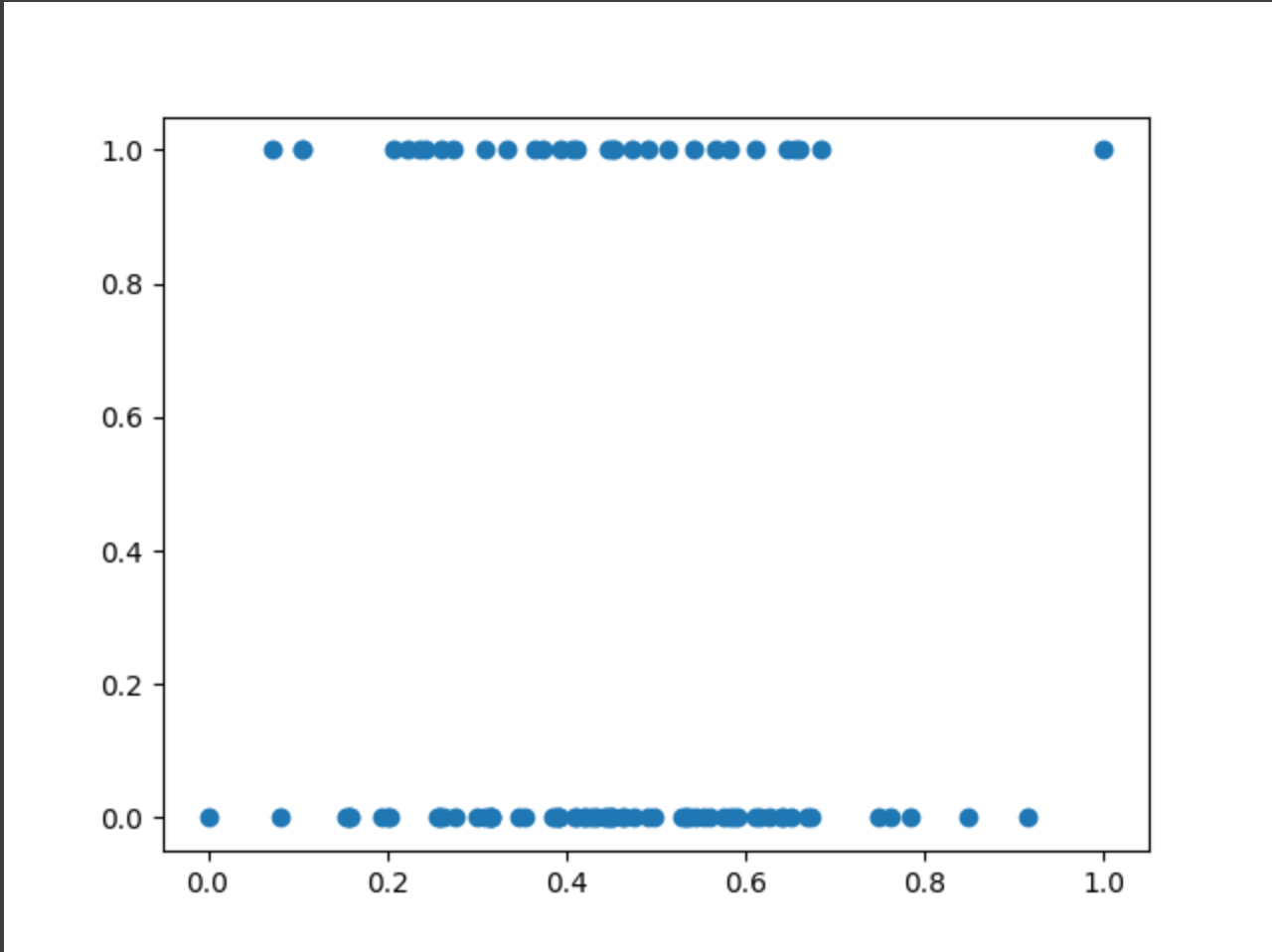


Рисунок 4.10 – Графік залежності жанру фентезі від кількості сюжетних поворотів

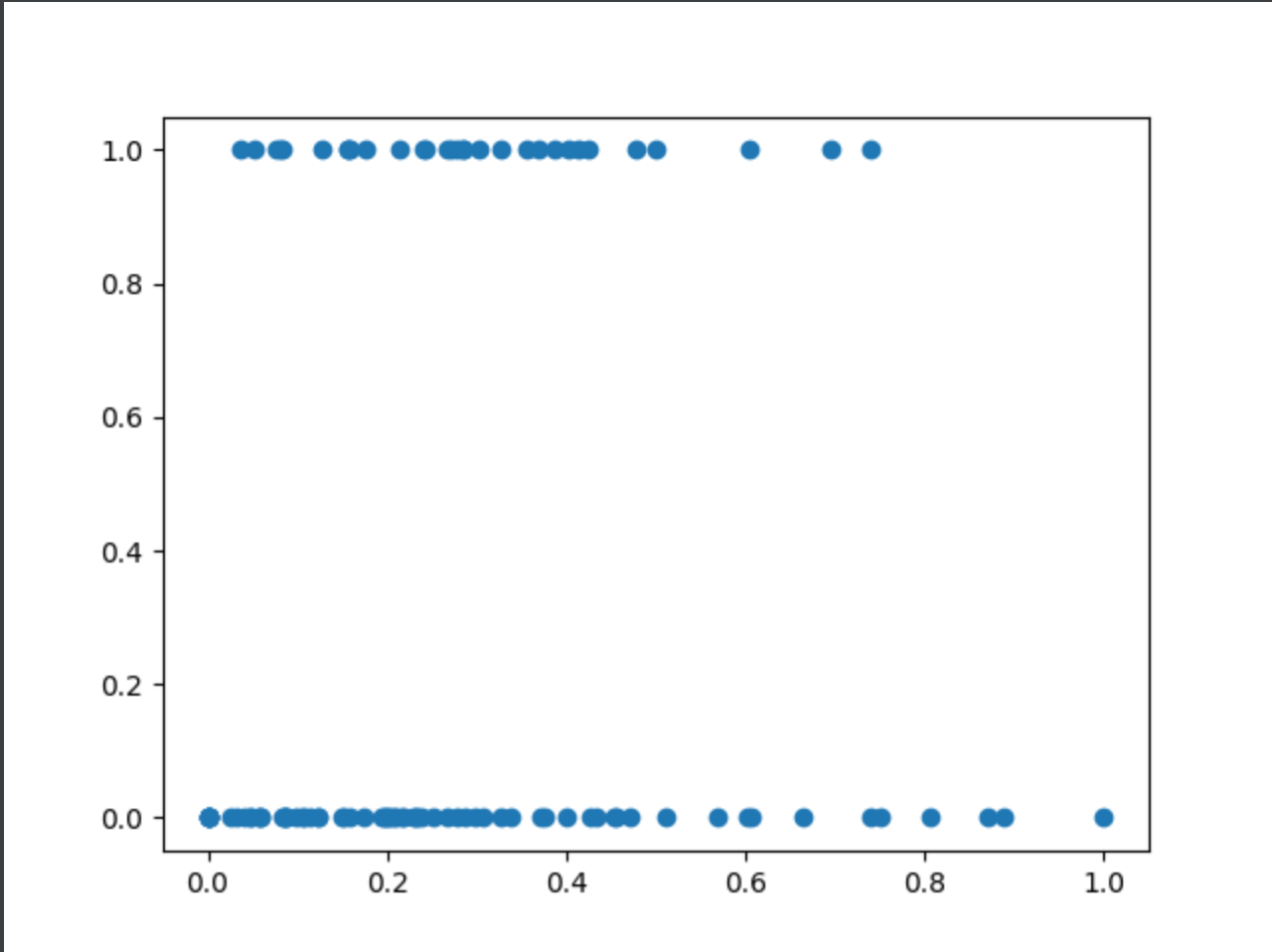


Рисунок 4.11 – Графік залежності жанру фентезі від кількості рас

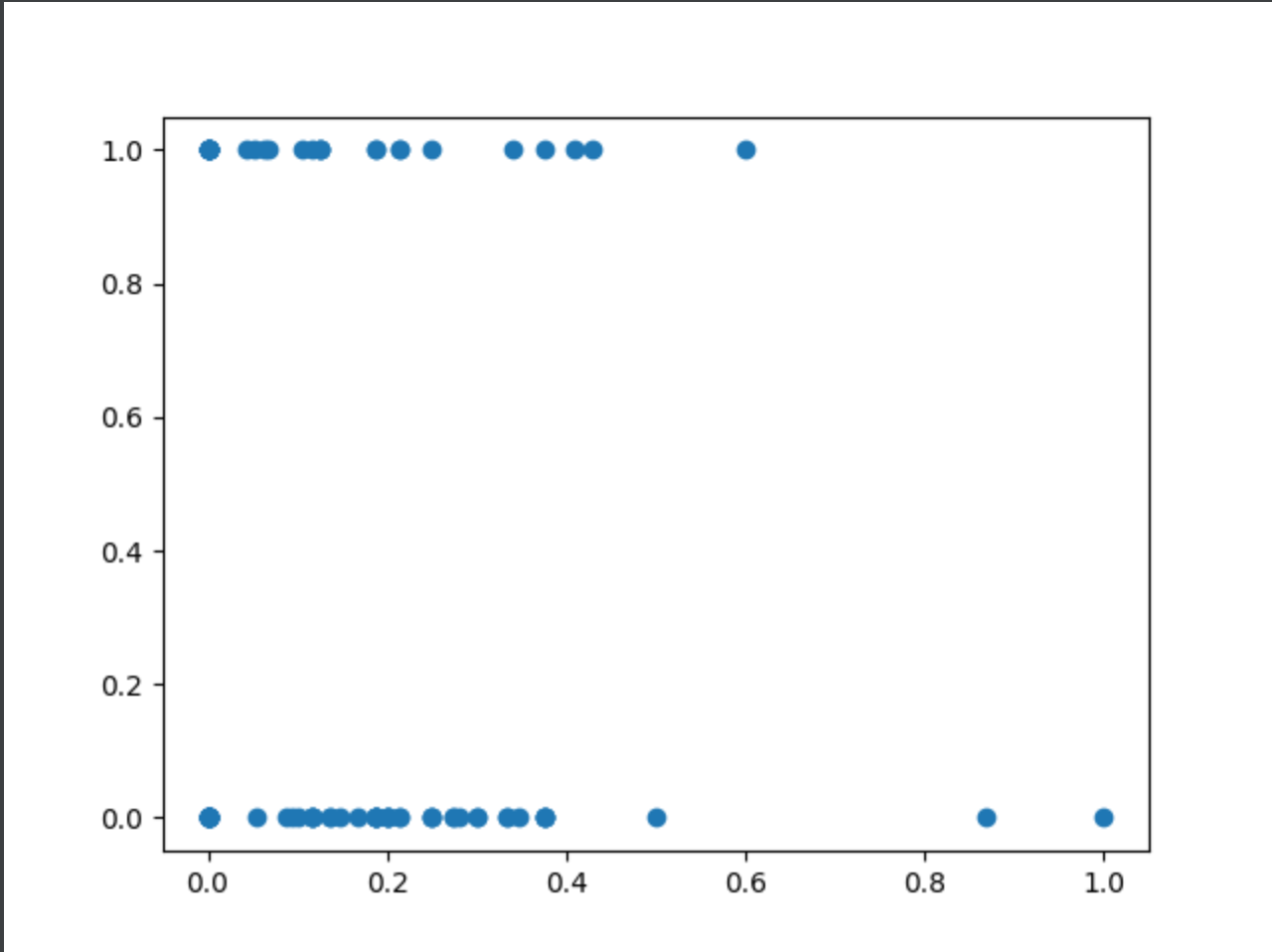


Рисунок 4.12 – Графік залежності жанру фентезі від проценту дружніх стосунків

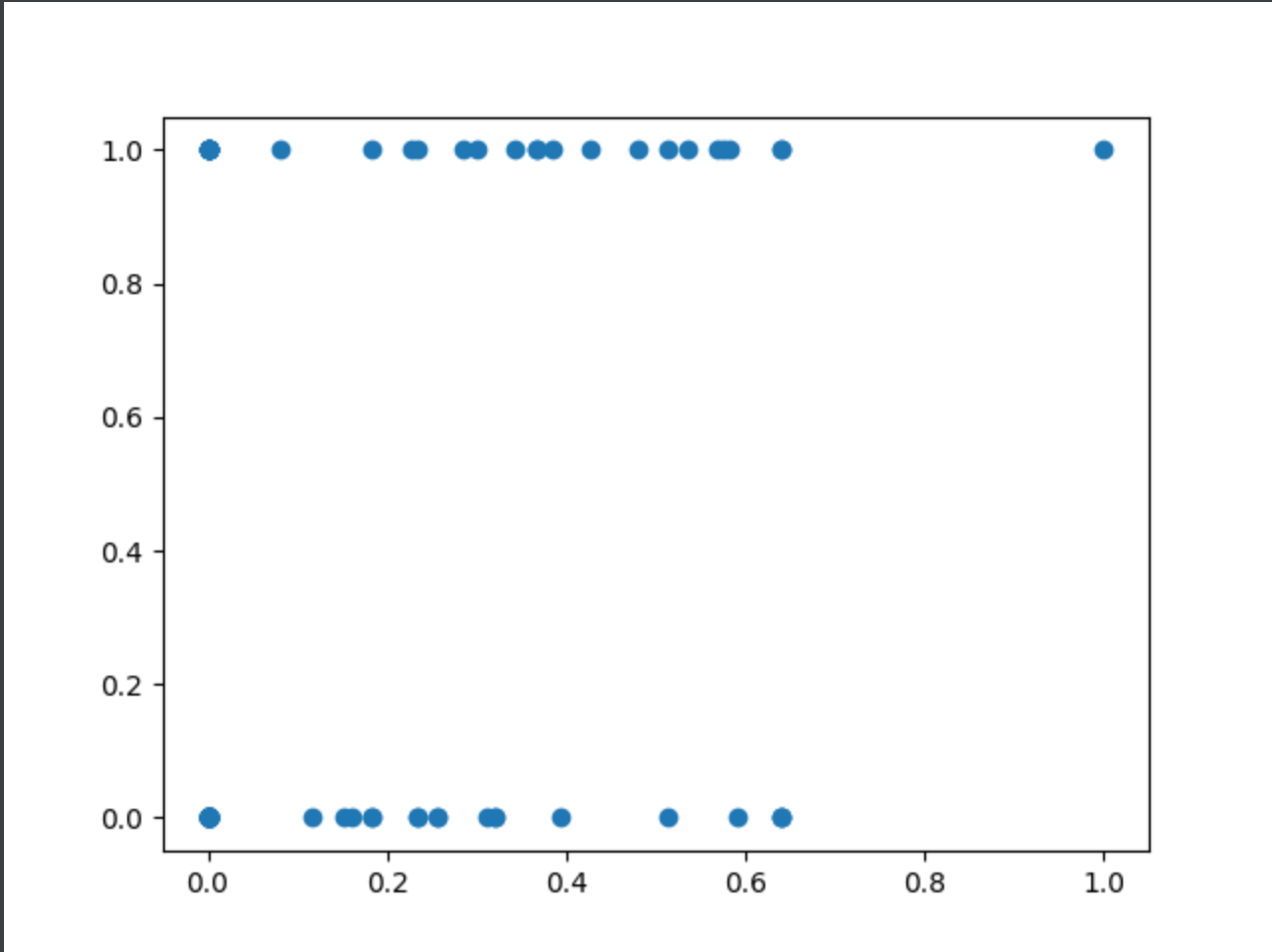


Рисунок 4.13 – Графік залежності жанру фентезі від проценту любовних стосунків

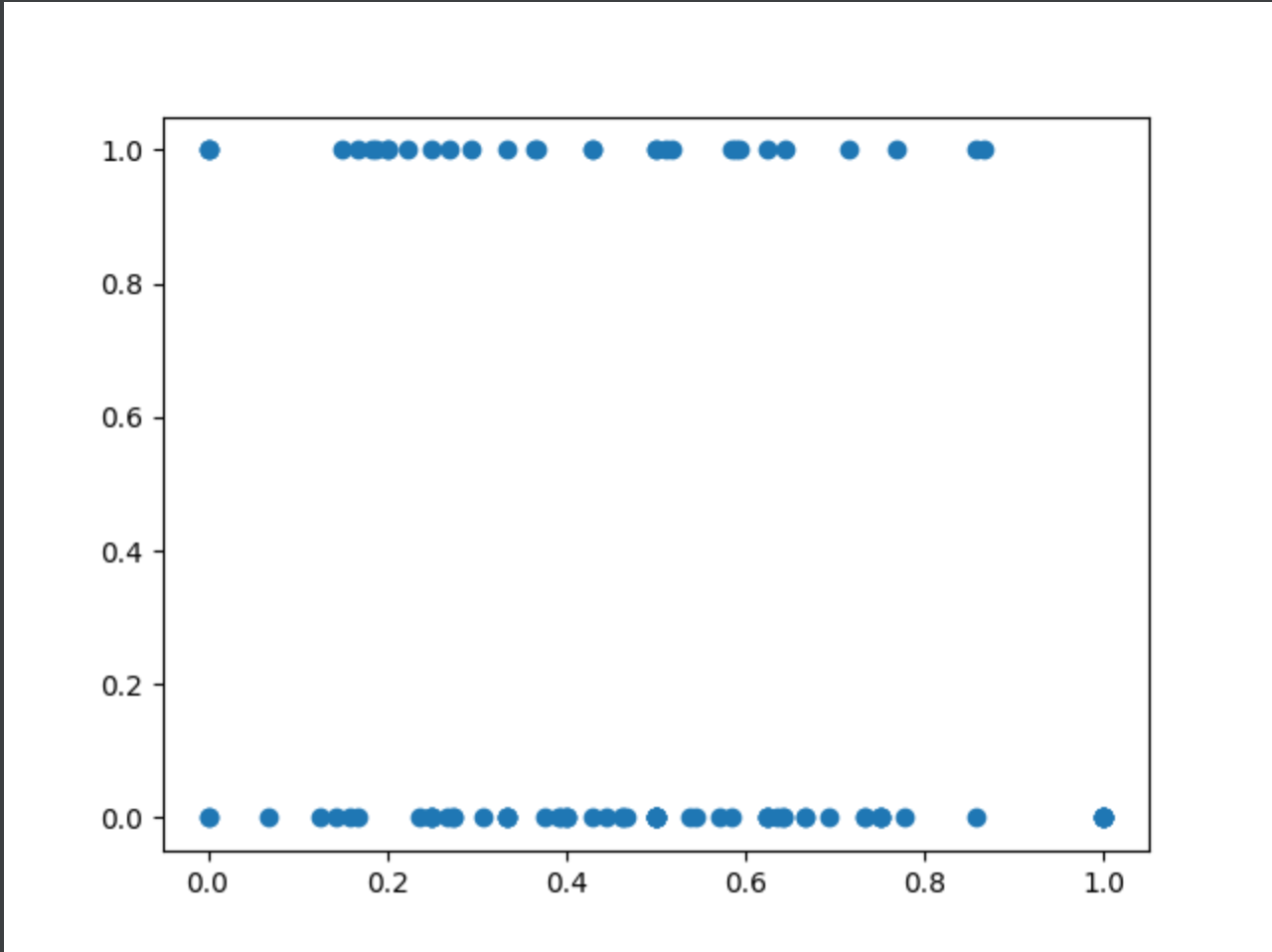


Рисунок 4.14 – Графік залежності жанру фентезі від проценту родинних стосунків

