**Отчет по лабораторной работе №4**

Студента 4 курса ПИ-20-1

Семчук Екатерины

Разработать программу (Python, C# или др.) для оценки тональности отзывов о 5 фильмах, оставленных пользователями на сайтах https://www.imdb.com/ или https://www.kinopoisk.ru/lists/movies/top250/. Выбрать и обосновать инструментальное средство (библиотеку).

Текст отзывов получен с помощью технологии web-скрапинга или Web API.

Выполнить анализ полученных результатов. Оценить зависимость указанного пользователем рейтинга от тональности отзыва. Рассчитать статистики, построить иллюстрирующие графики.

Исходный код разработанного приложения находится в репозитории: <https://github.com/EkatherinaS/Data-Analysis-Technologies-in-Internet.git>

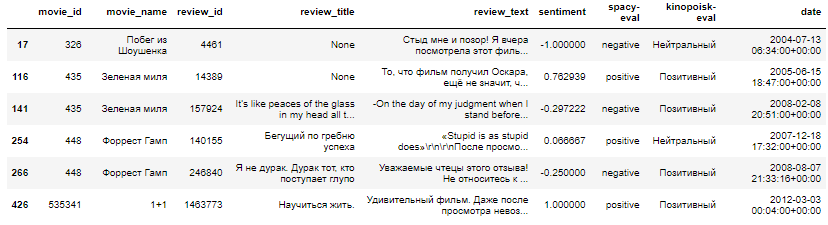
В папке: Lab4

1. **Разработать программу (Python, C# или др.) для оценки тональности отзывов о 5 фильмах, оставленных пользователями на сайтах. Выбрать и обосновать инструментальное средство (библиотеку).**

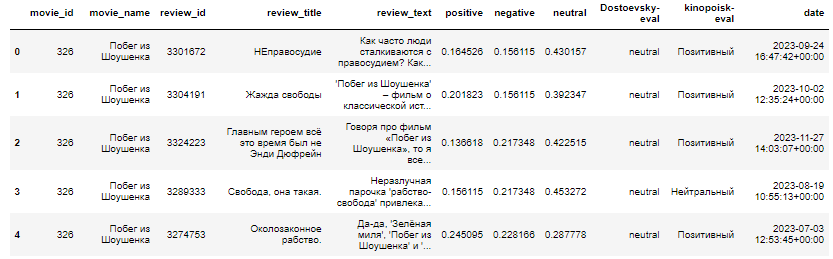
Для работы был выбран сайт «Кинопоиск», в связи с чем было важно, чтобы инструмент имел удобный функционал для работы с русском языком.

Из рассмотренных инструментов, были выбраны два наиболее подходящие к задаче: spaCy и Dostoevsky. Обе библиотеки умеют работать с русском языком и имеют достаточно удобный функционал.

Изначально была выбрана библиотека spaCy, так как ее последнее обновление было в 2023 году. Однако для данного набора отзывов библиотека не представила особенного интереса. Из 500 отзывов spaCy выделил как не нейтральные только 6 наиболее ярких:



Далее работа была продолжена с библиотекой Dostoevsky, преимуществом которой стало выделение отдельных показателей для уровня нейтральности, позитивности и негативности отзывов:



1. **Текст отзывов получен с помощью технологии web-скрапинга или Web API**

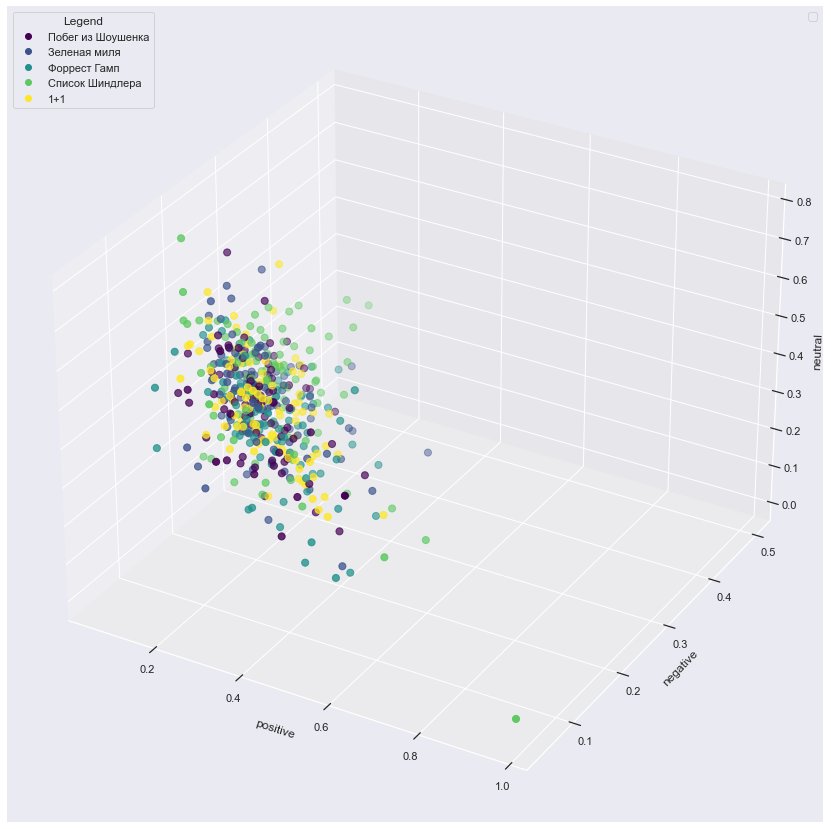
Парсинг сайта был выполнен при помощи инструмента для работы с неофициальным API КиноПоиска:

