Code

Итоговый проект

Группа N 23

Предобработка

В процессе текстового анализа будет осуществлен:

> Sentiment анализ (определение эмоциональной окраски отзывов читалетей, а так же описания каждого комикаса для того, чтобы оценить сюжет)

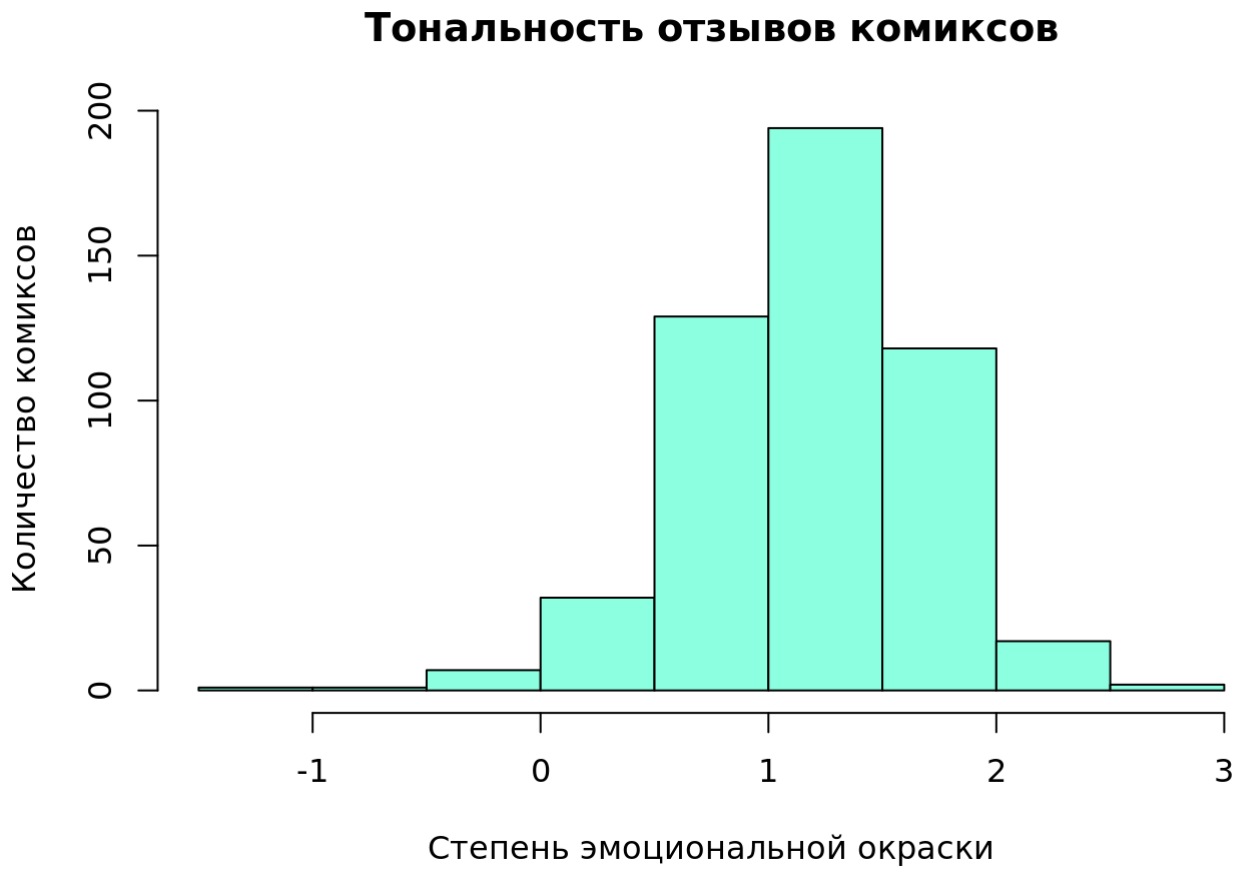
> Тематическое моделирование (разбиение комиксов на группы, сходные по тематике)

Оценка тональности отзывов

Для начала поработаем с отзывами: определим эмоциональную окраску каждого из них.

