Санкт-Петербургский государственный университет

КОРОТКОВА Анна Андреевна

Выпускная квалификационная работа

Оценка влияния уровня читабельности текста на популярность отзывов в сервисе «Кинопоиск»

Уровень образования: магистратура
Направление 45.04.01 «Филология»

Основная образовательная программа ВМ.5805. «Компьютерная и прикладная лингвистика»
Профиль «Компьютерная и прикладная лингвистика»

Научный руководитель:
Доцент, Кафедра математической лингвистики,
Митренина Ольга Владимировна
Рецензент:
Доцент, ФГАОУ ВО «Казанский (Приволжский) федеральный университет»,
Гурьянов Игорь Олегович

Санкт-Петербург
2021

Оглавление

Введение

Тексты даже по одной тематике могут быть очень разными. Какие-то читаются легко, какие-то требуют большого напряжения для чтения и понимания. С 30-х годов XX века исследователи пытались определить, по каким параметрам необходимо оценивать сложность восприятия того или иного текста читателем.

Было разработано более 200 формул для определения индекса удобочитаемости текста. Основная часть формул работает со статистическими и синтаксическими метриками текста (длина слова, количество слогов, средние показатели), однако некоторые формулы используют иные метрики, например, количество «трудных» слов, а также индекс абстрактности лексики.

Данная работа нацелена на определение степени влияния уровня удобочитаемости текстов рецензий на кинофильмы на сервисе «Кинопоиск». В работе проверяется, связано ли значение индекса удобочитаемости по пяти формулам (указанных ниже) с уровнем популярности отзывов. Наличие или отсутствие такой связи позволит оценить, насколько релевантно использование формул для определения индекса удобочитаемости текстов в данном дискурсе в целом.

Цель работы: определить, связан ли уровень удобочитаемости по существующим метрикам с популярностью отзывов на сервисе «Кинопоиск».

Для достижения поставленной цели были поставлены следующие **задачи**:

– изучить существующие метрики для определения индекса читабельности текста;

– собрать корпус рецензий на фильмы в сервисе «Кинопоиск»;

– провести статистический анализ полученных данных;

– провести корреляционный анализ читабельности и популярности собранных отзывов;

– проанализировать полученные результаты.

Материалом для исследования стали более трех тысяч текстов рецензий, написанных пользователями к фильмам на русском языке в сервисе «Кинопоиск».

Актуальность работы заключается в необходимости определения релевантности использования метрики удобочитаемости текстов в контексте ее влияния на популярность рецензий на кинофильмы в сервисе «Кинопоиск».

Новизну работы обуславливает отсутствие исследований, сконцентрированных на выявлении зависимости между популярностью пользовательских отзывов и их индексом удобочитаемости.

Структура работы: работа состоит из Введения, двух глав, Заключения, списка литературы и двух приложений. В первой главе рассматриваются теоретические предпосылки исследования. Вторая глава посвящена описанию проведенного эксперимента и анализу полученных в ходе работы результатов.

Глава 1. Теория и методология исследования роли читабельности в оценке популярности отзывов

1.1 Определение понятия «читабельность»

Одной из задач математической лингвистики является определение параметров, позволяющих охарактеризовать тексты с помощью неких индексов, а также создание формул для расчета этих индексов. В результате данного процесса возник такой параметр текста, как удобочитаемость (читабельность). Данное понятие возможно охарактеризовать с двух разных точек зрения. Во-первых, читабельность текста представляет собой характеристику содержания текста [21]. Во-вторых, удобочитаемость может трактоваться с точки зрения определенных типографических параметров [11]. В данной работе мы будем рассматривать удобочитаемость с первой, лингвистической точки зрения.

Единого взгляда на понятие читабельности текста не существует до сих пор, поскольку лингвисты определяют его с различных точек зрения. Одними из первых данное понятие употребили лингвисты М. Вогель и У. Уошберн в конце 1930-х годов. Они предположили, что удобочитаемость представляет количественную оценку объективных характеристик текста. Индекс читабельности по формуле, разработанной М. Вогелем и У. Уошберном, рассчитывался исходя из уравнения, построенного на основе проведенного тестирования, и устанавливал зависимость между параметрами исследуемого текста и уровнем его понимания [27].

Позднее, в 60-х годах XX столетия Дж. Клэр дал другое определение термину «удобочитаемость». Клэр определил читабельность как легкость понимания и восприятия текста, связанную со стилем письма [22]. Это определение фокусируется на стиле письма как параметре, не связанном с такими вопросами, как содержание, согласованность и организация печатного текста. Подобной точки зрения придерживалась и Г. Харгис, в 1998 году установившая, что читабельность текстов представляет собой легкость чтения слов и предложений и является характеристикой понятности [17].

В 1969 году Г. Маклафлин, создатель одной из наиболее популярных формул определения индекса удобочитаемости текстов под названием «SMOG readability formula», определил читаемость как ту степень, в которой данный класс людей находит данный материал для чтения убедительным и понятным. Это определение подчеркивает взаимосвязь между текстом и классом читателей, обладающих определенным набором характеристик, таких как умение читать, накопленный багаж знаний и мотивация [24].

По нашему мнению, наиболее полное определение удобочитаемости текстов предлагают Э. Дейл и Дж. Челл. Они считают, что читабельность представляет собой общую сумму тех элементов, которые оказывают влияние на успешность восприятия данного текста данными читателями. Успешность восприятия, согласно Э. Дейлу и Дж. Челлу, означает уровень того, насколько читатель понимает текст; скорость, с которой представляется возможным его читать, а также уровень заинтересованности читателей в тексте [16]. На наш взгляд именно данный подход к определению читабельности текстов является наиболее полным, так как он учитывает не только числовые характеристики текста, такие как длина предложения и слов, но и уровень понимания текста читателем, оцениваемый в количестве уже известных («простых») слов, а также заинтересованность в прочтении текстов, которая была оценена рядом тестов и скорость чтения, которая аналогично дает исследователям возможность оценить, насколько понятным текст представляется для читателя.

В последней четверти XX века многими исследователями было представлено большое число различных формул для определения индекса удобочитаемости текстов. Разработка формул, а также различных взглядов на читабельность дала мощный толчок развитию изучения читабельности текстов как одного из важнейших параметров текста, используемого в различных областях. Формулы и подходы представлены в следующем разделе.

Существующие формулы для определения индекса удобочитаемости текстов нашли широкое применение в различных областях. В частности, они используются при составлении учебных материалов: рабочих тетрадей, справочных пособий,

учебников. Проверка читабельности с целью ее дальнейшей корректировки необходима для того, чтобы учащиеся без труда понимали и усваивали представленную информацию. Также практика по использованию формул определения удобочитаемости текстов широко применяется в разработке текстов рекламных кампаний. Именно от уровня сложности восприятия представленного текста зависит успех в продвижении продукта [13].

1.2 Обзор формул, использующихся для оценки удобочитаемости

Для числового отражения уровня сложности текстов в 20-х годах XX столетия был разработан способ по использованию словарного содержания текста и длины предложений [18]. Через 60 лет, в 1980-х, было выпущено более тысячи исследовательских работ по читабельности и предложено около 200 математических формул. Среди них представлены: индекс туманности Ганнинга (Gunning fog index), уровень Колеман-Лиау (Coleman Liau idex), индекс Флеша-Кинкейда (Flesch Kincaid grade level), удобочитаемость по Флешу (Flesch reading ease), Индекс Флеша-Кинкейда, Формула FORCAST, Формула SMOG, график Фрая [2].

В своей работе под названием «Разработка метода автоматизированной оценки сложности учебных текстов для высшей школы» М. М. Невдах выявил 49 текстовых параметров, используемых для оценки сложности текстов [10]. Данные параметры составляют:

1. Длина текста:

a. В абзацах

b. В словах

c. В буквах

2. Средняя длина абзаца:

a. Во фразах

b. В словах

c. В буквах

d. В печатных знаках

3. Средняя длина предложения:

a. Во фразах

b. В Словах

c. В слогах

d. В буквах

e. В печатных знаках

4. Средняя длина самостоятельного предложения:

a. Во фразах

b. В ловах

c. В слогах

d. В буквах

e. В печатных знаках

5. Средняя длина фразы:

a. В словах

b. В слогах

c. В буквах

d. В печатных знаках

6. Средняя длина слов:

a. В слогах

b. В буквах

c. В печатных знаках

d. По Деверу

7. Процент слов длиной в:

a. 5 букв и больше

b. 6 букв и больше

c. 7 букв и больше

d. 8 букв и больше

e. 9 букв и больше

f. 10 букв и больше

g. 11 букв и больше

h. 12 букв и больше

i. 13 букв и больше

8. Процент слов в:

a. 3 слога и больше

b. 4 слога и больше

c. 5 слогов и больше

d. 6 слогов и больше

9. Процент неповторяющихся слов

10. Средняя частота повтора слов

11. Процент существительных:

a. Повторяющихся

b. Неповторяющихся

c. Конкретных

d. Абстрактных

12. Процент:

a. Прилагательных

b. Глаголов

c. Сложных предложений

d. Простых предложений

e. Придаточных предложений среди фраз (49 признаков)

Данные характеристики лежат в основе большинства существующих формул для определения индекса читабельности текстов [10].

Рассмотрим наиболее популярные в современной лингвистике формулы определения удобочитаемости, разработанные в 1970-х годах. Авторами данных распространенных формул являются Я. А. Микк и М. С. Мацковский.

Первая формула, формула Я. А. Микка, была создана в 70-х годах XX века. Изначально она была предназначена для определения индекса удобочитаемости

текстов на эстонском языке. Формула Я. А. Микка имела следующий вид F=0,131*X1+9,84*X2-4,5, где F – индекс удобочитаемости исследуемого текста, X1 – средняя длина предложений, выраженная в печатных знаках, а X2 – средняя абстрактность повторяющихся существительных. Абстрактность имен существительных определялась по трехбалльной шкале. Согласно данной дифференциации, существительные подразделялись на одушевленные и неодушевленные, воспринимаемые органами чувств; явления, также воспринимаемые органами чувств; конструкции мысли, не воспринимаемые органами чувств. Также данный показатель мог быть вычислен математически. Для этого использовался подсчет использования слов, содержащих морфемы абстрактности. Соответственно, X1 и X2 являются лингвистическими параметрами данной формулы [7].

Примерно в одно время с Я. А. Микком свою формулу удобочитаемости текстов разработал М. С. Мацковский. Формула использовалась для текстов на русском языке и имела следующий вид: F=0,62*X1+0,123*X2+0,051, где F – индекс читабельности анализируемого текста, X1 – средняя длина предложения, выраженная в словах, X2 – процент слов, состоящих из трех и более слогов. Таким образом, лингвистическими параметрами данной формулы являются показатели длины предложения и процента длинных слов [6].

В западной лингвистической парадигме распространена формула Дейла-Челла. Она определяет удобочитаемость, в отличие от большинства других формул, учитывая семантику использованных в англоязычном тексте слов. В данной формуле используется показатель количества «трудных» слов. Под этим термином подразумеваются слова, не присутствующие в разработанном авторами формулы списке «общих» слов, которые должны быть знакомы большинству учащихся 4-х классов по системе обучения K-12. Изначально в данном списке значилось 763 слова, однако позже, в 1995 году, Э. Дейл и Дж. Челл, переосмыслив свой труд, расширили данный список до 3000 слов и дали новое определение удобочитаемости текстов. Согласно новому взгляду ученых, читабельность представляет собой сумму всех элементов в исследуемом отрывке текста, которые

оказывают влияние на успех восприятия читателями данного отрывка. Под успехом лингвисты подразумевали уровень понимания текста читателем, способность прочесть отрывок с оптимальной скоростью, а также возможность определения содержания текста, как интересного [15, с. 80].

Итоговая формула Дейла-Челла, которая помимо значения слов учитывает также среднюю длину предложения, получила следующий вид: F1=0,1579*X1=0,0496*X2, где F1 – первичный балл, X1 – процент «трудных» слов, X2 – средняя длина предложения, выраженная в словах. Первичный балл используется для определения текстов, оптимальным уровнем обученности читателей которого являются ученики до третьего класса включительно. Для определения удобочитаемости текстов для читателей с высоким уровнем обученности, в текстах для которых встречается более 5% «трудных» слов, используется формула высчитывания уточненного балла, которая имеет вид: F2=F1+3,6365, где F2 – уточненный балл [15].

В представленной ниже Таблице 1 отображено соответствие между уровнем обученности в классах по системе K-12 и показателя уточненного балла.

Таблица 1 – Соответствие уточненного балла уровню обученности

Уточненный балл	Уровень обученности (класс)
4.9 и ниже	4 и ниже
5.0 до 5.9	5 - 6
6.0 до 6.9	7 - 8
7.0 до 7.9	9 - 10
8.0 до 8.9	11 - 12
9.0 до 9.9	13 - 15 (уровень колледжа)
10 и выше	16 и выше (уровень выпускника колледжа)

Широкое распространение получила формула определения удобочитаемости текстов на английском языке, разработанная Р. Ганнингом в 1952 году. Формула получила название «Индекс туманности Ганнинга». В основе данной формулы

лежит подсчет средней длины предложения и процента слов, длина которых превышает 3 слога. Результат расчета по данной формуле позволяет определить для читателей какого возраста данный текст будет понятен. Для текстов, которые будут просты в чтении для большинства людей, полученное значение индекса удобочитаемости не должно превышать 12 [25].

Чем выше полученный индекс, тем сложнее текст для прочтения. Для расчета индекса удобочитаемости по формуле Ганнинга берется два отрывка текста, в каждом из которых содержится около 100 слов. Также определяются основные параметры формулы, которые представлены в виде суммарного количества слов в тексте k, общего количества слов в данном тексте s, средней длины предложения w, количества слов длиннее трех слогов l. Сама формула имеет вид: $F = \sum_{i=l}^{s} \frac{w_i}{s} + 0.4 \sum_{i=l}^{s} \frac{l_i}{w_i} = w + 0.4l$, где F – индекс удобочитаемости, индекс i определяет очередность индексируемой единицы [25].

С помощью графика Фрая представляется возможным определить сложность текста по следующим параметрам: суммарное количество слов в тексте; общее количество предложений в тексте; количество слогов в тексте; средняя длина предложений; средняя длина слова, выраженная в количестве содержащихся в нем слогов [23].

Для определения индекса удобочитаемости по Фраю берется текст длиной не менее 100 слов. График Фрая представляет собой диаграмму зависимости количества слогов и предложений в данном тексте. Кривая диаграммы представляет собой показатели нормального текста. Чтобы получить значение индекса удобочитаемости по Фраю необходимо найти точки, соответствующие значениям средней длины предложения и средней длины слов в слогах. График удобочитаемости по Фраю представлен на Рисунке 1 [23].