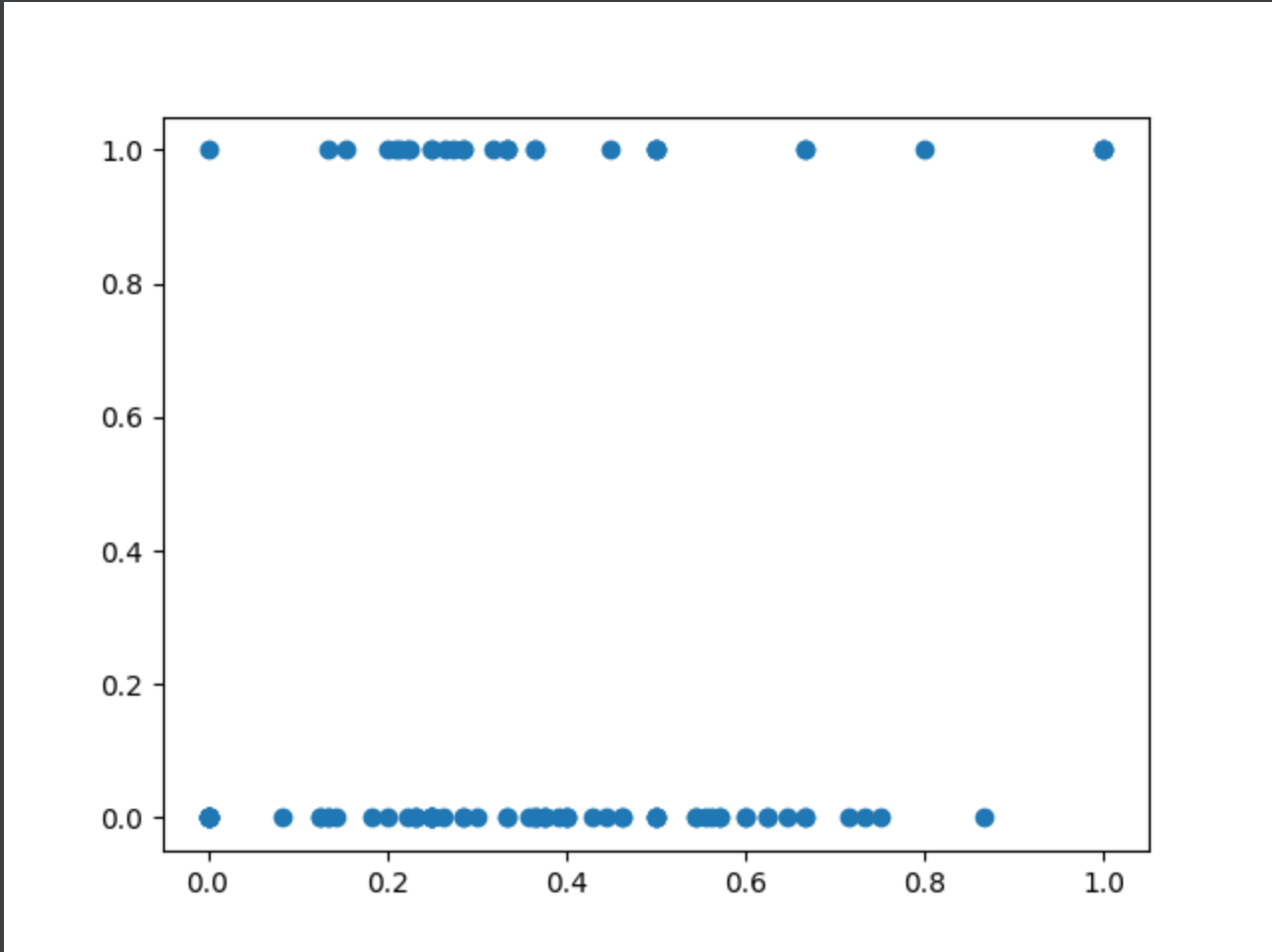


Рисунок 4.15 – Графік залежності жанру фентезі від проценту ворожих стосунків

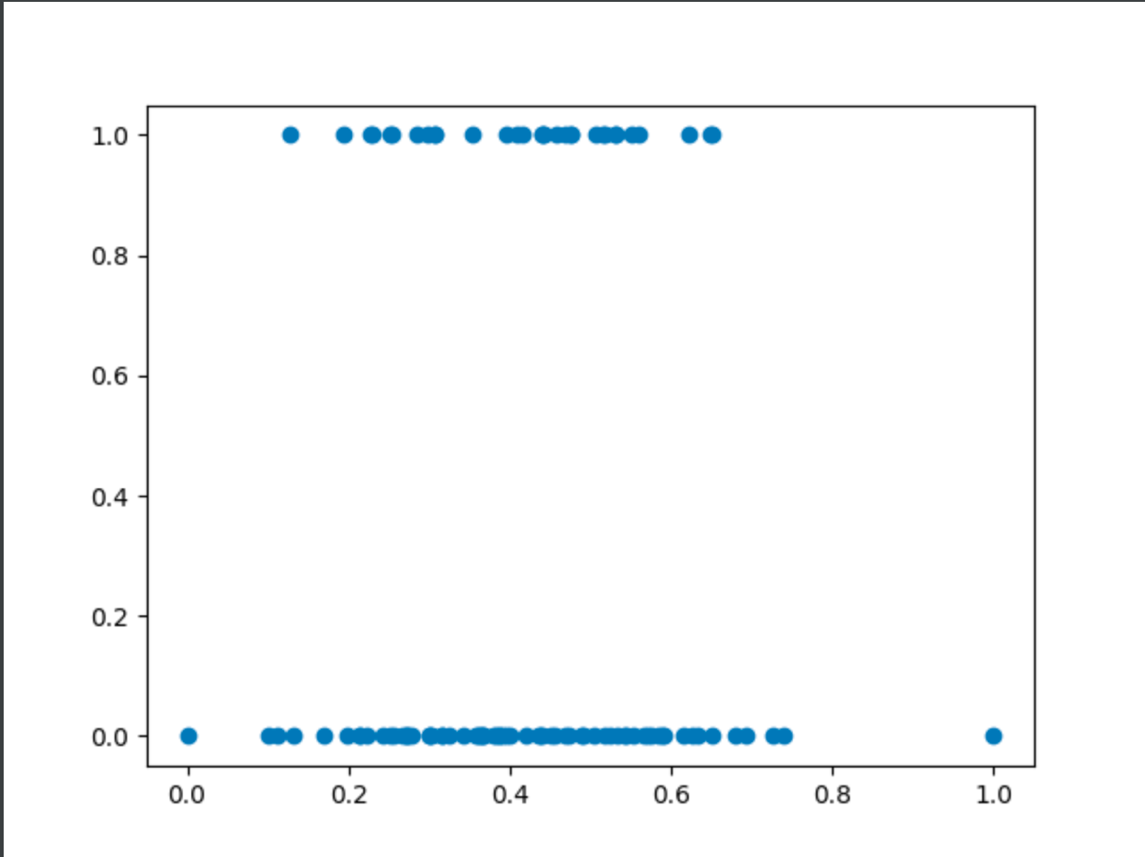


Рисунок 4.16 – Графік залежності жанру фентезі від кількості локацій

Таблиця аналізу факторів

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Назва | Середнє | Сер. при св. | Сер. при тем. | Чи значущий |
| Рік | 0.65 | 0.68 | 0.59 | 1 |
| К-сть сторінок | 0.28 | 0.23 | 0.39 | 1 |
| LiveLib | 0.51 | 0.5 | 0.53 | 0 |
| LitRes | 0.74 | 0.71 | 0.79 | 1 |
| Реалістичність | 0.43 | 0.39 | 0.51 | 1 |
| К-сть персонажів | 0.4 | 0.39 | 0.4 | 0 |
| Жінок | 0.35 | 0.35 | 0.34 | 0 |