Для этого нам необходимо разбить каждый отзыв на отдельные слова и с помощью словаря тональности выявить эмоционально окрашенные слова. Тональность отзыва будет представлять среднее значение тональности всех входящих в него слов.

Сразу сформируем новый признак, содержащий среднюю тональность каждого отзыва.

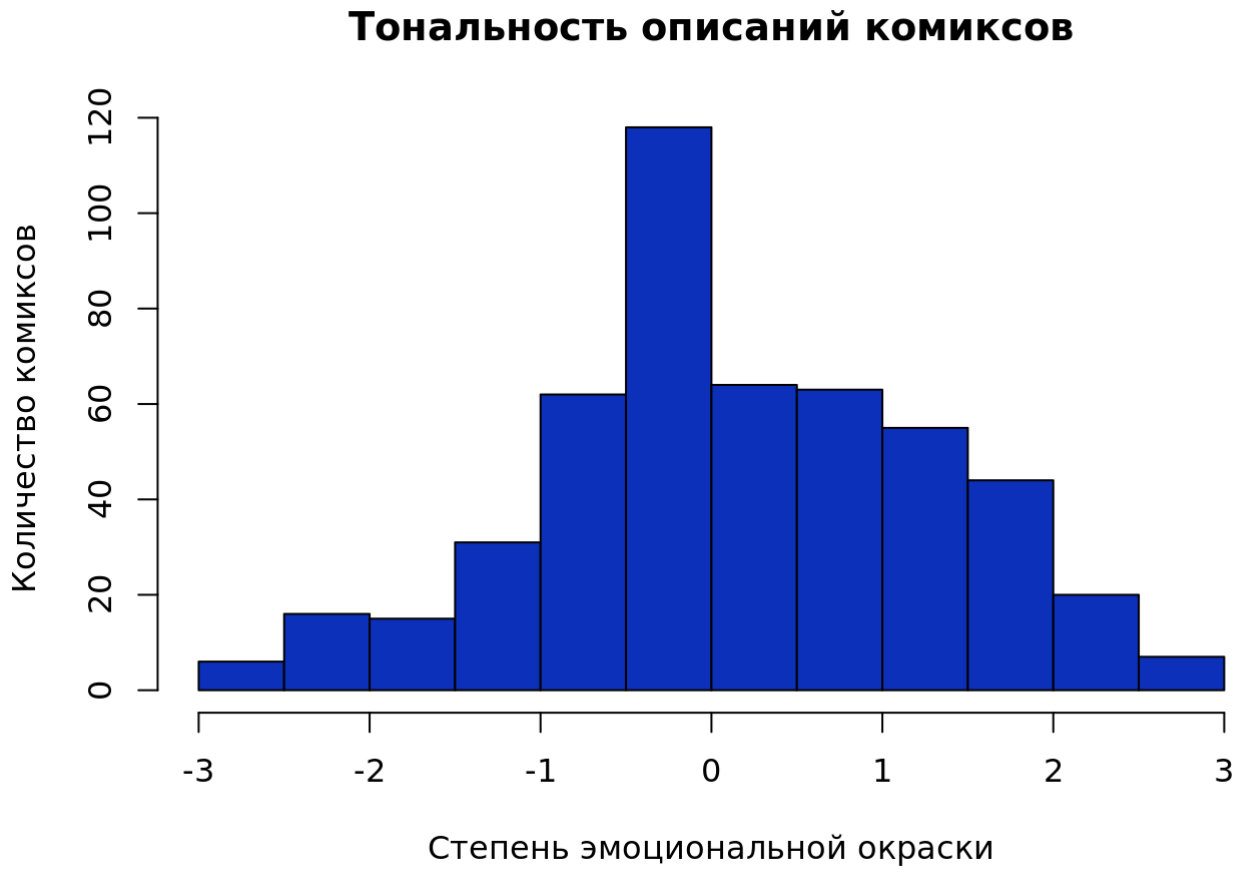


Значения в окрестностях -1 имеют комиксы с самыми негативно окрашенными отзывами, в 3 - самые позитивно окрашенные. Заметим, что у большинства комиксов тональность отзывов находится в районе 1 по эмоциональной окрашенности, что говорит о том, что многие отзывы не сильно эмоционально окрашены, однако больше положительны.