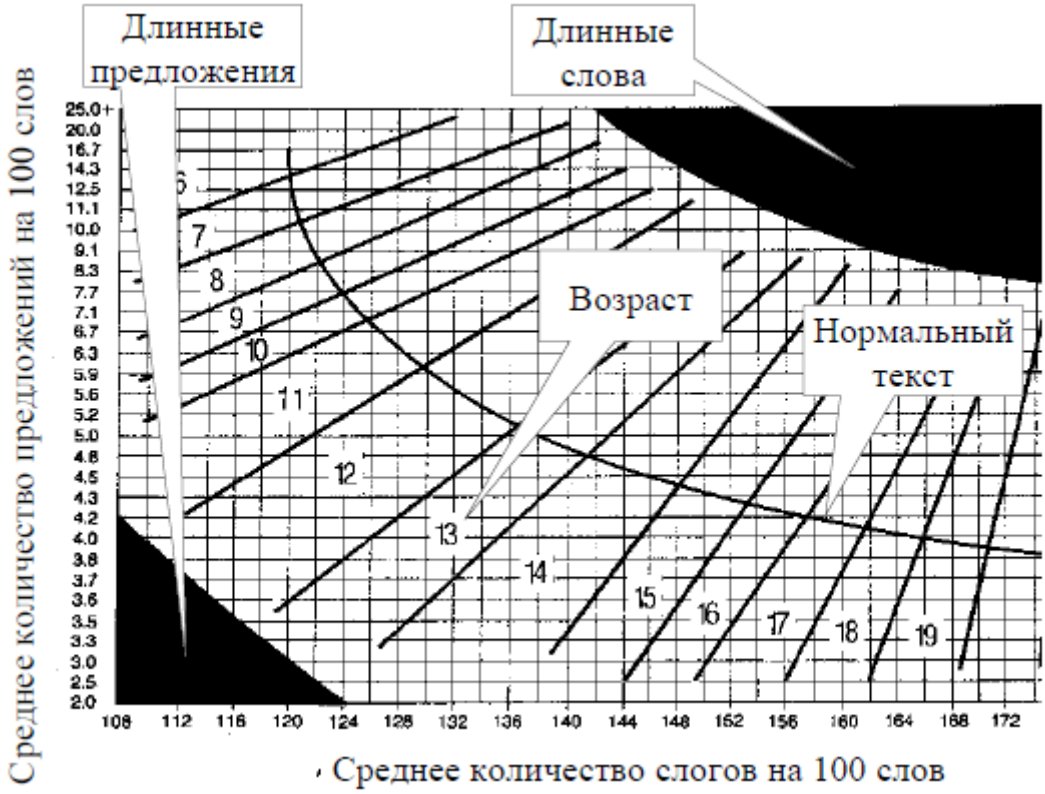


Рисунок 1 – График удобочитаемости Фрая

Индекс Колеман-Лиау, в отличие от большинства других формул определения читабельности текстов, базируется не на подсчете количества и среднего количества слогов в тексте, предложении и слове, а на подсчете количества символов. Так, параметрами данной формулы являются: суммарное количество символов в данном тексте x; общее количество слов k; число предложений s; средняя длина предложений w; средняя длина слов P, выраженная в символах [32].

Формула Колеман-Лиау имеет следующий вид: $F = 5{,}89 * \frac{x}{k} + 30 * \frac{s}{k} - 15{,}8$. Где F служит обозначением индекса удобочитаемости текстов по данной формуле [21].

Формула расчета удобочитаемости текстов на английском языке, получившая название «SMOG», была разработана лингвистом Маклаулином. Данная формула предназначена для прогнозирования абсолютного, 100% понимания текста. Длина текста s в этой формуле, в отличие от большинства других, ограничивается строго 30 предложениями. Другими параметрами по

формуле SMOG являются суммарное количество слов в данном тексте k; среднее количество слов, длина которых 3 и более слогов l; общее количество слов длиной 3 и более слогов L. Формула Маклаулина имеет следующий вид: $F = \sqrt{L/_S \sqrt{l}}+3$, где F представляет собой значение индекса читабельности по данной формуле [24].

Формула под названием FORCAST была разработана в целях определения уровня понятности текстов технического характера для вооруженных сил США. По этой причине данная формула не применяется для определения сложности учебных текстов. Ввиду характера текстов они могут не содержать полные предложения. По формуле FORCAST значимыми являются следующие параметры текста:

– Суммарное количество слов k

– Количество слов, состоящих из одного слога b

Формула имеет следующий вид: F=20-k*0,667*b, где F является значением индекса читабельности данного текста, рассчитанным по формуле FORCAST [26].

Несмотря на большое количество разработанных формул по определению удобочитаемости текстов, наиболее распространенной является формула читабельности по Флешу. Она была разработана в 40-х годах XX века Рудольфом Флешем. Шкала, согласно которой определялся уровень сложности исследуемого текста, состояла из 100 пунктов. Согласно данной дифференциации, чем ниже показатель, тем меньше людей могли без труда воспринять и усвоить представленный текст [19].

Последователем Флеша стал лингвист Кинкейд, который в 1975 году разработал формулу читабельности текстов, которая была основана на индексе Флеша, однако изменилась шкала определения уровня сложности. Так, согласно формуле Флеша-Кинкейда, итоговое значение индекса равно классу обучения по системе K-12. Лингвистическими параметрами данной формулы являются средняя длина предложения, выраженная в словах, и средняя длина слов, выраженная в слогах. Мы предполагаем, что изменение показателей данного параметра может повлиять на итоговый индекс удобочитаемости анализируемого текста [20].

Как уже было ранее отмечено, каждая формула определения индекса удобочитаемости текстов имеет определенный набор параметров, по которым вычисляется уровень сложности текстов. Для анализа текстов в данной работе мы будем использовать формулу Флеша-Кинкейда. Лингвистическими параметрами данной формулы определения читабельности текстов являются показатели:

1. Средней длины предложений в анализируемом тексте, выраженной в словах;

2. Средней длины слов в данном тексте, выраженной в слогах.

1.3 О сервисе «Кинопоиск»

Материалом для проводимого исследования стали рецензии на фильмы, которые оставляют пользователи в сервисе «Кинопоиск».

«Кинопоиск» представляет собой один из самых масштабных русскоязычный интернет-сервисов о кино. С 2018 года на сайте стал доступен онлайн-кинотеатр, где представлены тысячи фильмов, сериалов, мультфильмов. Более того, пользователи имеют доступ к премьерному и эксклюзивному контенту [29]. Также сайт предоставляет возможность не только узнать рейтинг того или иного произведения, но и самостоятельно повлиять на него, путем оставления оценок и написания рецензий. Всего на момент написания данной работы пользователи оставили 754282 рецензии [28].

С помощью сервиса пользователи имеют возможность получить информацию о кинокартинах, телесериалах, в том числе кадры, трейлеры, постеры, а также данные о персонах, связанных с кино- и телепроизводством: актерах, режиссерах, продюсерах, сценаристах, операторах, композиторах, монтажерах и художниках [29].

Регистрация на сайте дает пользователям дополнительные возможности. По сути, «Кинопоиск» является своего рода социальной сетью.

На странице зарегистрированного пользователя содержится информация о его деятельности на сайте:

1) Информация о пользователе:

– Никнейм

– Имя

– Дата рождения

– Пол

– Дата регистрации

2) Друзья. На самом деле под «друзьями» на «Кинопоиске» понимается своеобразная система подписок – поле «Друзья» показывает пользователей, за обновлениями которых следит пользователь, а поле «Добавили в друзья» позволяет увидеть пользователей, подписавшихся на обновления данной страницы.

Данный параметр будет в дальнейшем учитываться в работе, так как количество друзей увеличивает количество пользователей, которые увидят новые рецензии автора, что, в свою очередь, может повлиять на итоговую популярность отзыва.

3) Список понравившихся фильмов

4) Список понравившихся сериалов

5) Список ожидаемых фильмов

6) Список просмотренных фильмов

7) Список просмотренных сериалов

8) Список любимых актеров

9) Поставленные оценки

10) Написанные комментарии

11) Написанные рецензии

Пример страницы пользователя представлен на Рисунке 2.

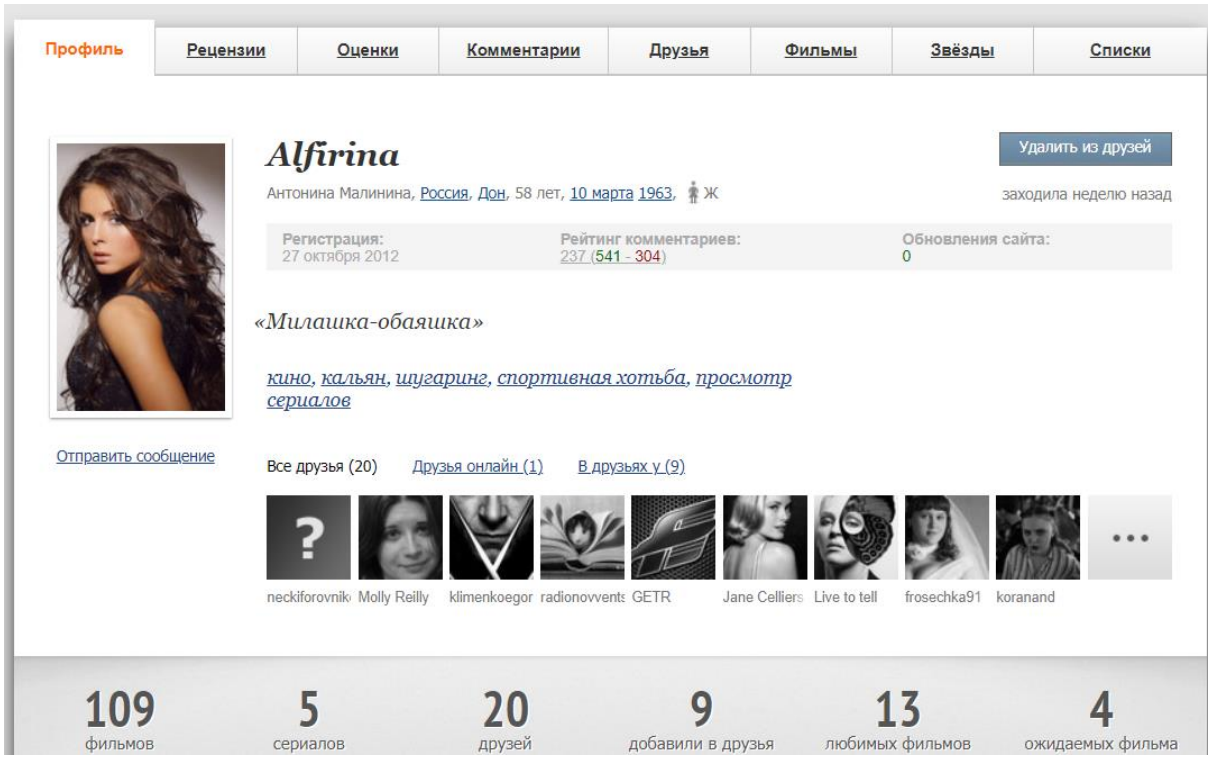


Рисунок 2

1.4 Рецензии

Рецензия – один из самых популярных жанров в литературно-художественной критике; в общем смысле это отчет, который дает чье-то мнение о качестве книги, производительности, продукте и т.д. [5; 32]. А. А. Тертычный также определяет данное понятие как «жанр, основу которого составляет отзыв (прежде всего – критический) о произведении художественной литературы, искусства, науки, журналистики и т.п.» [12]. Соответственно, исходя из указанных определений, мы можем прийти к выводу, что рецензии выполняют информационную и оценочную функции. В. А. Фомина представляет более расширенный список функций, выполняемых рецензиями. Так, исследовательница добавляет к вышеуказанному списку информационную, рекламную и развлекательную функции, соответственно, жанр рецензии является многофункциональным [14].

В «Кинопоиске» рецензии представляют собой мнение о фильме или сериале, которое выразил автор в поле «Рецензии» на странице профиля той или иной киноленты. Следует различать понятия «рецензия» и «комментарий», так как последние являются комментариями других пользователей на рецензию к фильму и ответом пользователя на комментарии к его рецензии.

На «Кинопоиске» представлена возможность оценки рецензий. Так, пользователи могут отмечать, понравилась им рецензия или нет. Поле для оценки рецензии показано на Рисунке 3 в левом нижнем углу – пользователь отмечает, была ли полезна рецензия в специальном поле.

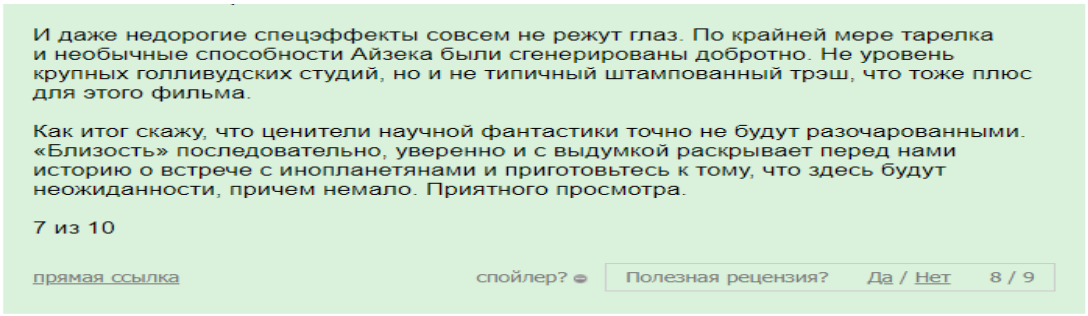


Рисунок 3

На странице автора есть поле «Рецензии в цифрах», где содержится информация об общем количестве написанных рецензий, датах написания первой и последней рецензии, среднем количестве написанных рецензий в месяц, а также о суммарном рейтинге рецензий. Пример данного поля представлен на Рисунке 4.

Рисунок 4

Суммарный рейтинг показывает, сколько положительных и отрицательных оценок на рецензии получил автор. Данный параметр будет учитываться в работе, так как именно он отражает популярность рецензий автора.

Под популярностью в работе понимается суммарное количество положительных и отрицательных оценок, так как нам важно видеть, вызвала ли та или иная рецензия автора отклик у других пользователей.

1.5 Понятие коэффициента корреляции

Корреляционный анализ занимается степенью связи между переменными.

Андрей Наследов в своей работе «IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных» дает следующее определение понятия корреляция/коэффициент корреляции: «Корреляция, или коэффициент корреляции, – это статистический показатель вероятностной связи между двумя переменными, измеренными в единицах количественной шкалы» [8]. Исследователи А. Бююль и П. Цефель в совместной работе «SPSS: искусство обработки информации» дополняют данное определение, утверждая, что «расчёты подобных двумерных критериев взаимосвязи основываются на формировании парных значений, которые образовываются из рассматриваемых зависимых выборок» [1].

Графическое представление корреляции между переменными отображается в виде так называемой диаграммы рассеяния. Пример подобной диаграммы представлен на Рисунке 5 ниже.