| Чоловіків | 0.64 | 0.64 | 0.65 | 0 |
| Поворотів | 0.43 | 0.44 | 0.42 | 0 |
| Рас | 0.27 | 0.26 | 0.29 | 0 |
| Дружніх зв. | 0.14 | 0.15 | 0.12 | 0 |
| Любовних зв. | 0.14 | 0.08 | 0.28 | 1 |
| Сімейних зв. | 0.47 | 0.51 | 0.39 | 1 |
| Ворожих зв. | 0.36 | 0.34 | 0.41 | 1 |
| Локацій | 0.41 | 0.41 | 0.41 | 0 |

Таким чином для визначення жанру фентезі нам потрібно оперувати лише такими параметрами: рік, кількість сторінок, рейтинг LitRes, реалістичність, кількість любовних зв’язків, кількість сімейних зв’язків та кількість ворожих зв’язків.

Далі дослідимо вибірку на мультиколінеарні фактори. Розрахуємо таблицю кореляцій за формулою Пірсона. (Формула 4.5)

* + - * 1. Формула Пірсона

Розраховані значення наявні в таблиці 4.2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | рік | сторінки | рейтинг | реаліст. | любовних | сімейних | ворожих |
| рік | 1 | 0.5 | -0.09 | -0.05 | 0.08 | -0.05 | 0.003 |
| сторінки | 0.5 | 1 | 0.1 | 0.12 | 0.3 | -0.07 | 0.006 |
| рейтинг | -0.09 | 0.1 | 1 | 0.0008 | 0.18 | 0.02 | -0.02 |
| реаліст. | -0.05 | 0.12 | 0.0008 | 1 | 0.09 | -0.09 | 0.11 |
| любовних | 0.08 | 0.3 | 0.18 | 0.09 | 1 | -0.33 | -0.02 |
| сімейних | -0.05 | -0.07 | 0.02 | -0.09 | -0.33 | 1 | -0.81 |