Средняя эмоциональная окраска отзывов для каждого комикса – новый признак, который будет использован в построении рекомендательной системы.

Эмоциональная окраска книг

Раскроем эмоциональную окраску описания самих комиксов (description). Это может помочь нам в оценке тональности сюжета. В результате мы получим новый признак, который добавим к исходному.



Весь диапозон эмоциональной окраски находится в промежутке от -3 до 3, где значения, близкие к 3, имеют наиболее позитивно эмоциональные отзывы, а к -3 – более негативно эмоциональные.

Из графика видно, что тональность большинства отзывов лежит в промежутке от -1 до 0. Это говорит о том, что описание многих из них не сильно эмоционально окрашены и больше склоняются в негативную сторону.

Тематическое моделирование

Разделим все комиксы на кластеры, каждый из который представляет собой группу комиксов относящихся к одной теме. Для этого используем тематическое моделирование, чтобы выявить кластеры (темы).

Количество кластеров - 8 (оптимальное значение исходя из оценки качества).

Приведем значения полок к единому формату (например, “dc-comic” и “dc”). Это снизит количество уникальных значений в данных. Также заменим пустые значения страниц на нули, а остальные данные приведем в числовой формат.

Добавим средние оценки и количество отзывов, что может быть ключевыми фичами для *content-based* модели.

В данных обнаружилось, что существуют отзывы с оценками 0, что не вписывается в нашу оценочную систему. Так как для CB модели наши оценки усредняются, то мы можем удалить такие отзывы без значительного искажения данных.

Content-based рекомендация

Подготовка данных для content-based модели

Content-based модель – это модель, рекомендующая комиксы по их похожести.

Мы должны наилучшим образом создать описание наших объектов, чтобы в дальнейшем похожесть комиксов имела реальный смысл.

Создадим матрицу описания комиксов, где мы будем:

> Нормировать значения непреревных переменных, чтобы они были однородными в пространстве.

> Кодировать категориальные переменные в форме one-hot encoding (создавая новые переменные, содержащие булевы значения, в зависимости от характеристик объектов).

Используемые признаки:

1. Количество страниц
2. Издатель
3. Количество оценок
4. Тональность отзывов
5. Тональность описания
6. Тема описания из тематического моделирования
7. Средний рейтинг

Создаем матрицу расстояния описания комиксов: дистанцию будем считать через косинусное расстояние.

Базовая content-based модель в виде функции

Content-based модель берет юзерские оценки 4/5 и выдает комиксы, похожие на положительно оцененные. Для юзеров, у которых нет оценок 4/5 мы предлагаем наиболее популярные комиксы. Популярность – средний рейтинг, фильтрованный по количеству отзывов (больше 25 квантиля).

Если юзер – новый пользователь, мы предлагаем ему ввести понравившийся комикс, по которому рекомендуем похожие. Если юзер вводит “Nope” (он не читал комиксы), то мы рекомендуем ему самые популярные.

В конце раздела будет представлена полноценная Content-based модель, но из-за формы предоставления отчета эта система не работает (она ожидает от юзера инпут, но его нет). Поэтому была создана базовая модель, которая имеет тот же функционал, что и полная версия, но может быть адаптирвоана под форму отчета.

Полноценная Content-based модель представлена в виде кода, но в отчете отображаться не будет.