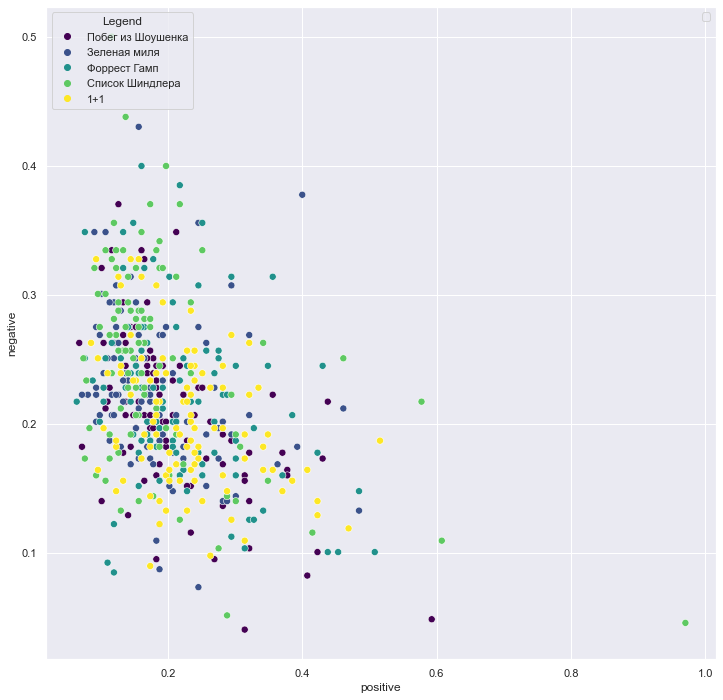
https://github.com/odi1n/kinopoisk\_dev/tree/master

1. **Выполнить анализ полученных результатов. Оценить зависимость указанного пользователем рейтинга от тональности отзыва. Рассчитать статистики, построить иллюстрирующие графики.**

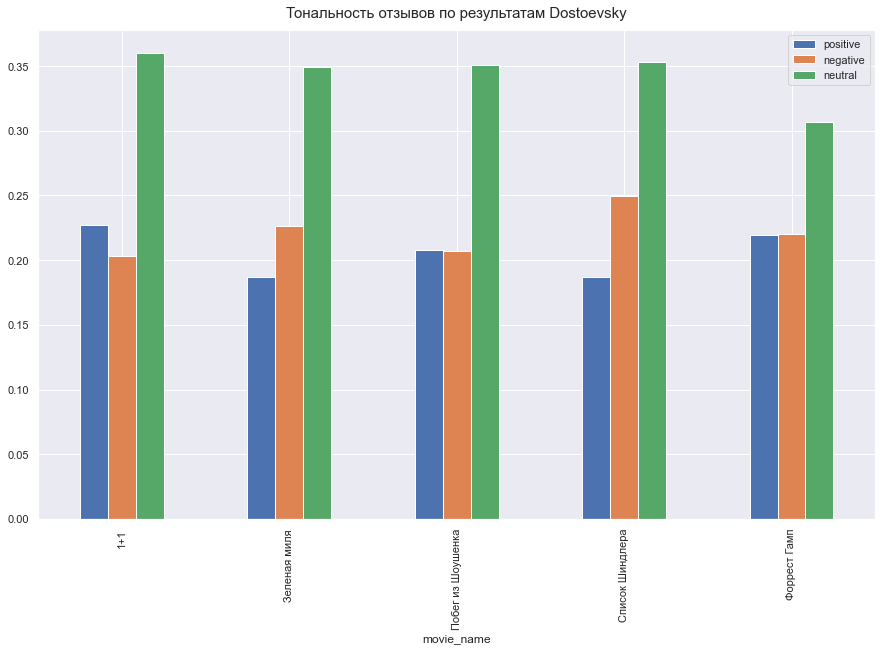
Для анализа результатов было построено несколько графиков.