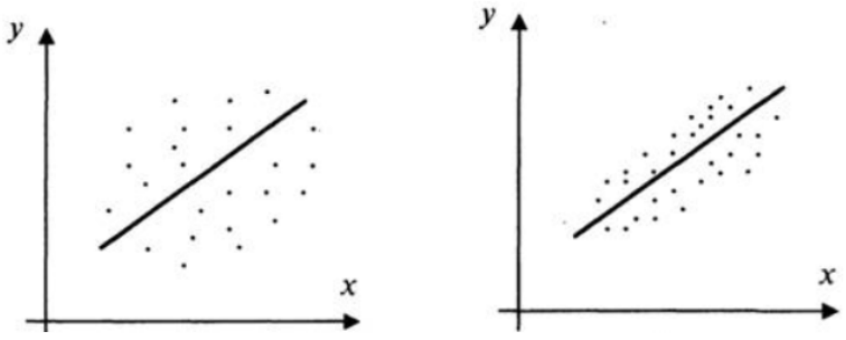


Рисунок 5

Большее скопление точек указывает на большую зависимость между исследуемыми переменными. Коэффициент корреляции указывает на «силу связи при помощи некоторого критерия зависимости» [1]. Данный коэффициент может принимать значения в промежутке от -1 до 1. Близкое к 1 полученное значение свидетельствует о высокой степени зависимости между переменными; чем ближе значение к нулю, тем слабее связь между значениями переменных. Отрицательность полученного значения говорит о наличии обратной связи между переменными.

Интерпретация значений коэффициента корреляции представлена на Рисунке 6 ниже [1].

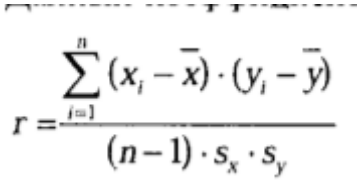
Значение	*Интерпретация*
до 0,2	Очень слабая корреляция
до 0,5	Слабая корреляция
до 0,7	Средняя корреляция
до 0,9	Высокая корреляция
свыше 0,9	Очень высокая корреляция

Рисунок 6

1.5.1 Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона представляет собой меру корреляции, которая подходит для двух метрических переменных, измеряемых на одной и той же выборке [9]. Является мерой прямолинейной связи, то есть в случае, когда точки на диаграмме рассеяния лежат на одной прямой линии, значения указанного коэффициента достигают своего максимума.

Коэффициент корреляции Пирсона вычисляется по формуле, указанной на Рисунке 7 [1].



$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y}$$

Рисунок 7

В представленной формуле x(i) и y(i) являются значениями двух переменных; x- и y- представляют собой их средние значения, а s(x) и s(y) – их стандартное отклонение; n отражает количество пар значений.

1.5.2 Коэффициенты корреляции Спирмена и Кендалла

В реальных задачах отношения между переменными часто оказываются не только прямолинейными, но и непрямолинейными, монотонными или немонотонными. В таком случае вместо коэффициента корреляции Пирсона следует использовать ранговые корреляции Спирмена или Кендалла [8].

Коэффициент корреляции Спирмена, как и коэффициент корреляции Кендалла, представляет собой меру корреляции, которая подходит для переменных, измеряемых в ранговой шкале или же переменных, распределение которых отличается от нормального [9].

Коэффициент корреляции Спирмена вычисляется по формуле, представленной на Рисунке 8 [3].

$$\rho = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2$$

Рисунок 8

В данной формуле R(i) является рангом наблюдения x(i) в ряду x; S(i) является рангом наблюдения y(i) в ряду y [3].

Коэффициент корреляции Кендалла вычисляется по формуле, представленной на Рисунке 9 [3].

$$\tau = 1 - \frac{4}{n(n-1)} R$$

Рисунок 9

В данной формуле под R скрывается значение, представленное на Рисунке 10, обозначающее количество инверсий, которое образуют величины Y(i), располагающимися в порядке возрастания x(i), соответствующих им [3].

$$R = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[\left[x_i < x_j \right] \neq \left[y_i < y_j \right] \right]$$

Рисунок 10

В следующей главе коэффициенты корреляции будут использоваться для выявления связей между различными оценками уровня читабельности текста и их популярностью.

Глава 2. Эксперимент по оценке влияния уровня читабельности на популярность отзывов в сервисе «Кинопоиск»

2.1. Общее описание эксперимента

В данной главе описывается серия экспериментов по поиску корреляции между популярностью отзывов в сервисе «Кинопоиск» и их индексами читабельности, полученных с помощью различный метрик по оценке читабельности.

Первый этап работы состоял в создании корпуса отзывов, в который вошло несколько тысяч отзывов, а также разметка по синтаксическим параметрам, индексам удобочитаемости и популярности отзыва.

Следующий этап работы заключался в вычислении коэффициентов корреляции между показателями индексов удобочитаемости текстов отзывов и их популярностью по формулам Пирсона, Спирмена и Кендалла.

Затем было проведено выделение признаков (feature selection), позволившее определить степень влияния метрик на показатель популярности.

Наконец, завершающим этапом анализа стал поиск аномалий в показателях популярности и соответствующих значениях индексов удобочитаемости текстов рецензий.

2.2. Сбор корпуса отзывов и его метаразметка

В данной работе корпус текстов состоит из текстов рецензий, написанных пользователями в сервисе «Кинопоиск». Сбор информации производился при помощи парсинга страницы сайта, где размещены все рецензии пользователей [28]. Помимо самого отзыва собирались его метаданные: никнейм пользователя на сайте, дата написания отзыва, id пользователя-автора отзыва, числовые показатели.

Программа по сбору корпуса работает на основе библиотек beautiful soup и Pandas. Изначально программой была собрана общая информация по странице сайта, важные пункты которой в дальнейшем были занесены в итоговую таблицу, которая будет описана ниже. Пример части программы представлен на Рисунке 11. Полный листинг программы по сбору корпуса представлен в Приложении 1.

```
soup = BeautifulSoup (r.content, features='html.parser')
users = soup.find_all('p', {'class':'profile_name'})
users_nick = []
for i in users:
    users_nick.append(i.text)
reviews = soup.find_all('div', {'class':'brand_words'})
date = soup.find_all('span', {'class':'date'})
dates = []
for i in date:
  dates.append(i.text)
id_users = []
# достаём ссылки на страницы авторов
for i in users:
    user = i.a['href']
    id_users.append(user)
```

Рисунок 11

Для каждого автора отдельно были собраны его показатели популярности: количество друзей и подписчиков.

Из-за редизайна и ограничений, появлявшихся в системе «Кинопоиск» в 2019-2021 годах, систему сбора отзывов пришлось неоднократно перестраивать.

2.3. Разметка сложности собранных отзывов

Для определения сложности текста был использован API, разработанный Иваном Бегтиным [31]. Метод его использования представлен на Рисунке 12.

```
for r in reviews:
    rew = r.text
    r_text.append(rew)
    response = requests.post("http://api.plainrussian.ru/api/1.0/ru/measure/", data={"text":rew})
    readability = response.json()
```

Рисунок 12

Разработка формул, подстроенных под русский язык, велась на основе текстов различных стилей. Так, были использованы тексты для школьного внеклассного чтения, размеченные тексты для людей возраста выпускников учебных заведений и старше, тексты повышенной сложности, например, тексты законов. На основе выборки был произведен подбор коэффициентов, позволяющих использовать формулы, изначально разработанные для работы с текстами на английском языке, для оценки удобочитаемости текстов на русском языке.

Программа дает возможность получить индексы читабельности по пяти формулам определения сложности текстов:

– Flesch-Kinkaid;

– Dale-Chale readability formula;

– Coleman-Liau index;

– SMOG;

– Automated readability index.

Особенности каждой из вышеприведенных формул были рассмотрены в теоретической части данной работы.

API находится в открытом доступе [30]. Для оценки сложности того или иного текста на русском языке программе необходимо передать либо непосредственно сам текст, либо ссылку на страницу, на которой находится текст, который будет подвержен дальнейшему анализу. Говоря на языке параметров, необходимо указать url (для ссылки) или text (для текста).

Параметр url может быть передан при помощи запроса GET, пример подобного обращения представлен на Рисунке 13.

```
{
    "metrics": {
        "wsyllabes": {
            "1": 3,
            "3": 3,
            "2": 1,
            "5": 4,
            "4": 5
        },
        "c_share": 25.0,
        "chars": 177,
        "avg_slen": 8.0,
        "spaces": 20,
        "n_syllabes": 54,
        "n_words": 16,
        "letters": 127,
        "n_sentences": 2,
        "n_complex_words": 4,
        "n_simple_words": 12,
        "avg_syl": 3.375
    },
    "status": 0,
    "indexes": {
        "grade_SMOG": "1 - 3 \u043a\u0443\u0440\u0441\u044b \u0412\u0423\u0417\u0430 (\u0432\u043e\u0437\u0440\u0430\u0441\u0442 \u043f\u0440\u0438\u043c\u0435\u0440\u043d\u043e: 17-19 \u043b\u0435\u0442)",
        "grade_ari": "\u0410\u0441\u043f\u0438\u0440\u0430\u043d\u0442\u0443\u0440\u0430, \u0432\u0442\u043e\u0440\u043e\u0435 \u0432\u044b\u0441\u0448\u0435\u0435
\u043e\u0431\u0440\u0430\u0437\u043e\u0432\u0430\u043d\u0438\u0435, phD",
        "index_fk": 19.968999999999994,
        "grade_cl": "\u0410\u0441\u043f\u0438\u0440\u0430\u043d\u0442\u0443\u0440\u0430, \u0432\u0442\u043e\u0440\u043e\u0435 \u0432\u044b\u0441\u0448\u0435\u0435
\u043e\u0431\u0440\u0430\u0437\u043e\u0432\u0430\u043d\u0438\u0435, phD",
        "grade_fk": "\u0410\u0441\u043f\u0438\u0440\u0430\u043d\u0442\u0443\u0440\u0430, \u0432\u0442\u043e\u0440\u043e\u0435 \u0432\u044b\u0441\u0448\u0435\u0435
\u043e\u0431\u0440\u0430\u0437\u043e\u0432\u0430\u043d\u0438\u0435, phD",
        "index_cl": 18.95125,
        "grade_dc": "4 - 6 \u043a\u0443\u0440\u0441\u044b \u0412\u0423\u0417\u0430 (\u0432\u043e\u0437\u0440\u0430\u0441\u0442 \u043f\u0440\u0438\u043c\u0435\u0440\u043d\u043e: 20-22 \u043b\u0435\u0442)",
        "index_dc": 15.984000000000002,
        "index_ari": 20.89275,
        "index_SMOG": 12.553279569776885
    }
}
```


Рисунок 13

Пример расчета параметров простейшего текста (1 предложение) показан на Рисунке 14.


```
{
    "metrics": {
        "wsyllabes": {},
        "c_share": 0,
        "chars": 61,
        "avg_slen": 0.0,
        "spaces": 8,
        "n_syllabes": 0,
        "n_words": 0,
        "letters": 0,
        "n_sentences": 1,
        "n_complex_words": 0,
        "n_simple_words": 0,
        "avg_syl": 0
    },
    "status": 0,
    "indexes": {
        "grade_SMOG": "\u043d\u0435\u0438\u0437\u0432\u0435\u0441\u0442\u043d\u043e (0)",
        "grade_ari": "\u043d\u0435\u0438\u0437\u0432\u0435\u0441\u0442\u043d\u043e (0)",
        "index_fk": 0,
        "grade_cl": "\u043d\u0435\u0438\u0437\u0432\u0435\u0441\u0442\u043d\u043e (0)",
        "grade_fk": "\u043d\u0435\u0438\u0437\u0432\u0435\u0441\u0442\u043d\u043e (0)",
        "index_cl": 0,
        "grade_dc": "\u043d\u0435\u0438\u0437\u0432\u0435\u0441\u0442\u043d\u043e (0)",
        "index_dc": 0,
        "index_ari": 0,
        "index_SMOG": 0.05
    }
}
```


Рисунок 14

Результат данного расчета, который будет представлен пользователю, показан на Рисунке 15.

{ metrics: { wsyllabes: { 1: 94, 2: 116, 3: 140, 4: 87, 5: 139, 6: 45, 7: 18, 8: 4, 15: 1 }, c_share: 32.142857142857146, chars: 6000, avg_slen: 46, spaces: 510, n_syllabes: 2232, n_words: 644, letters: 5170, n_sentences: 14, n_complex_words: 207, n_simple_words: 437, avg_syl: 3.4658385093167703 }, status: 0, indexes: { grade_SMOG: "Аспирантура, второе высшее образование, phD", grade_ari: "Аспирантура, второе высшее образование, phD", index_fk: 33.342906832298134, grade_cl: "Аспирантура, второе высшее образование, phD", grade_fk: "Аспирантура, второе высшее образование, phD", index_cl: 23.062857142857148, grade_dc: "Аспирантура, второе высшее образование, phD", index_dc: 30.300857142857147, index_ari: 32.11796894409938, index_SMOG: 34.046178356649776 } }

Рисунок 15

В нашей работе мы использовали параметр text для передачи анализируемого текста программе. Стоит отметить, что в работе использовался POST-запрос во избежание ограничения на размер URI. Как было показано выше, для проведения данного анализа применялась библиотека Requests.

Следуя из результатов анализа, представленных на Рисунке 15, необходимо отметить, что программа анализирует не только индексы удобочитаемости текстов, но и определенный ряд параметров, показатели которых непосредственно принимают участие в расчете итоговых значений. Среди данных параметров:

– Chars – общее количество знаков тексте;

– Spaces – количество пробелов в тексте;

– Letters – количество букв в тексте;

– N_words – количество слов;

– N_sentences – количество предложений;

– N_complex_words – количество слов с более чем 4-мя слогами;

– N_simple_words – количество слов до 4-х слогов включительно;

– Avg_slen – среднее число слов на предложение;

– Avg_syl – среднее число слогов на предложение;

– C_share – процент сложных слов от общего числа;

– W_syllabes – словарь из значений: число слогов и число слов с таким числом слогов в этом тексте.

Также необходимо расшифровать обозначения индексов удобочитаемости текстов:

– Grade_SMOG – уровень образования необходимый для понимания текста по формуле SMOG, выраженный в уровне обученности (количестве оконченных классов школы по системе K-12);

– Grade_ari – уровень образования необходимый для понимания текста по формуле automated readability index, выраженный в уровне обученности (количестве оконченных классов школы по системе K-12);

– Grade_cl – уровень образования необходимый для понимания текста по формуле Coleman-Liau, выраженный в уровне обученности (количестве оконченных классов школы по системе K-12);

– Grade_fk – уровень образования необходимый для понимания текста по формуле Flesch-Kinkaid, выраженный в уровне обученности (количестве оконченных классов школы по системе K-12);

– Grade_dc – уровень образования необходимый для понимания текста по формуле Dale-Chale, выраженный в уровне обученности (количестве оконченных классов школы по системе K-12);

– Index_SMOG – уровень образования необходимый для понимания текста по формуле SMOG, в годах обучения от 1 до бесконечности;

– Index_ari – уровень образования необходимый для понимания текста по формуле automated readability index, в годах обучения от 1 до бесконечности;

– Index_cl – уровень образования необходимый для понимания текста по формуле Coleman-Liau, в годах обучения от 1 до бесконечности;

– Index_fk – уровень образования необходимый для понимания текста по формуле Flesch-Kinkaid, в годах обучения от 1 до бесконечности;

– Index_dc – уровень образования необходимый для понимания текста по формуле Dale-Chale, в годах обучения от 1 до бесконечности.