Оценка рекомендации:

Теперь оценим наши рекомендации по такой логике:

1. Оцениваем только людей, последняя оценки которых это 4/5
2. Предполагаем, что эта последняя оценка и есть идеальный предикшен – скрываем ее
3. Делаем рекомендацию на остальных данных
4. Находим манхэттэнское расстояние между предсказанным комиксом и реальным (манхэттэнское расстояние здесь – аналог MAE).

В результате получаем расстояние предсказания до идеального предикшена. Так как это значение сложно интерпретировать, мы можем сделать константный предикшн популярными комиксами и сравнить, на сколько процентов мы улучшили результат, используя Content-Based модель.

Проверка для двух пользователей:

```
## MAE: [1] 13.60755
## Насколько лучше справились относительно константной модели (в %): [1] 21.53074
```

```
## MAE: [1] 15.31654
## Насколько лучше справились относительно константной модели (в %): [1] 15.30805
```

Выше можно увидеть, что наша формальная оценка действительно работает: она показывает, что модель неплохо справляется с задачей, что обосновывается лучшим качеством по сравнению с константной моделью популярных комиксов.

Коллаборативная фильтрация

За основу возьмем начальные датасеты, чтобы никакие изменения до этого не испортили данных по оценкам, ведь метод коллаборативной фильтрации строится именно на них.

Прежде всего, избавимся от нулевых оценок, которые могут олицетворять как хороший отзыв, так и плохой. А в данном случае это может навредить рекомендательной системе (хоть это и всего 2% имеющихся оценок).

Далее посчитаем количество оценок от одного пользователя и оценок по одному комиксу, чтобы отфильтровать данные, дабы они не были сильно смещены относительно друг друга.

Преобразуем в матрицу и отфильтруем таким образом, чтобы количество оценок, поставленных одним пользователем было больше 5 (при этом количество оценок одного комикса от 8 изначально). Число обосновано тем, что количество оценок одного пользователя варьируется от 1 до 126, причем почти половина пользователей оценили менее 7 комиксов. При параметре количество > 5 мы можем дать рекомендацию 357 пользователям из 501 доступных (с учетом нулей) в датасете.

Данные разделим на тестовую и обучающую выборки. На обучающей построим модель, для пользователей из тестовой будем рекомендовать комиксы.

Получить рекомендацию методом колаборативной фильтрации можно двумя способами: **IBCF** (Recommender based on item-based collaborative filtering) и **UBCF** (Recommender based on user-based collaborative filtering). Лучший метод мы определим с помощью оценки MAE.

Так или иначе, модель дает пользователю рекомендацию в виде 5 комиксов. В случае, если рекомендация не будет выдана, мы понимаем, что данных этого пользователя не хватает для составлений рекомендаций (аналогично, если это новый пользователь). Тогда ему предлагается список 10 самых популярных комиксов сайта, где популярность - это средний рейтинг, фильтрованный по количеству отзывов (больше 25 квантиля).

В результате полученные комиксы в рекомендации располагаются в порядке уменьшения средней оценки.

Оценка рекомендации:

Для правильной оценки разделим данные на тестовую и обучающую выборки.

Проверим оценку для IBCF метода:

```
##      RMSE       MSE       MAE
## 1.3002848 1.6907407 0.9791023
```

И то же самое для метода UBCF:

```
##      RMSE      MSE      MAE
## 1.227946 1.507851 0.950930
```

UBCF рекомендует лучше, так как значение MAE ниже. Поэтому в дальнейшем будем использовать метод UBCF.

Метод UBCF в виде функции

Для работы функции пользователю необходимо знать свой ID. Тогда он получит рекомендацию из 5 комиксов.

Если пользоватеть новый или сам оценил менее 5 комиксов, ему будет предложена рекомендация из 10 наиболее популярных комиксов.

Примеры

Примеры content-based

1. Система задает вопрос: новый ли юзер. Юзер отвечает, что он новый. Система просит вписать название любимого комикса или “Nope”, если такого нет. Юзер вводит “Nope”.