| ворожих | 0.003 | 0.006 | -0.02 | 0.11 | -0.02 | -0.81 | 1 |

Таким чином легко побачити, що кількість сімейних та кількість ворожих зв’язків є мультиколінеарними, а отже не будемо враховувати один з них – кількість ворожих.

Оскільки кількість світлих і темних фентезі у вибірці нерівна, то нам потрібно її доповнити. Скористаємося алгоритмом SMOTE для доповнення вибірки. Цей алгоритм на основі вже відомих значень, з урахуванням випадкового чинника генерує нові значення для вибірки. В результаті отримаємо 138 значень, 69 з яких – світле фентезі та 69 – темне фентезі.

Код аналізу даних представлено у додатку А, файл LogisticRegresion.py.

## Логістична регресія

Приступимо до регресійного аналізу. Параметри логістичної регресії знаходяться за формулою, представленою у 2му пункті. Код представлено у додатку А, файл LogisticRegresion.py. В результаті для функції z отримуємо наступні параметри. (Формула 4.6)

* + - * 1. Отримана функція

## Дерево рішень

Код для побудови дерева рішень представлено у додатку А, файли DecisionTree.py, DecisionCell.py. Отримане дерево рішень зображене на рисунку 4.17. У хорошій якості доступно за посиланням: https://drive.google.com/file/d/1JXbIuxBhItHYHYBsJJY\_tlFypeAh-JBU/view?usp=sharing