2.4. Проблемы при сборе корпуса и их решение

В процессе сбора информации с сайта «Кинопоиск», нам пришлось столкнуться с некоторыми трудностями. В частности, сервис воспринимал парсер

как автоматическую программу, которая может нанести вред работе сайта. Было принято решение использовать user agent в целях определения сервисом парсера в качестве зарегистрированного пользователя, что давало возможность собирать информацию в неограниченном количестве, так как сайт больше не воспринимал работу программы в качестве взлома.

User agent представляет собой идентификационную строку клиентского приложения, которая использует сетевой протокол. Данная строка применяется в работе приложений, которые, в свою очередь, осуществляют доступ к веб-сервисам, например, в браузерах, поисковых роботах и мобильных приложениях.

При посещении того или иного веб-сервиса, как правило, веб-серверу посылается информация о приложении, использующем данную идентификационную строку. Информация представляет собой текстовую строку, которая является частью HTTP-запроса. Обычно она включает в себя:

– Название приложения;

– Версию приложения;

– Операционную версию используемого компьютера;

– Язык.

Ввиду того, что сам сервис «Кинопоиск» является проектом Яндекса, для идентификации в системе используются логин и пароль, которые были использованы пользователем при регистрации в системе Яндекс.

Пример того, как было осуществлено решение вышеуказанной проблемы в нашей работе представлен на Рисунке 16.

```
for y in range (1, 500):
    link = 'https://www.kinopoisk.ru/reviews/type/comment/period/month/perpage/200/page/'+str(y)+'/#list/'
    headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Safari/537.36'}
    r = requests.get(link, headers=headers, auth=('ЛОГИН', 'ПАРОЛЬ'))
```

Рисунок 16

Данное решение позволило собрать информацию о рецензиях и пользователях, необходимую для дальнейшей работы. Частный разбор информации, а также обоснование релевантности ее использования в исследовании будет представлено далее.

Еще одной сложностью в сборе корпуса стало ограниченное количество информации, находящейся в открытом доступе на сервисе «Кинопоиск». Так, сайт предоставляет пользователям доступ лишь к рецензиям, написанным за последний календарный месяц. Общее количество рецензий, когда-либо написанных пользователями на сайте «Кинопоиск», составляет 754473. Однако среднее количество отзывов, написанное в месяц, составляет примерно 2000. Таким образом, общий объем итогового корпуса составил 3872 рецензии, что, по нашему мнению, достаточно для проведения дальнейших исследований в рамках данной работы.

Также необходимо отметить, что количество пользователей отличается от количества написанных рецензий, из чего вытекает следующая проблема. Так, программа для парсинга страницы с отзывами не позволила сразу собирать информацию о друзьях и подписчиках пользователей, что, в свою очередь, могло бы оказать отрицательное влияние на качество работы, так как данный параметр, как было указано ранее, является одним из ключевых в нашем исследовании. Для решения проблемы было принято решение написать дополнительную программу, которая получала на вход ID пользователя в системе сайта, а далее собирала необходимые данные.

Далее нам предстояло свести две таблицы:

1) Таблицу, содержащую всю основную информацию о рецензиях;

2) Таблицу, содержащую информацию о пользователях, написавших рецензии.

Как было указано ранее, таблицы имели разный размер, так как многие пользователи оставляли более одной рецензии, соответственно, необходимо было каждой рецензии найти и обозначить количество друзей и подписчиков пользователей, оставивших их.

Решение данной проблемы было проведено в программе Excel. Так, была проведена сортировка данных по ID пользователей в обеих таблицах. Далее было необходимо копировать информацию о количестве друзей на такое количество строк вниз, чтобы итоговое количество строк совпадало с количеством рецензий,

написанных данным пользователем. Для этого была использована формула ВПР. Данная формула используется для поиска данных в таблице или же в заданном диапазоне по строкам. Так, в нашей работе данная формула была использована для поиска совпадений ID пользователя из данных таблицы с информацией исключительно по пользователям и данных таблицы с информацией о рецензиях. Затем итоговые данные по друзьям были скопированы в основную таблицу.

Таким образом, были решены все возникшие при сборе информации для корпуса проблемы.

2.5. Описание корпуса текстов

Нами был собран корпус текстов, содержащий 3872 рецензии на фильм, а также данные о пользователях, их написавших.

Итоговый корпус представляет собой таблицу, содержащую следующую информацию:

1) Никнейм пользователя на сайте, который использовался для дальнейшего поиска ID пользователя. Пример записи, а также наименование столбца для дальнейшего успешного форматирования представлен на Рисунке 17.

users_nick
AlexIFreedom
modus_exciter
panfilov.4lexej
BlueOkulus
ShivvaRudra
cyberlaw
suckmyska@gmail.cor
cyberlaw
n0ooneex

Рисунок 17

2) ID пользователя, который необходим для дальнейшего получения информации о количестве друзей и подписчиков пользователя. Пример записи, а также наименование столбца для дальнейшего успешного форматирования представлен на Рисунке 18.

id_users
/user/61495
/user/13820
/user/15806
/user/76204
/user/85852
/user/20260
/user/76210
/user/20260
/user/72275

Рисунок 18

3) Текст отзыва, который подавался на вход API с параметром text, описанному выше. Способ хранения текстов рецензий пользователей представлен на Рисунке 19.

r_text
Начало фильма походит на типичный американский блокбастер, где родитель(и) показывает себя в хорошем свете и в самом начале фильма закладывает
Уве Болла называют худшим режиссёром в мире. На мой взгляд, это незаслуженно, у него немало отличных фильмов, и даже «Один в темноте» и «Бладрейн»,
Знаю, очень многих пугает это цифра. Триста двадцать восемь серий для сериала это действительно очень много, но поверьте, это того стоит и сейчас я объясню
Современное русское кино, увы и ах, нас не балует. Нам предлагают фаст-фуд, щедро приправленный незатейливым юмором, более-менее качественным
Сериал оставил весьма двоякое ощущение. С одной стороны он красив и хорошо снят. Операторская работа на очень высоком уровне. Актерский состав силён.
В техническом аспекте перед нами скучное и пресное кино. Нет изысков. Нет эстетства. Нет даже увлекательного рассказа. Серые тона, беседы о заводе,
Это моя первая рецензия на фильм в жизни и думаю начать её с забавного факта, что фильм «Мертвые ласточки» мне посоветовала мама, которая и привила
Помните фильм «Отпуск, который не состоялся». Он будто предтеча этой ленте. Тот же герой., только постаревший. Та же профессия. То же беспокойство. Те же
Вот-вот посмотрела последнюю серию сериала, и сейчас могу уже здраво оценить всю картину. Сразу хочется похвалить операторскую работу, все на высшем

Рисунок 19

4) Показатели различных параметров текста, которые были использованы для расчета индексов удобочитаемости текстов по различным формулам (Flesch-Kinkaid; Dale-Chale readability formula; Coleman-Liau index; SMOG; automated readability index). Пример хранения данной информации представлен на Рисунке 20. Каждая ячейка, изначально имевшая текстовый формат, была преобразована в числовой формат с целью возможности дальнейшей работы с показателями.

Расшифровка метрик: процент сложных слов от общего числа; общее количество знаков тексте; среднее число слов на предложение; общее количество пробелов в тексте; общее количество слогов в тексте; общее количество слов в тексте; общее количество букв в тексте; общее количество предложений в тексте;

количество слов с более чем 4-мя слогами; количество слов до 4-х слогов (включительно); средняя длина слов, выраженная в слогах.

c_share	chars	avg_slen	spaces	syll	words	letters	sent	complex_w	simple_worc	avg_syl
9	2615	15	285	880	349	2088	24	30	319	3
7	2117	19	227	693	299	1668	16	21	278	2
6	4921	15	558	1695	703	3973	46	44	659	2
7	2685	10	295	918	377	2120	39	27	350	2
8	1130	7	124	390	157	897	21	12	145	2
12	2947	10	310	1013	387	2382	38	45	342	3
7	1897	10	211	651	261	1528	27	19	242	2
7	1275	8	149	425	192	996	25	13	179	2
8	1873	9	201	646	263	1490	30	20	243	2

Рисунок 20

5) Также таблица содержит показатели индексов удобочитаемости текстов, выраженные в необходимых для понимания текста классах обученности по системе K-12. Данный показатель указывает на минимальное количество успешно оконченных читателем классов. Примеры столбцов, содержащих данную информацию, представлены на Рисунке 21.

grade_smog	grade_ari	grade_cl	grade_fk	grade_dc
10 - 11-й кла	10 - 11-й кла	10 - 11-й кла	10 - 11-й кла	7 - 9-й класс
10 - 11-й кла	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс
7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс
7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс
7 - 9-й класс	7 - 9-й класс	4 - 6-й класс	7 - 9-й класс	4 - 6-й класс
10 - 11-й кла	10 - 11-й кла	10 - 11-й кла	10 - 11-й кла	7 - 9-й класс
7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс
4 - 6-й класс	4 - 6-й класс	4 - 6-й класс	1 - 3-й класс	4 - 6-й класс
7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс	7 - 9-й класс

Рисунок 21

6) Стоит отметить, что столбцы с обозначением «idex_» несут в себе ту же информацию, что и столбцы с обозначением «grade_», однако имеют числовой формат, необходимый для проведения дальнейших исследований. Пример столбцов с указанием числовых индексов удобочитаемости текстов представлен на Рисунке 22.

index_fk	index_cl	index_dc	index_ari	index_smog
9,93	10	9	10	10
8,35	8	9	9	10
8,6	8	8	9	9
7,15	7	7	7	7
7,15	6	6	7	7
9,91	10	9	10	10
7,99	8	7	8	7
3,37	4	6	4	6
7,17	7	7	7	7
7,56	8	7	8	8

Рисунок 22

7) Далее в таблице представлены столбцы популярности, показатели которых являются объектом нашего исследования. Изначально программой были собраны показатели положительных и отрицательных оценок каждой рецензии. Далее при помощи формулы Excel «СУММ» были получены значения итоговой популярности отзыва, которая представляет собой общее количество положительных и отрицательных оценок. Пример данных столбцов в таблице представлен на Рисунке 23.

positive	negative	popularity
0	0	0
0	0	0
0	2	2
9	26	35
1	1	2
0	0	0
0	3	3
0	1	1
2	5	7
9	28	37

Рисунок 23

8) Завершающие столбцы таблицы представляют собой список «друзей» и «добавили в друзья». В работе использовался второй показатель, который демонстрирует количество «подписчиков», имеющих возможность просмотра

активности пользователя на сайте. Данный параметр показан в последнем столбце, пример представлен на Рисунке 24.

friends_all	friends_with
0	0
4	4
0	0
0	0
19	7
61	522
0	0
61	522
0	0
0	0

Рисунок 24

2.6. Расчет корреляции

Для проведения расчета корреляции была использована библиотека Pandas. Для успешной работы была создана сокращенная версия изначальной таблицы. В итоговую базу данных была включена следующая информация:

1) ID пользователя, который возможно использовать для дальнейшей идентификации пользователя в базе данных и на сайте с целью оценки его рецензий;

2) Непосредственно индексы удобочитаемости:

– Индекс Флеша-Кинкейда;

– Индекс Колемана-Лиау;

– Индекс Дейла-Чела;

– ARI-индекс;

– Индекс SMOG.

В итоговой версии таблицы не использовались показатели с пометкой «grade_», так как они лишь дублируют значение индекса на естественном языке, а программа может оценивать корреляцию и работать только с числовыми значениями.

3) Показатель популярности. В итоговую версию таблицы было принято решение не включать отдельные показатели положительных и отрицательных оценок рецензии, так как в работе оценивалась итоговая популярность отзыва.

4) Количество друзей и подписчиков.

В целях осуществления дальнейшего анализа данных был создан датафрейм на основе итоговой таблицы. Первые пять строк получившегося датафрейма представлены на Рисунке 25.

	id_users	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
0	/user/61495332/	9,93	10	9	10	10	0	0	0
1	/user/13820846/	8,35	8	9	9	10	0	4	4
2	/user/15806801/	8,6	8	8	9	9	2	0	0
3	/user/76204605/	7,15	7	7	7	7	35	0	0
4	/user/858522/	7,15	6	6	7	7	2	19	7

Рисунок 25

Всего датафрейм содержит 3910 строк и 9 столбцов. Такого количества данных достаточно для проведения дальнейшего исследования.

Были собраны статистические показатели данных, содержащихся в датафрейме. Сводная таблица показателей представлена на Рисунке 26.

	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
count	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000
mean	9.405627	8.930435	8.495908	9.409463	9.502046	16.324808	33.586445	57.129923
std	3.166670	2.476137	2.343895	3.098017	2.585528	65.421031	107.538082	177.743091
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	7.000000	7.000000	7.000000	7.000000	8.000000	1.000000	0.000000	0.000000
50%	9.000000	9.000000	8.000000	9.000000	9.000000	12.000000	1.000000	1.000000
75%	11.000000	10.000000	10.000000	11.000000	11.000000	19.000000	15.000000	20.750000
max	31.000000	19.000000	26.000000	30.000000	29.000000	2732.000000	1487.000000	2733.000000

Рисунок 26

Как видно из таблицы, значения популярности имеют большой разброс, так, при минимальном значении 0, максимальное – 2732. Мы предполагаем, что

подобная разница в показателях может в дальнейшем оказать влияние на значение корреляции.

2.7. Общие показатели корреляции

В целях выявления зависимости показателей популярности рецензии на фильмы от индекса удобочитаемости данной рецензии, нами был проведен корреляционный анализ данных, описанных выше. Для анализа была использована библиотека Pandas, а для визуализации применялись библиотеки Matplotlib и Seaborn.

Корреляционный анализ данных был проведен с использованием трех методов.

Во-первых, данные были проанализированы при помощи метода Пирсона. Результаты анализа представлены на Рисунке 27.

	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
index_fk	1.000000	0.930032	0.916232	0.970944	0.911944	0.011374	0.201818	0.218826
index_cl	0.930032	1.000000	0.823592	0.960748	0.829124	0.009283	0.141285	0.172292
index_dc	0.916232	0.823592	1.000000	0.888009	0.978099	0.002004	0.199218	0.219296
index_ari	0.970944	0.960748	0.888009	1.000000	0.885060	0.003529	0.197081	0.215573
index_smog	0.911944	0.829124	0.978099	0.885060	1.000000	0.005858	0.202339	0.219831
popularity	0.011374	0.009283	0.002004	0.003529	0.005858	1.000000	-0.005136	-0.005958
friends_all	0.201818	0.141285	0.199218	0.197081	0.202339	-0.005136	1.000000	0.837383
friends_with	0.218826	0.172292	0.219296	0.215573	0.219831	-0.005958	0.837383	1.000000

Рисунок 27

Исходя из полученных данных, мы можем сделать вывод, что показатель популярности слабо коррелирует с показателями индексов удобочитаемости (максимальное модульное значение корреляции равно 0.011374 – с индексом Флеша-Кинкейда). Более того, в данном анализе не была выявлена зависимость популярности рецензии фильма и количества друзей и подписчиков пользователя.