Ожидаем увидеть самые популярные комиксы.

```
## [1] "Try out what's popular among users!"
```

title <chr>
Green Arrow, Volume 2: Triple Threat
Green Arrow, Volume 3: Harrow
Legion of Super-Heroes, Vol. 1: Hostile World
Dark Empire I
Daredevil: Season One

5 rows

Действительно, получаем самые популярные комиксы.

2. Система задает вопрос: новый ли юзер. Юзер отвечает, что он новый. Система просит вписать название любимого комикса или “Nope”, если такого нет. Юзер вводит название любимого комикса.

Например, введенный комикс: “Batman Incorporated, Volume 1: Demon Star”. Ожидаем увидеть комиксы вселенной Бэтмэна/издателя DC.

title <chr>
Birds of Prey, Volume 1: Trouble in Mind
Justice League, Volume 5: Forever Heroes
Plutona
All-Star Batman, Volume 1: My Own Worst Enemy

title
<chr>
Justice League, Volume 1: Origin

5 rows

Действительно, получаем комиксы DC/про бетмена.

3. Система задает вопрос: новый ли юзер. Юзер отвечает, что он НЕ новый. Система просит вписать user_id. Юзер вводит свой id.

Например, введенный id: “ff05755454e5c477c5cc79011c8ada6f”. Так как уюзера есть комикс с оценкой 5, ожидаем увидеть комиксы издателя Марвел про Мстителей.

title
<chr>
Invincible: Ultimate Collection, Vol. 1
Queen and Country: The Definitive Edition, Vol. 1
DC Comics: The New 52
The Animal Man Omnibus
Spider-Man: Spider-Island

5 rows

Действительно, получаем Марвел комиксы про супергероев.

4. Система задает вопрос: новый ли юзер. Юзер отвечает, что он НЕ новый. Система просит вписать user_id. Юзер вводит свой id.

Например, введенный id: “003f7ff55fde1b9717dc1f90bd47cb1e”. Так как у юзера нет комиксов с оценкой 4/5, ожидаем увидеть рекомендацию популярных комиксов.

title
<chr>
Green Arrow, Volume 2: Triple Threat
Green Arrow, Volume 3: Harrow
Legion of Super-Heroes, Vol. 1: Hostile World
Dark Empire I
Daredevil: Season One

5 rows

Действительно, получаем самые популярные комиксы.

Примеры collaborative filtering

1. Пользователь “90ee0eac78765a906c34e63d0e080a3f” оценил данные комиксы:

title	rating
<chr>	<dbl>

title <chr>	rating <dbl>
Wolverine and the X-Men, Volume 1	5
Daredevil, Volume 3	5
Thor: God of Thunder, Volume 2: Godbomb	5
New Avengers, Volume 2: Infinity	5
Uncanny X-Men: Lovelorn	4
Justice League, Vol. 1: Origin	4
Indestructible Hulk, Volume 1: Agent of S.H.I.E.L.D.	4
Star Wars: Purge	4
Cable and X-Force, Volume 2: Dead or Alive	4
Five Ghosts, Volume One: The Haunting of Fabian Gray (Five Ghosts, #1)	4

1-10 of 20 rows Previous **1** 2 Next

Так как оценено 20 комиксов, этих данных хватает для рекомендации:

title <chr>	average_rating <chr>	pul <ch
The Absolute Sandman, Volume Two	4.69	
Locke & Key, Vol. 5: Clockworks	4.50	IDW
Fables (The Deluxe Edition, #3)	4.42	Ver
Swamp Thing, Vol. 3: The Curse	4.34	Ver
Buffy the Vampire Slayer: Wolves at the Gate (Season 8, #3)	4.05	Da

5 rows

В начальных данных и предложенной рекомендации присутсвуют одинаковые издатели и тематики. Будем считать, что рекомендация адекватна.