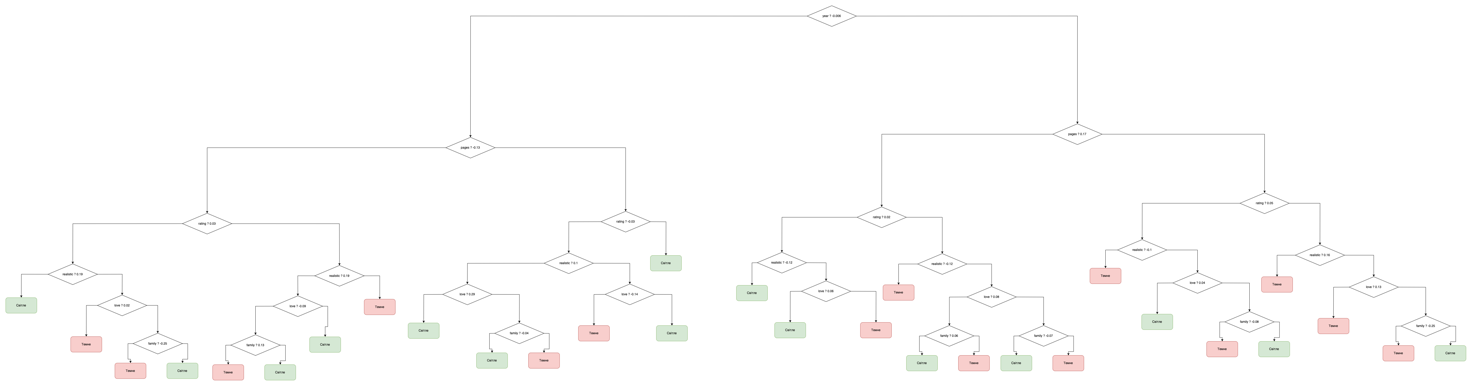


Рисунок 4.17 - Загальний план дерева рішень

## Тестування методів

Для тестування алгоритмів було відібрано 25 книг, що не брали участь у навчанні. На рисунку 4.18 та 4.19 зображено графік результатів рівняння логістичної регресії.

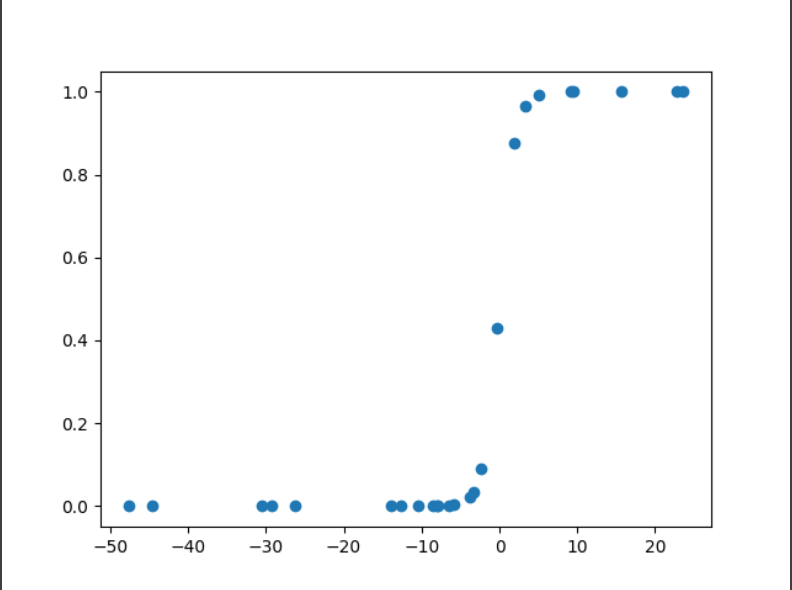


Рисунок 4.18 – Графік залежності прогнозованого y від z

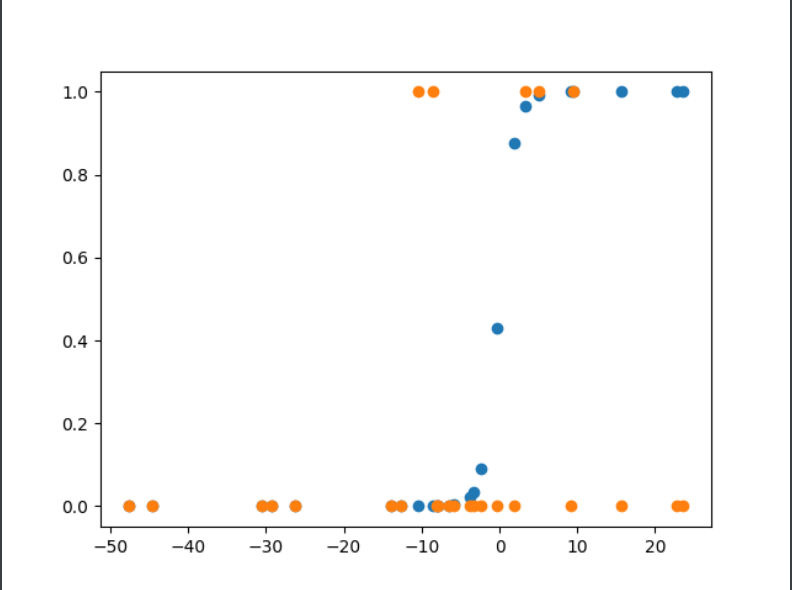


Рисунок 4.19 - Графік залежності прогнозованого y від z, з накладанням реального y

Проведемо аналіз залишків для логістичної регресії. Гістограма залишків зображена на рисунку 4.20