Во-вторых, корреляционный анализ данных был проведен при помощи метода Спирмена. Результаты анализа продемонстрированы на Рисунке 28.

	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
index_fk	1.000000	0.938442	0.895677	0.963325	0.893621	0.155788	0.217587	0.221779
index_cl	0.938442	1.000000	0.824495	0.972183	0.829754	0.150523	0.215702	0.222643
index_dc	0.895677	0.824495	1.000000	0.863430	0.977725	0.137947	0.195661	0.202482
index_ari	0.963325	0.972183	0.863430	1.000000	0.861326	0.151344	0.217641	0.222855
index_smog	0.893621	0.829754	0.977725	0.861326	1.000000	0.143804	0.204130	0.210772
popularity	0.155788	0.150523	0.137947	0.151344	0.143804	1.000000	0.045164	0.045758
friends_all	0.217587	0.215702	0.195661	0.217641	0.204130	0.045164	1.000000	0.971126
friends_with	0.221779	0.222643	0.202482	0.222855	0.210772	0.045758	0.971126	1.000000

Рисунок 28

Проанализировав выходные данные корреляционных показателей, мы пришли к выводу, что, аналогично использованию предыдущего метода, метод Спирмена не выявляет сильной корреляции показателя «popularity» с индексами читабельности (максимальное модульное значение корреляции равно 0.155788 – с индексом Флеша-Кинкейда). Тем не менее, стоит отметить, что использование данного метода повысило значения корреляции, которые, несмотря на этот факт, остаются недостаточными, чтобы заявлять о существовании корреляции между популярностью кинорецензий и их индексами удобочитаемости. Модульные значения корреляции популярности отзывов на фильмы и количества друзей и подписчиков пользователей между двумя вышеуказанными методами достаточно близки, однако при учитывании знака показателя обнаруживается, что зависимость, при исследовании методом Пирсона являвшаяся обратной, в методе Спирмена стала прямой. Тем не менее, значения слишком низкие для обнаружение явной корреляции между показателями.

В-третьих, был использован метод Кендалла. Результаты исследования представлены на Рисунке 29.

	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
index_fk	1.000000	0.857346	0.792032	0.900414	0.785180	0.157198	0.164643	0.167852
index_cl	0.857346	1.000000	0.705131	0.925890	0.706477	0.147909	0.164479	0.169966
index_dc	0.792032	0.705131	1.000000	0.748727	0.942045	0.139125	0.150303	0.155499
index_ari	0.900414	0.925890	0.748727	1.000000	0.742134	0.153539	0.164842	0.168805
index_smog	0.785180	0.706477	0.942045	0.742134	1.000000	0.144254	0.156171	0.161042
popularity	0.157198	0.147909	0.139125	0.153539	0.144254	1.000000	0.033989	0.034312
friends_all	0.164643	0.164479	0.150303	0.164842	0.156171	0.033989	1.000000	0.889958
friends_with	0.167852	0.169966	0.155499	0.168805	0.161042	0.034312	0.889958	1.000000

Рисунок 29

Выходные показатели близки к полученным при использовании вышеуказанного метода (максимальное модульное значение корреляции равно 0.157198 – с индексом Флеша-Кинкейда), соответственно, также недостаточно оснований для утверждения факта о наличии зависимости между показателями «популярность» – «читабельность».

Для наглядности была составлена тепловая карта, отражающая зависимость между данными в таблице с использованием метода Пирсона. Данная карта представлена на Рисунке 30.

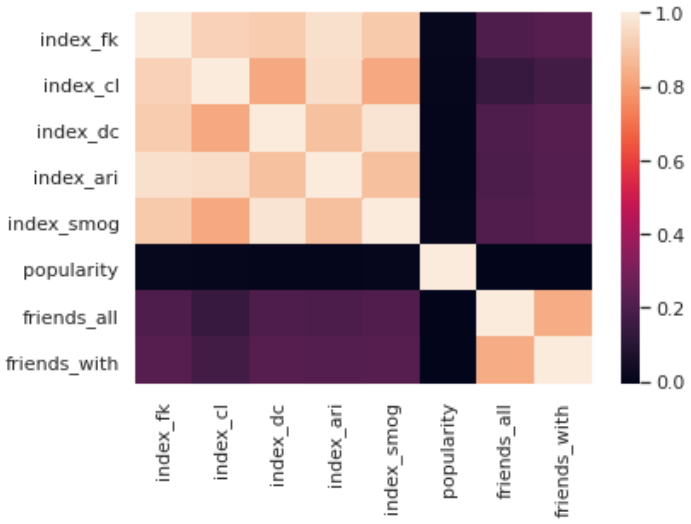


Рисунок 30

Как видно из представленных на рисунке данных, показатель популярности не коррелирует ни с какими данными в таблице.

Далее будет описан анализ корреляции каждого индекса удобочитаемости текстов отдельно с показателями популярности рецензии и популярности пользователя, написавшего данную рецензию.

2.8. Частные показатели корреляции

В основе дальнейшего корреляционного анализа лежит выявление зависимостей между отдельными элементами исходной базы данных. Так, в данном анализе используется:

– Отдельный индекс удобочитаемости;

– Количество подписчиков пользователей («friends_with»);

– Показатель популярности рецензии.

Первым анализируемым индексом удобочитаемости стал индекс SMOG. Результаты анализа представлены на Рисунке 31.

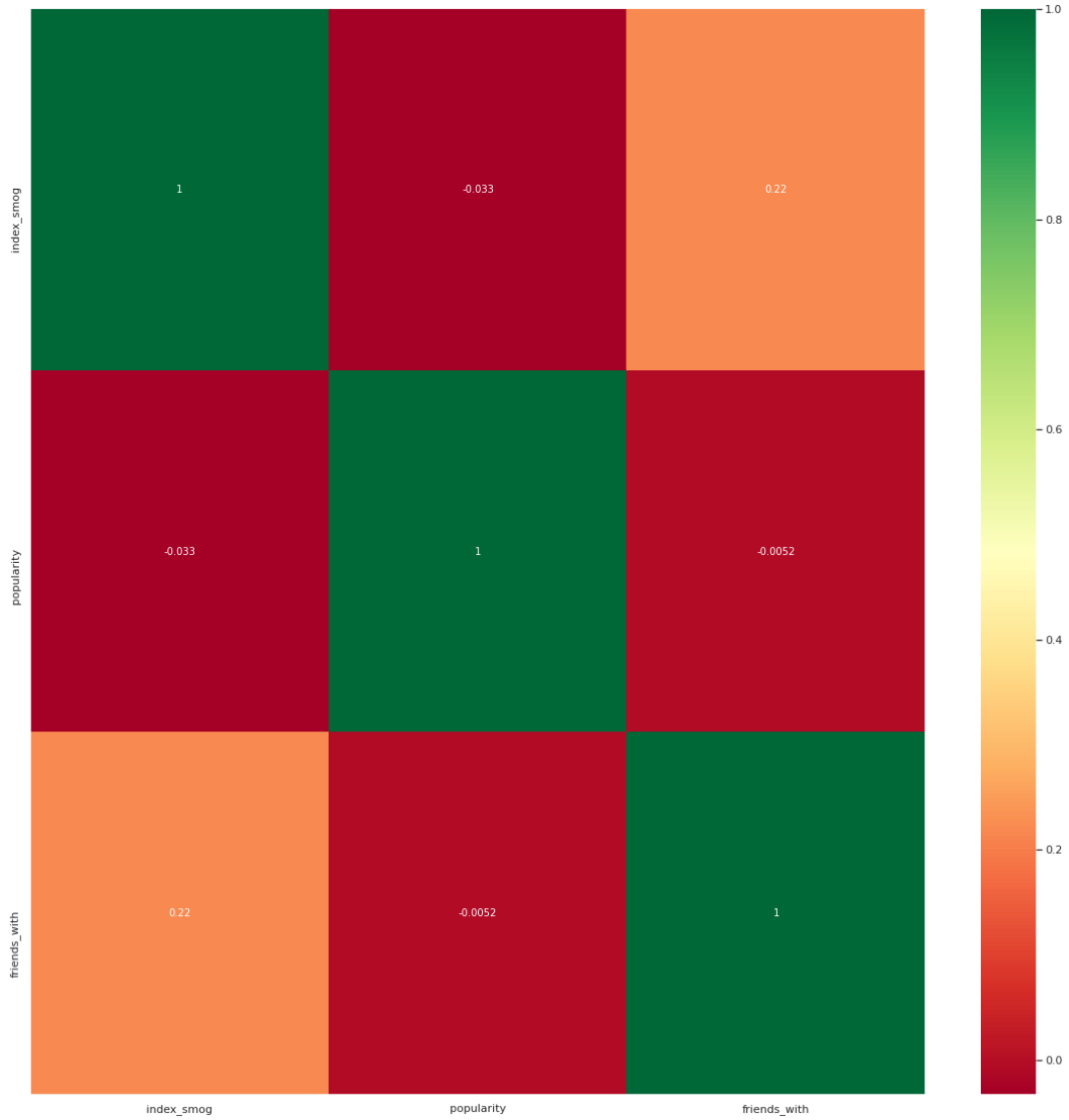


Рисунок 31

Исходя из представленного анализа, мы можем прийти к выводу, что показатель популярности не скоррелирован с показателями читабельности и количеством подписчиков пользователя. Стоит отметить, что небольшая корреляция между индексом удобочитаемости и количеством подписчиков, по нашему мнению, является случайной.

Далее был проведен аналогичный анализ с использованием индекса удобочитаемости ARI. Тепловая карта итогового анализа представлена на Рисунке 32.

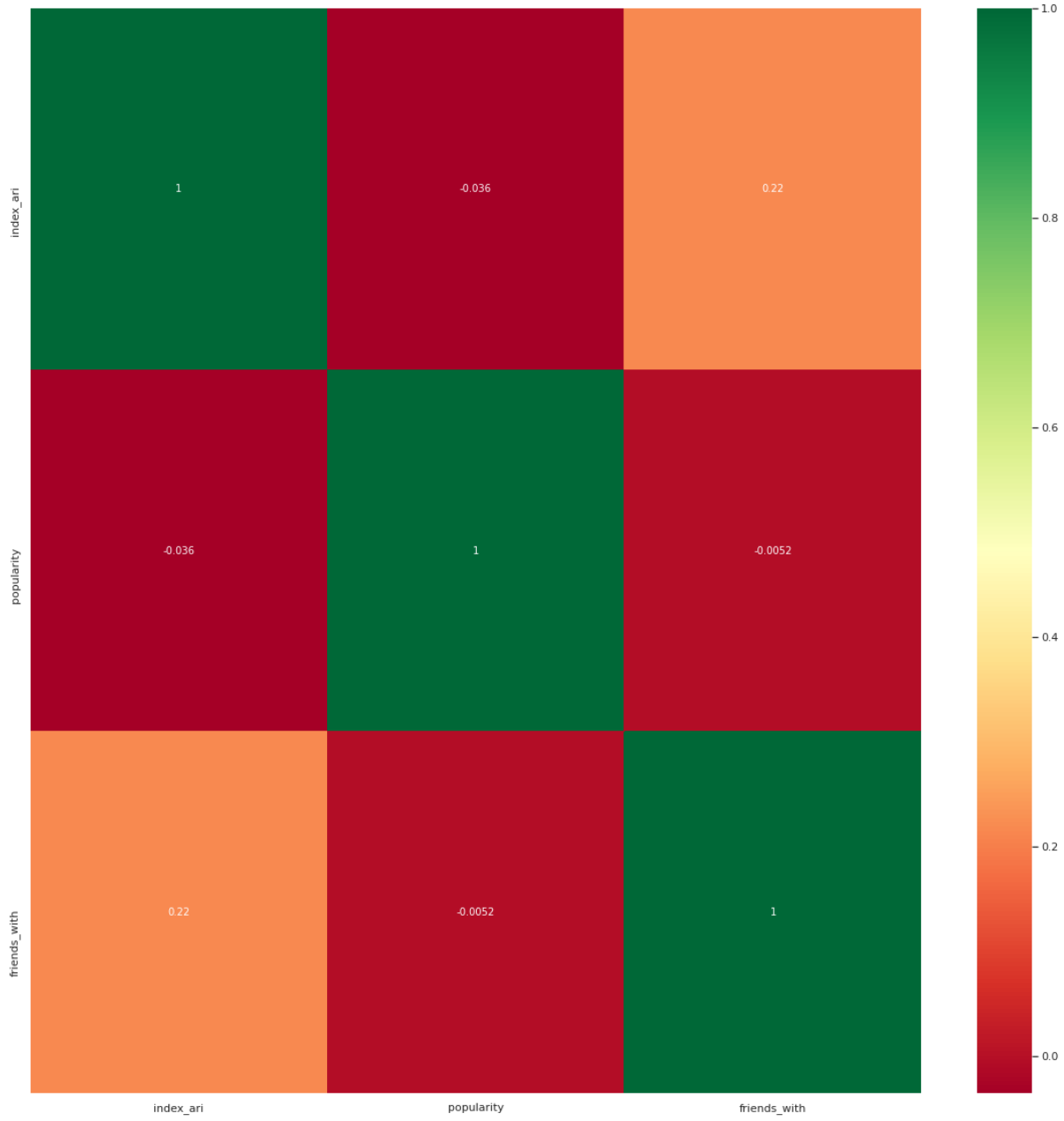


Рисунок 32

Анализ показал минимальную, даже отсутствующую корреляцию между популярностью рецензии и количеством подписчиков пользователя. Модульный показатель корреляции по схеме «популярность» – «читабельность» немного выше, но также недостаточен для заключения о наличествовании зависимости.

Далее проводился корреляционный анализ, в основу которого был положен индекс читабельности Дейла-Челла. Итоги анализа показаны на Рисунке 33.