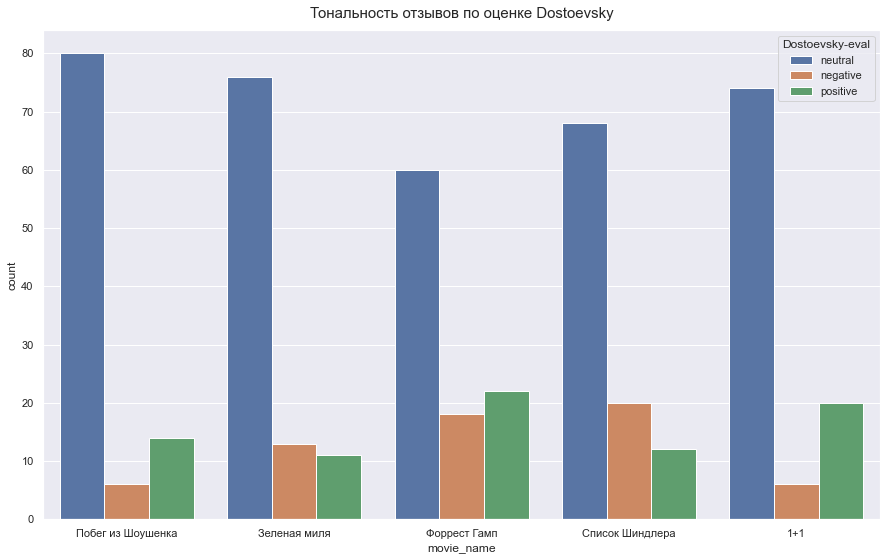
Первая группа графиков - scatterplots, из них можно сделать вывод, что в среднем отзывы на выбранные фильмы пишутся достаточно нейтральные, хотя есть несколько с сильной эмоциональной окраской (о чем говорил и инструмент spaCy).

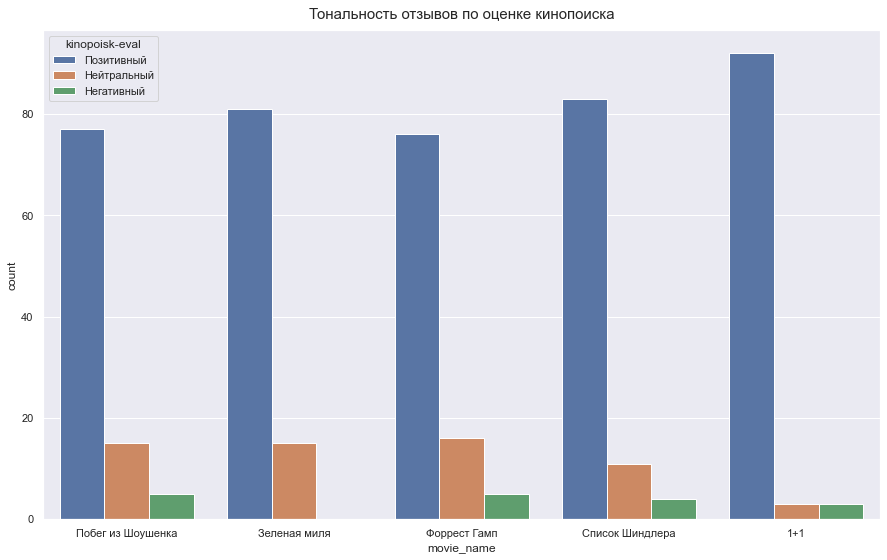




Вторая группа графиков - barplots и countplots.На них наблюдается интересный момент: кинопоиск оценивает большую часть отзывов как позитивные, хотя обе библиотеки считают, что преобладают нейтральные.







Третья группа графиков - тональность отзывов на временной шкале. По ней можно заметить, что для каждого из фильмов было много отзывов в 2008 году (предположительно - момент, когда кинопоиск стал популярен) или в год близко к выходу фильма (для 1+1 - 2012 год). После чего на длительный период отзывы отсутствуют или их очень мало. И следующий пик наступает в 2023 году.