2. Пользователь “6baf45d03466a5858403d892286ff222” оценил комиксы:

title <chr>	rating <dbl>	publisher <chr>
V for Vendetta	4	Vertigo

1 row

Так как оценен всего 1 комикс, то данных слишком мало для рекомендации. Пользователю будут предложены самые популярные комиксы:

title <chr>
March: Book Three (March, #3)

title
<chr>
The Absolute Sandman, Volume Two
Saga: Book Two
Skip Beat!, Vol. 26
Fullmetal Alchemist, Vol. 14 (Fullmetal Alchemist, #14)
The Sandman, Vol. 9: The Kindly Ones (The Sandman, #9)
The Complete Peanuts, Vol. 6: 1961-1962
Skip Beat!, Vol. 17
Fullmetal Alchemist, Vol. 7 (Fullmetal Alchemist, #7)
Daredevil by Brian Michael Bendis & Alex Maleev Ultimate Collection, Book 2

1-10 of 10 rows | 1-2 of 3 columns

3. Пользователь с ID “85af94303466a5abc403d811286ff111” - новый, оцененных комиксов нет. Поэтому ему также будут предложены самые популярные комиксы:

title
<chr>
March: Book Three (March, #3)
The Absolute Sandman, Volume Two
Saga: Book Two
Skip Beat!, Vol. 26
Fullmetal Alchemist, Vol. 14 (Fullmetal Alchemist, #14)
The Sandman, Vol. 9: The Kindly Ones (The Sandman, #9)
The Complete Peanuts, Vol. 6: 1961-1962
Skip Beat!, Vol. 17
Fullmetal Alchemist, Vol. 7 (Fullmetal Alchemist, #7)
Daredevil by Brian Michael Bendis & Alex Maleev Ultimate Collection, Book 2

1-10 of 10 rows | 1-2 of 3 columns

Примеры peer review

Вопрос: Если бы был пользователь, которому нравятся комиксы с более чем 300-ми страницами, был бы ему рекомендован комикс паблишера Image Comics?

title
<chr>
Injustice: Gods Among Us: Year Three, Vol. 2
I, Vampire, Vol. 2: Rise of the Vampires

title
<chr>
New Avengers, Volume 2: Infinity
Moon Knight, Volume 1: Lunatic
The Vision, Volume 1: Little Worse Than A Man

5 rows

Ответ: Нет. Логичный ответ, у Image Comics мало комиксов с кол-вом страниц 300+

Вопрос: Если бы был пользователь, которому нравятся комиксы с высокой средней тональностью отзывов, то ему рекомендовались бы комиксы с высокими средними оценками?

title
<chr>
Samurai Jack, Vol. 1
American Vampire, Vol. 2
Anita Blake, Vampire Hunter: The First Death
Mockingbird, Vol. 1: I Can Explain
The Strain, Volume 1
The X-Files: Season 10, Volume 1

6 rows

Ответ: Нет.

Вопрос: “Ввожу ID ffecee234f84555b8598155449e0cdf4 - оценил единственный комикс”A User’s Guide to Neglectful Parenting" на 4. Ожидаю увидеть “Gotham Central, Vol. 2: Half a Life”.

title
<chr>
Injection, Vol. 1 (Injection, #1)
American Vampire, Vol. 2
Will O' the Wisp
Death Note, Vol. 1: Boredom (Death Note, #1)
Angel: After the Fall, Volume 1

5 rows

Ответ: Конкретно “Gotham Central, Vol. 2: Half a Life” выведен не был, так как его в наших данных нет, но “Injustice: Gods Among Us: Year Two, Vol. 2” - похож на него.

Вопрос: Мне интересно, порекомендовали бы пользователю, которому нравится вселенная DC (что часто отражает publisher и что включено в рекомендательную систему content based), комиксы из той же вселенной.

title
<chr>

title
<chr>
Injustice: Gods Among Us: Year Three, Vol. 2
I, Vampire, Vol. 2: Rise of the Vampires
New Avengers, Volume 2: Infinity
Moon Knight, Volume 1: Lunatic
The Vision, Volume 1: Little Worse Than A Man

5 rows

Ответ: Да, 4/5 рекомендованных комиксов - вселенная DC.