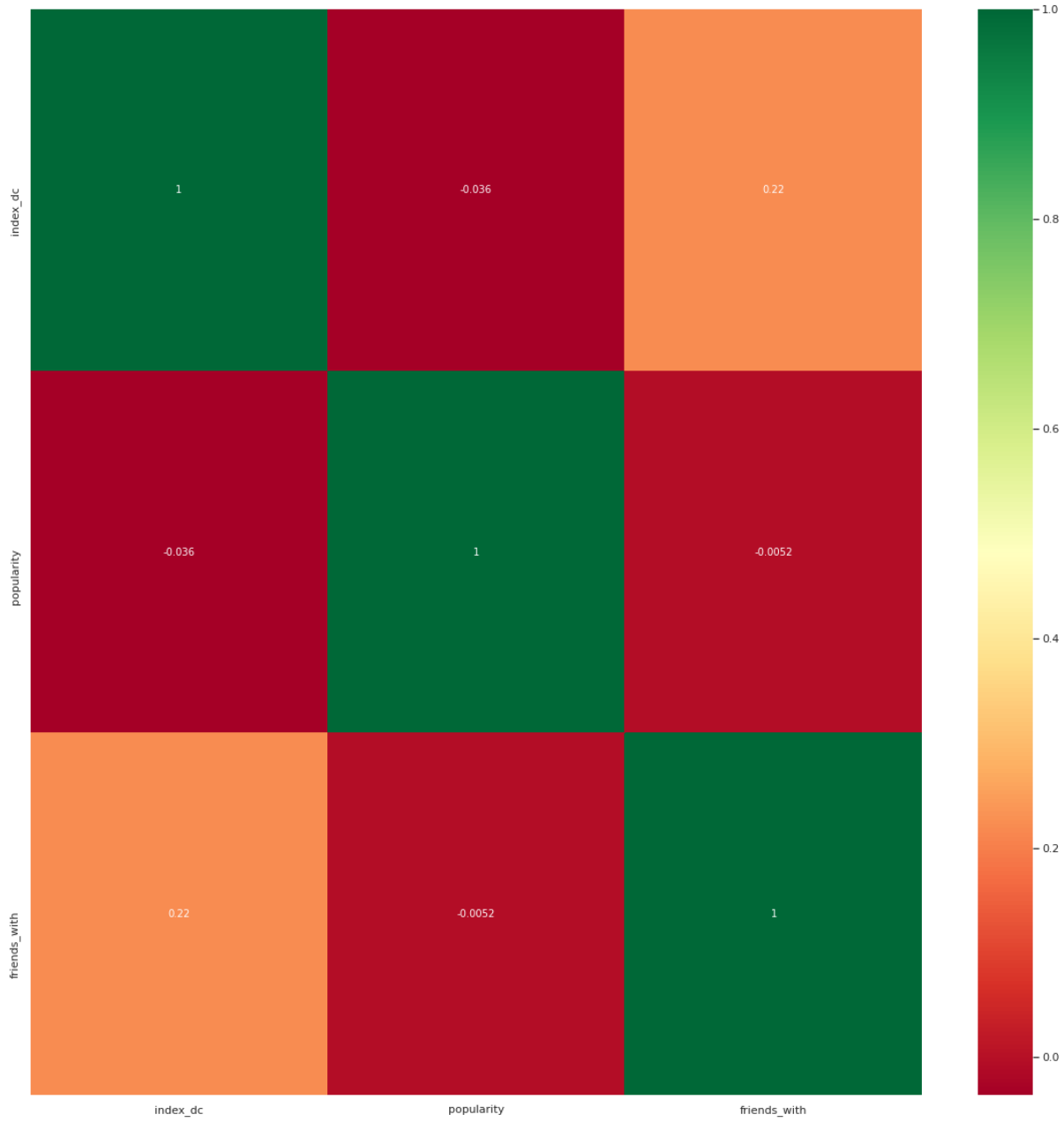


Рисунок 33

Выходные показатели полностью совпали с показателями при анализе с использованием индекса ARI.

Результаты корреляционного анализа с использованием формулы Флеша-Кинкейда представлены на Рисунке 34.

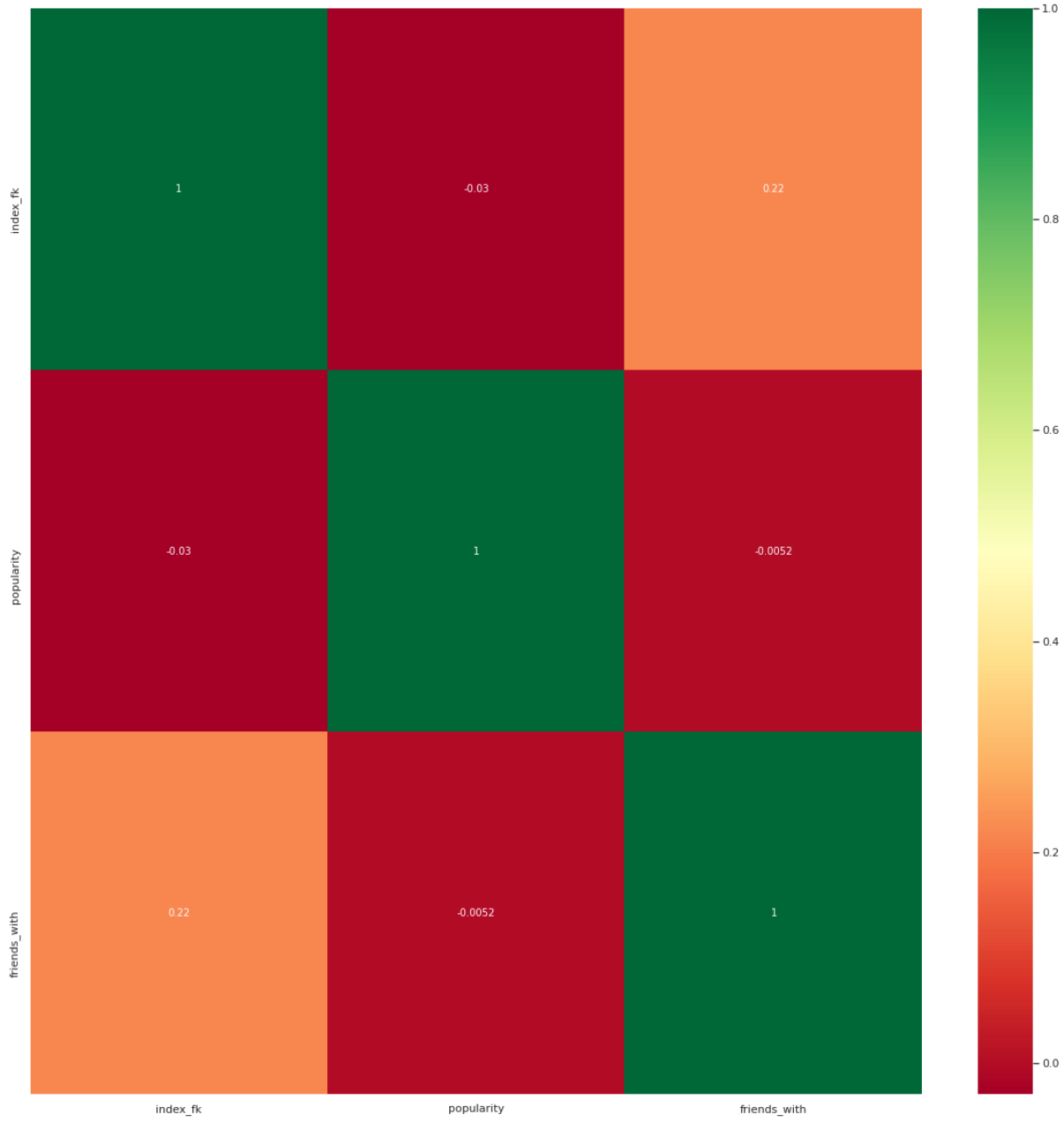


Рисунок 34

Показатели зависимости популярности отзыва от количества подписчиков пользователя совпали с вышеуказанными; выявление корреляции между популярностью отзыва и индексом удобочитаемости текста, рассчитанным по формуле Флеша-Кинкейда, не является успешным, показатель остается стабильно низким.

В завершающем частном анализе была использована формула Колемана-Лиау. Результаты представлены на Рисунке 35.

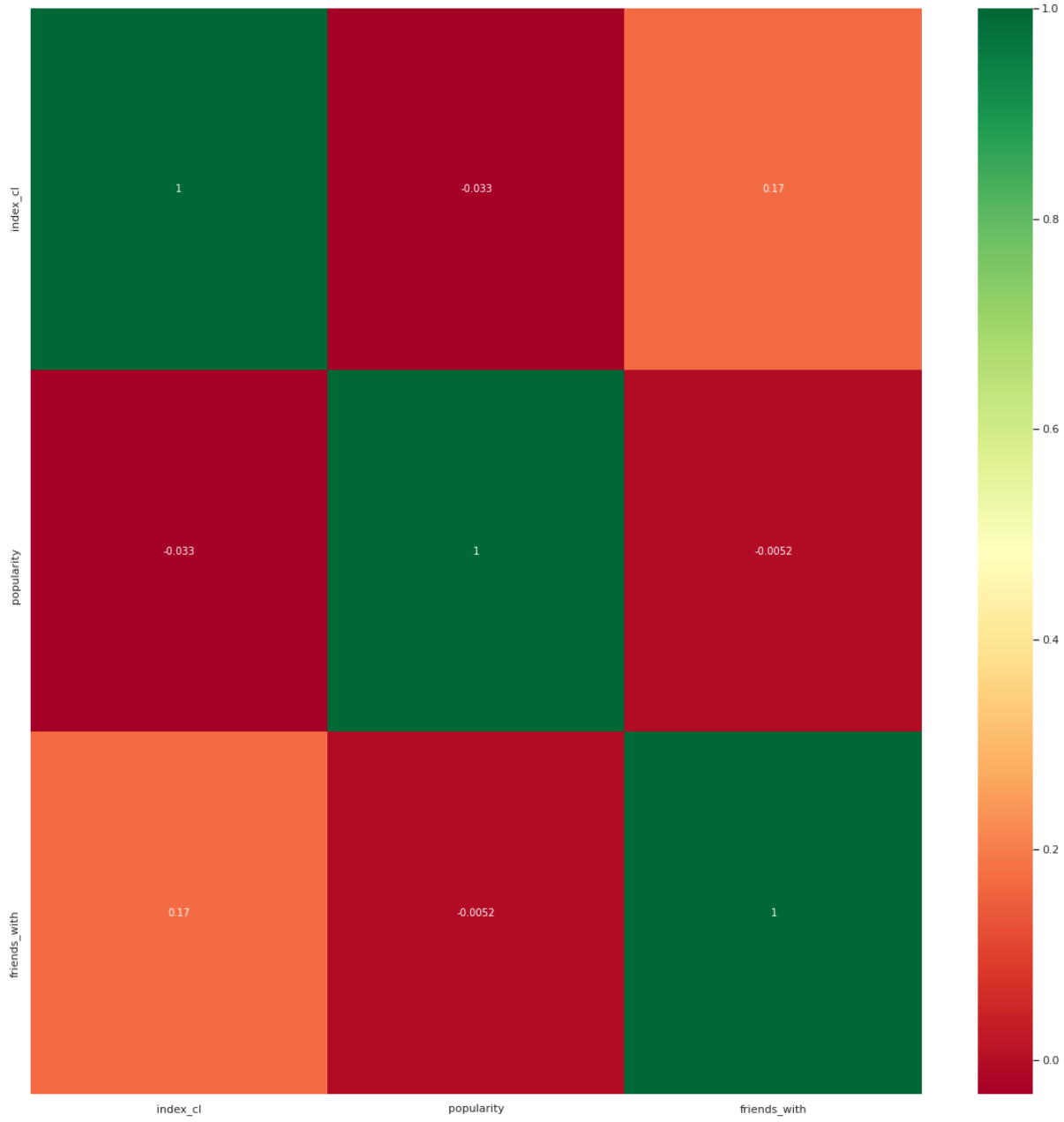


Рисунок 35

В данном случае снова совпадает низкий уровень зависимости популярности рецензии от количества подписчиков пользователя. Уровень корреляции популярности отзыва и его читабельности также обладает близким показателем.

Таким образом, наибольшим модульным показателем зависимости популярности отзыва от показателя индекса удобочитаемости является показатель равный 0.036.

2.9. Алгоритм выделения признаков

Выделение признаков (feature selection) является важнейшей частью процесса машинного обучения. Стоит отметить, что в данной работе мы использовали данный алгоритм не для обучения и дальнейшего предсказания, а для определения важности параметров в базе данных. Было принято решение считать показатель популярности рецензии целевой переменной. Анализ был проведен на тех же отдельных наборах данных, что и в пункте выше, таким образом, было проведено пять анализов, различительным критерием являлся показатель индекса удобочитаемости.

С технической точки зрения, в работе использовалась библиотека sklearn, в частности, модель extra trees classifier и ее модуль feature_importances.

Результаты исследования представлены:

1) Для индекса SMOG на Рисунке 36

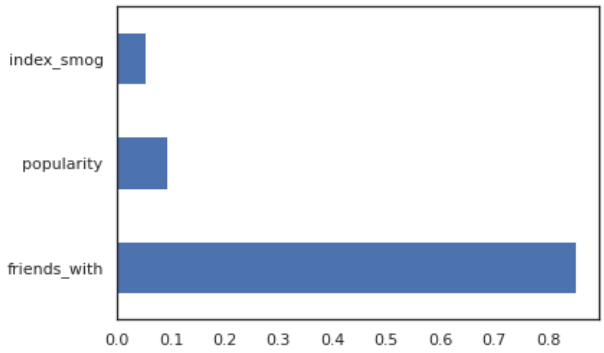


Рисунок 36

2) Для индекса ARI на Рисунке 37

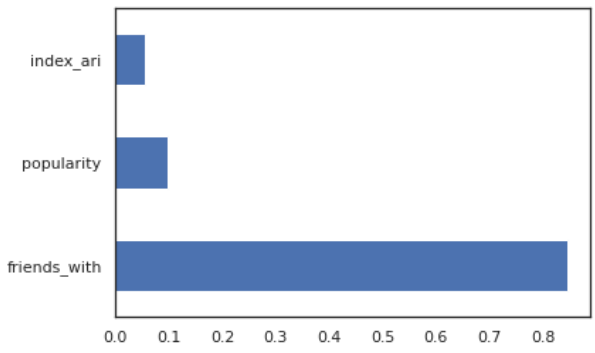


Рисунок 37

3) Для индекса Дейла-Челла на Рисунке 38

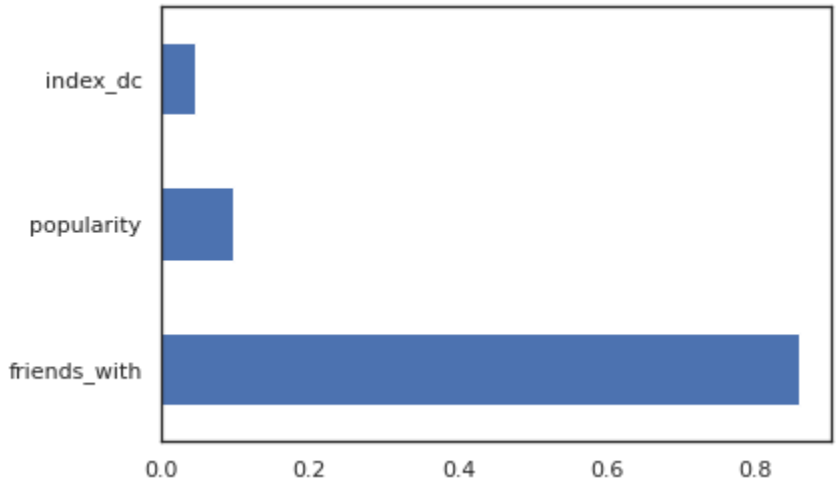


Рисунок 38

4) Для индекса Флеша-Кинкейда на Рисунке 39

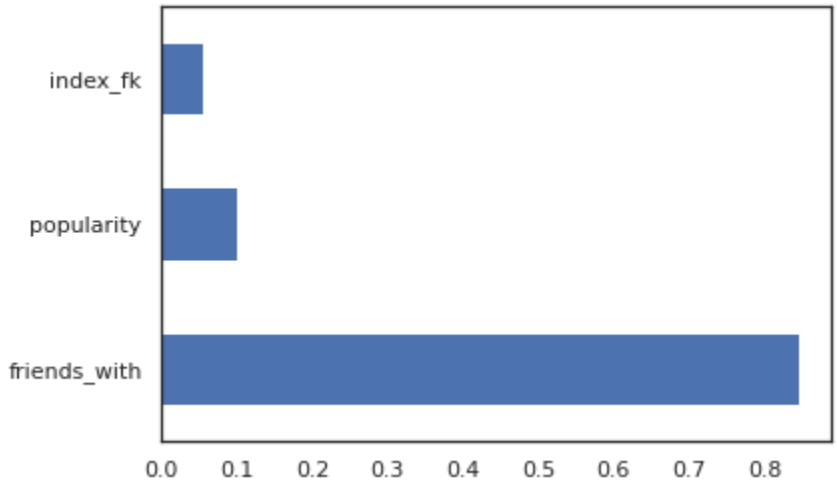


Рисунок 39

5) Для индекса Колемана-Лиау на Рисунке 40

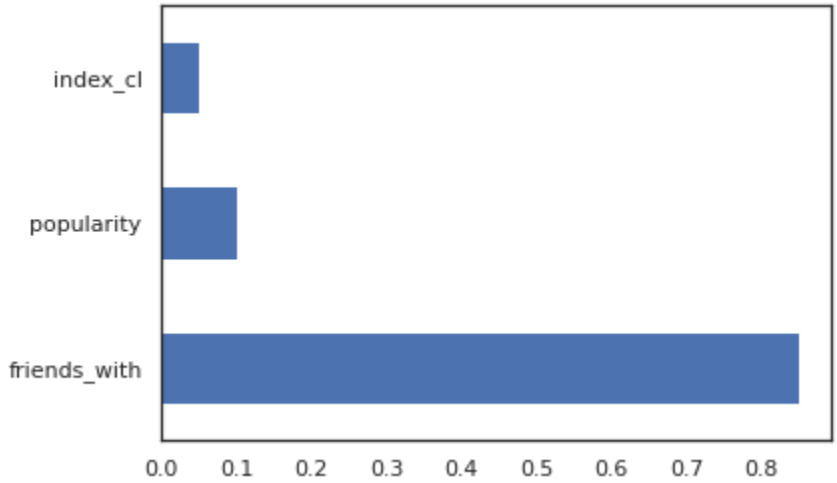


Рисунок 40

Результаты оценки важности параметров на всех пяти наборах данных достаточно близки. Исходя из них, мы можем прийти к выводу, что индекс удобочитаемости является наименее важным параметром. Большое влияние оказывает количество подписчиков, более того, именно с данным показателем была выявлена лучшая корреляция показателя популярности рецензии на предыдущем этапе. Соответственно, мы предварительно можем прийти к выводу, что количество подписчиков пользователя оказывает большее влияние на популярность рецензии, чем удобочитаемость текста отзыва.

2.10. Поиск аномалий в зависимости цепи «популярность» – «читабельность»

В ходе работы не было выявлено сильной корреляции между параметром популярности отзыва на кинофильм и его читабельностью. Данный факт натолкнул нас на мысль о том, что зависимости между данными показателями нет. Чтобы доказать это, представляется необходимым обнаружить аномалии в характеристиках рецензий.

Под аномалиями в данном контексте следует понимать отклонение от ожидаемых показателей. Так, изначально мы предполагали, что существует зависимость между популярностью рецензии и читабельностью ее текста. Соответственно, в наборе данных должны быть рецензии с низкими показателями удобочитаемости и высокими показателями популярности, так как мы основываемся на предположении о том, что чем проще читать текст отзыва, тем он будет популярнее. В ходе работы явной корреляции выявлено не было, соответственно, необходимо обнаружить отзывы с высокими показателями читабельности и при этом высокими показателями популярности и, наоборот, с низкими показателями читабельности и низкими показателями популярности.

Для начала необходимо представить статистическую информацию по всему корпусу. На данном этапе работы необходимо ориентироваться на среднее

значение параметров, показанное в строке «mean» в нижепредставленном Рисунке 41.

	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
count	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000	3910.000000
mean	9.405627	8.930435	8.495908	9.409463	9.502046	16.324808	33.586445	57.129923
std	3.166670	2.476137	2.343895	3.098017	2.585528	65.421031	107.538082	177.743091
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	7.000000	7.000000	7.000000	7.000000	8.000000	1.000000	0.000000	0.000000
50%	9.000000	9.000000	8.000000	9.000000	9.000000	12.000000	1.000000	1.000000
75%	11.000000	10.000000	10.000000	11.000000	11.000000	19.000000	15.000000	20.750000
max	31.000000	19.000000	26.000000	30.000000	29.000000	2732.000000	1487.000000	2733.000000

Рисунок 41

Ориентируясь на среднее значение параметра популярности текстов, равное 16, и среднее значение параметра читабельности, расположенное в промежутке от 8,4 до 9,5 единиц, мы приняли решение сделать сортировку значений в сокращенной версии базы данных, создав сортировку по максимальному значению индекса удобочитаемости по Флешу-Кинкейду, так как именно данная формула на сегодняшний день является наиболее популярной в научных исследованиях.

Отрывок таблицы представлен на Рисунке 42. Желтое выделение свидетельствует об обнаружении аномалии в показателях данной рецензии. Примеры отзывов с аномальными показателями представлены в Приложении 2.

	id_users	index_fk	index_cl	index_dc	index_ari	index_smog	popularity	friends_all	friends_with
2	/user/1479/	31,42	18,98	25,99	30,45	29,04	59,49	1487,00	2733,00
3	/user/1479/	25,85	13,15	22,73	24,16	23,72	47,88	1487,00	2733,00
4	/user/49576	25,05	17,79	19,34	23,11	22,06	3,00	3,00	10,00
5	/user/66162	23,71	17,26	19,49	23,26	22,23	45,48	246,00	526,00
6	/user/65711	23,59	18,57	20,10	23,98	22,88	46,86	246,00	526,00
7	/user/31514	23,29	10,96	19,67	21,18	18,38	39,56	246,00	526,00
8	/user/14818	23,25	12,69	18,98	22,03	18,63	40,66	246,00	526,00
9	/user/49576	23,17	17,89	19,92	22,30	22,42	0,00	3,00	10,00
10	/user/14501	22,16	17,28	17,57	21,51	20,02	11,00	23,00	26,00
11	/user/29349	21,96	18,40	15,58	19,79	13,75	33,54	23,00	26,00
12	/user/49576	21,89	18,59	19,86	21,53	21,18	13,00	3,00	10,00
13	/user/49576	21,87	16,62	16,86	21,03	19,25	19,00	3,00	10,00
14	/user/49576	21,52	18,11	17,40	21,06	19,22	58,00	3,00	10,00
15	/user/49576	21,31	18,06	18,23	20,88	19,82	48,00	3,00	10,00

Рисунок 42

Первый же отобранный отзыв (с максимальным значением читабельности) имеет популярность, значительно превышающую среднее значение (59,49 при среднем 16). Стоит также отметить, что в основном большинство популярных отзывов оставляют пользователи с большим количеством друзей. Тем не менее, социологические исследования на данную тему не являются целью настоящей работы, однако дают возможность для дальнейшего изучения в данной сфере.

На Рисунке 43 представлена другая сторона аномалий, описанная выше.

/user/57886	0,67	1,40	3,90	0,56	4,25	4,81	0,00	0,00
/user/78661	0,92	2,35	3,87	1,86	4,38	6,24	0,00	0,00
/user/55865	0,93	1,93	3,25	1,63	3,59	5,22	0,00	0,00
/user/23921	1,57	1,92	3,71	1,70	4,26	5,96	0,00	0,00
/user/16898	1,58	1,38	3,77	0,99	4,31	5,30	0,00	0,00
/user/14358	1,81	3,66	3,98	3,07	4,30	25,00	4,00	3,00
/user/36684	1,82	1,82	4,63	1,14	5,32	7,00	0,00	0,00
/user/73193	1,91	2,17	3,44	2,54	3,96	0,00	6,00	6,00
/user/37033	1,92	0,71	4,60	1,26	5,06	6,32	0,00	0,00
/user/67106	1,96	2,64	4,02	2,00	4,47	0,00	23,00	12,00
/user/13520	2,12	2,48	4,21	2,61	4,82	7,43	1,00	9,00
/user/56640	2,19	4,09	3,39	3,45	2,03	5,47	0,00	0,00
/user/57181	2,24	2,16	5,09	1,97	5,74	2,00	5,00	4,00
/user/65028	2,36	2,28	3,88	2,37	4,47	1,00	1,00	1,00

Рисунок 43

Исходя из представленных данных, мы можем прийти к выводу, что тексты с максимально низкими показателями читабельности не имеют тенденции к тому, чтобы стать популярными.

2.11. Анализ результатов

Проведенные эксперименты показали отсутствие корреляции между популярностью отзывов на фильмы в сервисе «Кинопоиск» и их читабельностью, вычисленной по пяти наиболее распространенным формулам: SMOG, ARI, Coleman-Liau, Flesch-Kinkaid, Dale-Chale. Этот результат хорошо соотносится с полученными мною ранее выводами о невозможности применения формул

удобочитаемости при работе с экзаменационными текстами на английском языке [4].

Возможно, более адекватная оценка удобочитаемости текстов на русском языке может быть проведена с использованием метрики И. В. Обороневой, которая была целенаправленно разработана для русского языка [21]. Однако в открытом доступе отсутствует инструмент для автоматического определения удобочитаемости с помощью этой метрики. Создание такого инструмента и оценка и усовершенствование данной метрики могут стать отдельной исследовательской задачей.

На результаты работы могла повлиять не совсем корректная работа API для автоматического определения уровня удобочитаемости. Данная программа была разработана с применением алгоритмов машинного обучения на материале корпуса текстов русского языка различной направленности, а в основу были положены формулы для определения удобочитаемости текстов на английском языке. Мы предполагаем, что такой подход к выработке коэффициентов для текстов на русском языке мог показать неверные результаты оценки простоты текстов.

В процессе работы было обнаружено большое количество аномальных рецензий, показатели которых обратны изначально предполагавшимся: некоторые «неудобочитаемые» тексты оказывались очень популярными. Примеры таких рецензий приводятся в Приложении 2.

Заметное влияние на популярность отзыва оказывает количество подписчиков пользователя. Однако изучение данного вопроса в рамках ведущегося исследования невозможно ввиду его более социологической, нежели лингвистической направленности.

Заключение

Читабельность текста является важным параметром в определении простоты понимания текста читателем. Большинство формул определения индекса удобочитаемости текстов было разработано для определения читабельности текстов на английском языке. Тем не менее, исследователи не оставляют попыток модификации данных формул с целью применения их к русскоязычным текстам. Именно с использованием подобных «русифицированных» формул было проведено данное исследование.

Материалом для исследования выступили более трех тысяч текстов рецензий, написанных пользователями к фильмам на русском языке в сервисе «Кинопоиск». Корпус отзывов был собран и размечен по ряду параметров, среди которых основные показатели для вычисления читабельности текста (длина слов, предложений; количество слогов, слов в тексте; средние показатели), которые были в дальнейшем использованы для определения индексов читабельности по пяти наиболее распространенным формулам:

- SMOG;
- ARI;
- Coleman-Liau;
- Flesch-Kinkaid;
- Dale-Chale.

Результаты вычислений также были занесены в корпус. Наконец, в таблицу были включены данные о популярности рецензий. Под популярностью в данной работе мы понимаем общий отклик пользователей на отзыв, то есть суммарное количество положительных и негативных отметок.

Далее мы провели корреляционный анализ, целью которого стало выявление зависимости между показателями индексов удобочитаемости рецензии и ее популярностью. Данный анализ был проведен при помощи коэффициентов Пирсона, Спирмена и Кендалла. Анализ показал отсутствие корреляций между

удобочитаемостью отзыва, посчитанной по указанным формулам, и его популярностью.

В работе также был проведен анализ отзывов с помощью алгоритма выделения признаков (feature selection), который опять же указал на отсутствие значимости такой метрики как удобочитаемость текста в контексте связи с популярностью рецензии на кинофильм.

Наконец, нами был выполнен поиск аномалий в связи «читабельность»-«популярность». Было выявлено, что существует большое количество рецензий, имеющих хорошие (низкие) показатели читабельности, при этом плохие (низкие) показатели популярности. Более того, часто встречается и противоположная ситуация, когда встречаются тексты с плохими (высокими) показателями удобочитаемости, при этом имеющие хорошие (высокие) показатели популярности.

Исследование показало, что использование указанных индексов удобочитаемости текстов на русском языке в качестве метрики в контексте популярности отзывов не совсем корректно, так как наблюдается отсутствие связи между показателями читабельности и показателями отклика у отзыва. Мы предполагаем, что причина данного явления может быть связана с отсутствием разработанных и протестированных формул именно для русского языка.

Наконец, под сомнение была поставлена корректность работы используемого API, так как изначально данное программное обеспечение опиралось на формулы, используемые для исследования текстов на английском языке. Тем не менее, применение формул удобочитаемости в исследовании англоязычных текстов также не всегда может давать валидные результаты [4].

Полученный результат указывает на то, что формулы для определения индексов удобочитаемости текстов необходимо совершенствовать, в особенности, формулы, применяемые для определения читабельности текстов на русском языке. Для усовершенствования существующих (разработки новых коэффициентов на основе имеющихся «англоязычных» формул) или же создания новых формул необходимо проведения ряда исследований, в частности, с применением алгоритмов

машинного обучения, а также психо- и нейролингвистических исследований вопроса простоты усваивания прочитанного текста реципиентом.

Список литературы

1. Бююль А. SSPS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. / А. Бююль, П. Цёфель. – СПб. : ООО «ДиаСофтЮП», 2005. – 608 с.

2. Кисельников А. С. Параметры сложности экзаменационных текстов / А. С. Кисельников, М. И. Солнышкина // Вестник Волгоградского государственного университета. Сер. 2: Языкознание. – 2015. – №1 (25). – С. 99-107.

3. Кобзарь А. И. Прикладная математическая статистика / А. И. Кобзарь. – М. : Физматлит, 2006. – 816 с.

4. Короткова А. А. Определение сложности текстов ЕГЭ по английскому языку при помощи цифровых технологий / А. А. Короткова, О. С. Сафонкина // Нижегородское образование. – 2019. – № 2. – С. 107-111.