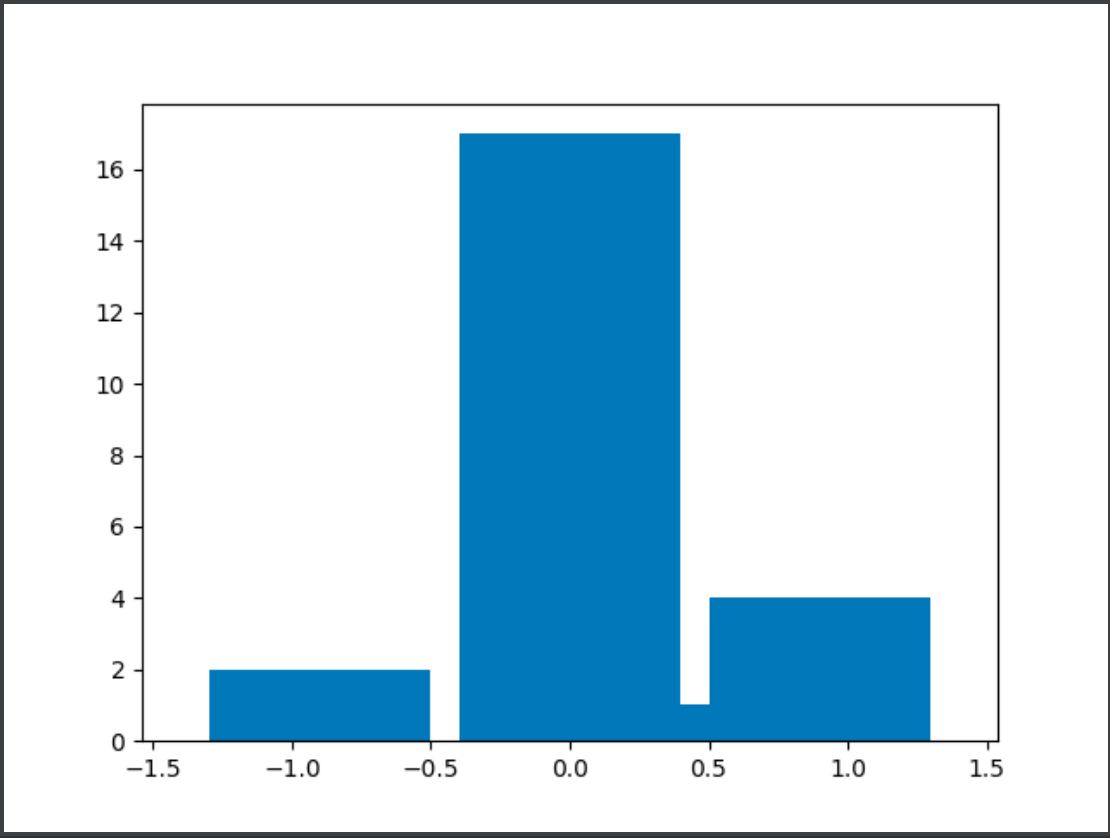


Рисунок 4.20 – Діаграма залишків

Як видно з діаграми – розподіл залишків – нормальний.

Знайдемо коефіцієнт детермінації та точність передбачення. Застосуємо формули 4.7 – 4.9

* + - * 1. Формула знаходження коефіцієнту детермінації
        2. Формула знаходження RSS
        3. Формула знаходження TSS

Результати запишемо до таблиці 4.2

Результати аналізу результатів

|  |  |  |
| --- | --- | --- |
|  | Логістична регресія | Дерево рішень |
| RSS | 4 | 4 |
| TSS | 12 | 6.96 |
| R2 | 0.66 | 0.42 |
| Точність  (к-сть правильних/  к-сть усіх) | 0.72 | 0.52 |

## Порівняння методів

Таким чином, протестувавши обидва методи, можемо зазначити, що метод логістичної регресії є точнішим, має більший коефіцієнт детермінації, а отже є кращим. Але, варто зазначити, що метод дерева рішень навчається набагато швидше, отже також має право на існування.

Висновки

В результаті виконання курсової роботи було оброблено та проаналізовано короткі описи книг, завантажено дані до сховища даних. Для реалізації даних задач було використано MySQL та мову Python3.

На основі проведеного аналізу предметної області інтелектуального аналізу даних для прогнозування піджанру фентезі книг було реалізовано 2 методи інтелектуального аналізу: логістична регресія за допомогою градієнтного спуску та дерево прийняття рішень. Перед реалізацію методів з вибірки було відсортовано незначущі та мультикореляційні параметри.

Було протестовано обидва методи. В результаті тестування виявилося, що метод логістичного аналізу працює краще ніж метод дерева прийняття рішень.

Отже, поставлені задачі були виконані, а також планується розширення функціоналу, за рахунок збільшення кількості даних та кількості порівнюваних алгоритмів.

Перелік посилань

1. Шахи [Електронний ресурс] Доступ за посиланням:

<https://uk.wikipedia.org/wiki/Шахи>

1. Правила гри в шахи ФІДЕ [Електронний ресурс] Доступ за посиланням:

<http://ukrchess.org.ua/arbiter/laws_07_2017.pdf>

Додаток А Тексти програмного коду

*Тексти програмного коду програмного забезпечення Прогнозування успішності фентезі-літератури*

(Найменування програми)

*CD-RW*

(Вид носія даних)

*57*

(Обсяг програми)

Студента групи ІП-01 2 курсу

Храмченко А.С.