Вопрос: Если бы был пользователь, который смотрим Marvel, что бы еще ему порекомендовало?

title
<chr>
Invincible: Ultimate Collection, Vol. 1
Queen and Country: The Definitive Edition, Vol. 1
DC Comics: The New 52
The Animal Man Omnibus
Spider-Man: Spider-Island

5 rows

Ответ: Да, пользователю часто приходит рекомендация комиксов Marvel.

Вопрос: Мне бы хотелось проверить рекомендательную систему основанную на Content Based. Поскольку я новый пользователь и меня нет в системе, но при этом я большой любитель комиксов про Россомаху (wolverine). Поэтому мне бы было интересно, какие комиксы мне выдала система, если бы я вбил комикс под названием: “Wolverine: X weapon”.

title
<chr>
X-23, Vol. 1: The Killing Dream
Justice League Dark, Volume 2: The Books of Magic
Anita Blake, Vampire Hunter: The First Death
Wolverine and the X-Men, Volume 2
Wolverine and the X-Men, Volume 1

5 rows

Ответ: Комикса “Wolverine: X weapon” в наших данных не было, но есть комикс “Wolverine: Origin”. Введя его мы получаем комиксы про Россомаху.

Вопрос: Если меня нет в системе и я напишу, что мне нравится атака титанов, то что мне порекомендуют?

title <chr>
Batman vs. Superman: The Greatest Battles
Cowboy Ninja Viking Volume 1 (Cowboy Ninja Viking, #1)
Flop to the Top!
Black Metal: Volume 1
The Adventures of Tintin, Vol. 1: Tintin in America / Cigars of the Pharaoh / The Blue Lotus

5 rows

Ответ: был введен комикс “Attack on Titan Anthology”. Атака Титанов рекомендована не была. Это произошло из-за того, что в наших данных только 3 комикса этой вселенной, и все они сильно отличаются друг от друга.

Вопрос: Если бы был пользователь, которому нравится комикс “Deadpool, by Daniel Way: The Complete Collection, Volume 3”, были бы ему рекомендованы комиксы “Black Panther: World of Wakanda (2016-) #1”, “Hulk, Volume 1: Banner DOA”, “Doctor Strange, Vol. 3: Blood in the Aether”?

Ответ: Комикса “Deadpool, by Daniel Way: The Complete Collection, Volume 3” в данных нет.

Выводы

Текстовый анализ:

В результате проведения текстового анализа были определены эмоциональная окраска отзывов пользователей и тональность общего описания сюжета каждого комикса. Также комиксы были разделены на кластеры, каждый из которых отражает принадлежность к той или иной теме.

Все новые переменные были использованы в дальнейшем построении рекомендательных систем.

Также в процессе анализа был сделан вывод о том, что большинство комиксов, а так же отзывов имеют относительно нейтральную эмоциональную окраску.

Content-based:

На основе примеров можно сказать, что модель работает неплохо и логично. Она действительно предлагает пользователю те объекты, которые похожи на его предпочтения. Также можно увидеть, что и формально модель прошла проверку: она стабильно предсказывает лучше, чем константная модель.

Тем не менее, проблема модели в том, что для конкретного предсказания она не использует информацию об остальных оценках пользователя. Это может исправить симбиоз модели CF и Content-Based, когда юзеру с достаточным количеством оценок система предсказывает комиксы с помощью CF.

Также проблема нынешней Content-Based модели заключается в создании пузыря для пользователя (он может быть окружен однотипными рекомендациями и не сможет открыть для себя что-то новое). Однако данная проблема не может решаться в статике, так как мы не можем реагировать на действия пользователя.

Коллаборативная фильтрация:

На примерах видно, что модель работает исправно. Пользователь получает рекомендацию из комиксов, близких к уже оцененным или же самые популярные комиксы в случае нехватки данных.