5. Мальчевская Е. А. Трансформация жанра рецензии / Е. А. Мальчевская // Веснік Беларускага дзяржаўнага універсітэта. Сер. 4, Філалогія. Журналістыка. Педагогіка. – 2011. – № 1. – С. 74-77.

6. Мацковский М. С. Проблемы читабельности печатного материала / М. С. Мацковский// Смысловое восприятие речевого сообщения в условиях массовой коммуникации. – М. : Наука, 1976. – С. 126-142.

7. Микк Я. А. О факторах понятности учебного текста: автореф. дис. … канд. пед. наук / Я. А. Микк. – Тарту, 1970. – 22 с.

8. Наследов А. Д. IBM SP SS Statistics 20 и AMOS: профессиональный статистический анализ данных / А. Д. Наследов. – СПб. : Питер, 2013. – 416 с.

9. Наследов А. Д. Математические методы психологического исследования. Анализ иинтерпретация данных / А. Д. Наследов. – СПб. : Речь, 2012. – 392 с.

10. Невдах М. М. Разработка метода автоматизированной оценки сложности учебных текстов для высшей школы / М. М. Невдах // Международная научная конференция: «Теория вероятностей, случайные процессы, математическая статистика и приложения». – 2008. – С. 239-243.

11. Осетрова О. Шрифт в рекламном дизайне / О. Осетрова // Коммуникации. – 2005. – № 11.

12. Тертычный А. А. Жанры периодической печати: Учеб. пособие для студентов вузов / А. А. Тертычный. – М. : Аспект Пресс, 2000. – 310 с.

13. Филиппова А. В. Управление качеством учебных материалов на основе анализа трудности понимания учебных текстов: автореф. дис. … канд. техн. наук / А. В. Филиппова. – Уфа, 2010. – 20 с.

14. Фомина В. А. Виды и маркеры интердискурсивности в текстах кинорецензий [Электронный источник] / В. А. Фомина // Вестник Балтийского федерального университета им. И. Канта / Научная электронная библиотека «Киберленинка». – 2011. – Режим доступа: http://cyberleninka.ru/article/n/kinoretsenziya-v-sisteme-diskursnyh-vzaimodeystviy. – Загл. с экрана.

15. Chall J. S. Readability revisited: The New Dale-Chall Readability Formula / J. S. Chall, E. Dale. – Cambridge, Mass. : Brookline Books, 1995. – 149 p.

16. Dale E., Chall J. S. A formula for predicting readability / E. Dale, J. S. Chall // Educational Research Bulletin. – 1948. – № 27 (2). – P. 11-54.

17. A handbook for writers and editors / G. Hargis, A. K. Hernandez, P. Hughes, J. Ramaker, S. Rouiller, E. Wilde. – Upper Saddle River, NJ: Prentice Hall, 1998.

18. Dubay W. H. The Principles of Readability / W. H. Dubay. – Costa Mesa: Impact Information, 2004. – 72 p.

19. Flesch R. A new readability yardstick / R. Flesch // Journal of Applied Psychology. – 1948. – № 32 (3). – P. 221-233.

20. Gamble L. G. Ease of Comprehension of Standard and Readable Insurance Policies as a Function of Reading Ability / L. G. Gamble, J. P. Kincaid // Journal of Reading Behavior. – 1977. – № 1. – P. 87-95.

21. Harkova E. V. Unified (Russian) state exam in English: Reading comprehension tasks / E. V. Harkova, A. S. Kiselnikov, M. I. Solnyshkina // English Language Teaching. – 2014. – № 12. – P. 1-11.

22. Klare G. R. The measurement of readability / G. R. Klare. – Ames, Iowa: Iowa State University Press, 1963.

23. Long A. Calculating Reading Level / A. Long // Tameri Guide for Writers [Электронный ресурс] Режим доступа: http://www.tameri.com/edit/levels.html. – Загл. с экрана.

24. Mclaughlin G. H. SMOG grading – a new readability formula / G. H. Mclaughlin // Journal of reading. – 1969. – № 12 (8). – P. 639-646.

25. Miles T. H. The fog index: a practical readability scale / T. H. Miles // In Critical Thinking and Writing for Science and Technology. Harcourt Brace Jovanovich. – 1990. – P. 280–284.

26. Sticht T. G. Research towards the design, development and evaluation of a job-functional literacy training program for the US Army / T. G. Sticht //Literacy Discussion. – 1973. – № 4. – P. 339-369.

27. Washburne C. Grade Placement of Children's Books / C. Washburne, M. Vogel // Elementary School Journal. – 1938. – Vol. XXXVII. – P. 335-364.

28. Рецензии пользователей на «Кинопоиске» [Электронный ресурс]. – Режим доступа: https://www.kinopoisk.ru/reviews/. – Загл. с экрана.

29. Частые вопросы о «Кинопоиске» – «Кинопоиск». Справка [Электронный ресурс]. – Режим доступа: https://yandex.ru/support/kinopoisk/index.html. – Загл. с экрана.

30. API для определения читабельности текста [Электронный ресурс]. – Режим доступа: http://api.plainrussian.ru/api/1.0/ru/measure/. – Загл. с экрана.

31. API: документация для определения читабельности текста [Электронный ресурс]. – Режим доступа: https://github.com/ivbeg/readability.io/wiki/API. – Загл. с экрана.

32. Coleman–liau Index [Электронный ресурс] Режим доступа: http://en.wikipedia.org/wiki/Coleman-Liau_Index. – Загл. с экрана.

33. Review. Definition of Review by Merriam-Webster [Электронный ресурс] Режим доступа: http://www.merriam-webster.com/dictionary/review. – Загл. с экрана.

Приложение 1. Листинг программы сбора корпуса

```
import pandas as pd
import requests
from bs4 import BeautifulSoup

final = []  # итоговый список для добавления фрейма
for y in range (1, 500):
    link = 'https://www.kinopoisk.ru/reviews/type/comment/period/month/perpage/200/page/'+str(y)+'/#list/'
    headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Safari/537.36'}
    r = requests.get(link, headers=headers, auth=('ЛОГИН', 'ПАРОЛЬ'))

    soup = BeautifulSoup (r.content, features='html.parser')
    users = soup.find_all('p', {'class':'profile_name'})
    users_nick = []
    for i in users:
        users_nick.append(i.text)
    reviews = soup.find_all('div', {'class':'brand_words'})
    date = soup.find_all('span', {'class':'date'})
    dates = []
    for i in date:
      dates.append(i.text)
    id_users = []
    # достаём ссылки на страницы авторов
    for i in users:
        user = i.a['href']
        id_users.append(user)
```

```
    # оценка читабельности текста отзыва
    r_text = []
    wsyllabes = []
    c_share = []
    chars = []
    avg_slen = []
    spaces = []
    syll = []
    words = []
    letters = []
    sent = []
    cw = []
    simw = []
    avg_syl = []
    grade_smog = []
    grade_ari = []
    index_fk = []
    grade_cl = []
    grade_fk = []
    index_cl = []
    grade_dc = []
    index_dc = []
    index_ari = []
    index_smog = []
    for r in reviews:
        rew = r.text
        r_text.append(rew)
        response = requests.post("http://api.plainrussian.ru/api/1.0/ru/measure/", data={"te
xt":rew})
        readability = response.json()
```

```
    wsyllabes.append(readability['metrics']['wsyllabes'])
    c_share.append(readability['metrics']['c_share'])
    chars.append(readability['metrics']['chars'])
    avg_slen.append(readability['metrics']['avg_slen'])
    spaces.append(readability['metrics']['spaces'])
    syll.append(readability['metrics']['n_syllabes'])
    words.append(readability['metrics']['n_words'])
    letters.append(readability['metrics']['letters'])
    sent.append(readability['metrics']['n_sentences'])
    cw.append(readability['metrics']['n_complex_words'])
    simw.append(readability['metrics']['n_simple_words'])
    avg_syl.append(readability['metrics']['avg_syl'])
    grade_smog.append(readability['indexes']['grade_SMOG'])
    grade_ari.append(readability['indexes']['grade_ari'])
    index_fk.append(readability['indexes']['index_fk'])
    grade_cl.append(readability['indexes']['grade_cl'])
    grade_fk.append(readability['indexes']['grade_fk'])
    index_cl.append(readability['indexes']['index_cl'])
    grade_dc.append(readability['indexes']['grade_dc'])
    index_dc.append(readability['indexes']['index_dc'])
    index_ari.append(readability['indexes']['index_ari'])
    index_smog.append(readability['indexes']['index_SMOG'])


# оценка отзывов
positive = []
negative = []
reviews_rate = soup.find_all("div", {"class": "reviewItem userReview"})
for item in reviews_rate:
    review_id = item["data-id"]
    review_rating = item.find("li", {"id": f"comment_num_vote_{review_id}"})
```

```
        rev_rate = review_rating.text.split(" / ")
        rev_rate_pos = rev_rate [1]
        rev_rate_neg = rev_rate [0]
        positive.append(rev_rate_pos)
        negative.append(rev_rate_neg)

    for a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a18, a19, a
20, a21, a22, a23, a24, a25, a26, a27, a28 in zip(users_nick, dates, id_users, r_text, wsyl
labes, c_share, chars, avg_slen, spaces, syll, words, letters, sent, cw, simw, avg_syl, gra
de_smog, grade_ari, index_fk, grade_cl, grade_fk, index_cl, grade_dc, index_dc, index_
ari, index_smog, positive, negative):
        frame = [a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, a15, a16, a17, a1
8, a19, a20, a21, a22, a23, a24, a25, a26, a27, a28]
        final.append(frame)

df = pd.DataFrame(final, columns=['users_nick', 'dates', 'id_users', 'r_text', 'wsyllabes', 'c
_share', 'chars', 'avg_slen','spaces', 'syll', 'words', 'letters', 'sent', 'complex_w', 'simple_wo
rds', 'avg_syl', 'grade_smog', 'grade_ari', 'index_fk', 'grade_cl', 'grade_fk', 'index_cl', 'gra
de_dc', 'index_dc', 'index_ari', 'index_smog', 'positive', 'negative'])
df.to_csv('Films.csv', sep=';')

from google.colab import files
files.download('Films.csv')
```

Приложение 2. Примеры аномальных рецензий

Пример отзыва с хорошими (низкими) показателями читабельности и плохими (низкими) показателями популярности:

Текст рецензии: «От сериала на Нетфликсе с названием «Люцифер» и с голым мужиком на постере не ждешь ничего, кроме этого голого мужика. Серьезно, никаких избыточных ожиданий: окей, посмотрим, посмеемся, 18+, substances, nudity, поехали дальше. И…

И это как сесть в поезд метро, а вместо этого попасть на американские горки. Это вау. Это просто masterpiece. Начнем от самого простого до самого невероятного:

1. Голос Тома Эллиса. Боже, это не голос, это мёд, это елей, это конец первого сезона — и этот голос озвучивает все английские книги в моей голове. Вот с этим британским акцентом и бархатцей. Вот с этими низкими нотами и игрой слов. Он настолько характерный, настолько яркий, настолько запоминающийся — ух ты, серьезно. После первой же спетой им песни я полезла в Яндекс. музыку слушать все остальные альбомы.

2. Игра. Игра на фортепиано. Игра актерская. Он просто хорош. И не только он. Весь каст — класс. Просто класс. Это первый раз, когда смотришь на вот эту подборку: чтобы разный цвет кожи, разный разрез глаз, чтобы все разные, и… — окей. И классно. И вообще не важно, потому что играют просто чудесно. Ко второму сезону чуть ли не над каждой серией сидишь и плачешь, потому что ох. Ах. Ух. Забавный факт, что у Тома Эллиса отец, сестра и дядя — священники.

3. Сюжет. Бог с ними, с этими murders, но глобальная идея… Это самая лучшая психосказка, которую я когда либо видела. Просто самая лучшая. Это терапия для смотрящего, это философия и это история в одном флаконе. И пьётся этот флакон одним глотком. Это настолько глубоко, настолько прекрасно, настолько правильно, настолько интересный взгляд — нет слов, серьезно. Вау и все тут.

О том, что мы сами решаем, кто мы. О том, что двери в аду всегда открыты и любой может выйти. О том, как важны друзья. О том, как проходить через вину,

как учиться чувствовать, как меняться, как находить ответы на свои вопросы, как… Как вообще жить. С верой и правдой. И как это неоднозначно. И так легко, так ненавязчиво все это — ювелирная работа.

И олд-мобили, и виды с небоскребов, и LA, и… Красиво. Сказка. Чудесная сказка.»

Показатели читабельности:

- index_fk – 1,91;
- index_cl – 2,17;
- index_dc – 3,44;
- index_ari – 2,54;
- index_smog – 3,96.

Показатель популярности – 0. Количество друзей / подписчиков – 6 / 6.

Пример отзыва с плохими (высокими) показателями читабельности и хорошими (высокими) показателями популярности:

Текст рецензии: «Этот фильм может показаться сплошной путаницей, нагромождением сбивчивых мыслей о действительном и мнимом, хаосом панического сознания в котором пытается разобраться героиня картины, психотерапевт, прервав карьеру в Штатах, приезжающая на замену в пасмурную Шотландию, где принимает под опеку пациентов отошедшего от дел доктора, пытается вникнуть в суть их проблем и упирается в глухую стену неразговорчивого мальчика Эммануэля, одержимого рисованием мрачных картинок, сюжеты которых удивительно совпадают с реальностью, включая личную жизнь женщины, не понимающей, как остановить наваждение, ломающее арифметику ума, отдавая его под контроль безумию.

Сюжет усложняется новыми знакомствами страдающего от депрессии врача, проводящего вечера в обнимку с бутылкой крепкого виски, приводящими его в клуб интеллектуалов, погружённых в анализ квантовой механики и мысленных экспериментов Шрёдингера, что превращает картину в совершенную галиматью,

презрев которую, героиня ищет объяснимые истоки поведения мальчика, считая, что, установив причину, можно прекратить вызванные ею последствия.

Потребуется пережить, по крайней мере, одну ложную развязку, чтобы понять, что математическая головоломка была ложным следом, а идея картины имеет концептуальное родство с громоподобным «Шестым чувством» М. Найт Шьямалана, построенном на скрытом изъяне представлений персонажа о своём состоянии, искажающем его взгляд на природу вещей и событий, что становится очевидным на пути к окончательному раскрытию карт, подчинённом внутренней логике возбуждённого сознания субъекта, исключающего альтернативу собственной реальности, но здесь это не производит оглушительного эффекта из-за куда более прозаичного диагноза, определяемого по меняющимся декорациям и статусу действующих лиц, что отчасти снимает претензии к невразумительности их образов, оказывающихся вполне подходящими для созданной сценаристами расщеплённой реальности, где встречаются другой пациент и другой доктор.»

Показатели читабельности:

- index_fk – 31,42;
- index_cl – 18,98;
- index_dc – 25,99;
- index_ari – 30,45;
- index_smog – 29,04.

Показатель популярности – 59,49. Количество друзей / подписчиков – 1487 / 2733.