Минус модели в том, что предпочтения выдаются исключительно на основе уже проставленных оценок. Чем меньше таких оценок (напомню, что мы фильтровали таким образом, чтобы их было больше 5), тем, вероятно, хуже может быть составлена рекомендация. Однако, это зависит в том числе и от пользователя: он мог оценить как комиксы из различных категорий с разными оценками, так и комиксы только одной категории, при этом рекомендации могут совпадать.

Таким образом, можно сказать, что на данный момент для нынешнего проекта с имеющейся базой данных обе модели довольно успешно выполняют свои функции.

Ответы на вопросы

1. **Вопрос:** Может, лучше пояснить ваши метрики (one hot например) для тех, кто не знает

Ответ: One-hot – способ кодирования категориальной переменной в булевой форме

2. **Вопрос:** Почему решили использовать мэнхэттенское расстояние?

Ответ: Манхэттенское расстояние – аналог MAE в пространстве, хорошо интерпретируется

3. **Вопрос:** Почему даете рекомендации пользователям, оценившим комиксы на 4 (ведь 4 это не высшая оценка, т.е. комикс понравился не идеально)?

Ответ: 4 как оценка для рекомендации позволяет сделать рекомендацию более персонализированной, не дожидаясь, когда пользователь сам найдет то, что ему идеально нравится

4. **Вопрос:** Не совсем понятно, как в контент-бэйсд пользователь должен ввести какой-то комикс, если он только зашел в систему и не знает названий комиксов

Ответ: Пофиксили

5. **Вопрос:** Зачем вводить айди пользователя, если он понимает, что его нет в системе (возможно лучше сразу ввести желаемый комикс)

Ответ: Пофиксили

6. **Вопрос:** Также, не очень понятно, какие комиксы рекомендуются новым пользователям. Участники проекта сказали, что выдается список самых популярных комиксов, однако возникает вопрос, что значит популярный комикс (большое кол-во отзывов или высокие оценки)? "

Ответ: В системе content-based изначально была использована синтетическая переменная, выражающая популярность (кол-во отзывов * средняя оценка). Однако в результате проверки качества такая оценка оказалась хуже, чем простое предложение самых высокорейтинговых комиксов, которые имею достаточно большое количество оценок (> 0.25 квантиль). В системе

7. **Вопрос:** Почему так специфически была произведена оценка (оценка ТОЛЬКО пользователей, у которых несколько оценок и последняя ОБЯЗАТЕЛЬНО 4 или 5?

Ответ: Оценка пользователей с обязательной последней (одной из последних) 4/5 сделана из логики: Предположим, эта последняя оценка и есть идеальный предикшен, так как комикс человеку понравился. Мы можем скрыть эту оценку, сделать предикт на остальных данных и понять, насколько в среднем наш предикт отличается от идеального (от реальной оценки)

8. **Вопрос:** Участники проекта сказали, что выдается список самых популярных комиксов, однако возникает вопрос, что значит популярный комикс?

Ответ: Популярность комикса определяется средним рейтингом, который фильтрованный по количеству отзывов (больше 25 квантиля).

9. **Вопрос:** Если человек введет только начало названия комикса (орфографически корректно), но не укажет, например, часть, что выдаст система? Технически человек все сделал правильно, но система может сказать что комикса нет.

Ответ: Система выдаст ошибку в названии и предложит наиболее популярные комиксы. Это происходит из-за того, что юзер вводит тот комикс, которому он бы поставил 5.

10. **Вопрос:** Как вы оценивали качество тематической модели? Насколько содержательными получились темы?

Ответ: Разбиение на топики было оценено с помощью рассмотрения топовых по частоте использования слов в каждой теме. Изначально вручную были заданы рамки разбиения от 3 до 10 тем, чтобы разбиение было и содержательным, но в то же время не сформировалось распыление по темам. Далее методом перебора было выявлено наиболее оптимальное количество топиков, при котором топовые слова в каждый теме были хорошо отличимы дргу от друга.