цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Выявление элементов культурного кода россиян в отзывах о произведениях художественной литературы»

Введение

Одной из задач Вятского государственного университета в рамках программы «Приоритет-2030» является изучение культурного кода россиян. Культурный код – это коллективное бессознательное, которое явно не осознается, но предопределяет наши поступки. Существует гипотеза ученых, что элементы культурного кода можно найти в отзывах о литературных произведениях. Представленный вам датасет содержит отзывы о произведениях художественной литературы, получивших признание в виде российских литературных премий за 2000-2021 гг., размещенных на сайте «Литрес».

Условие задачи

Цель — разработать модель, выявляющую отзывы, в которых содержатся элементы культурного кода. Отзыв считается релевантным, если он обладает ценностно-смысловым содержанием, и его автор вписывает эти смыслы и ценности в свой жизненный и культурный контекст, идентифицирует себя с героями, содержанием, исторической судьбой народа.

Описание входных значений

- train.csv — файл содержащий данные с отзывами для обучения
- test.csv — файл с отзыва, для предсказания
- submission.csv — пример файла для отправки.

Пояснение к столбцам:

- RecordNo — уникальный идентификатор отзыва
- Название книги — название книги, на которую оставлен отзыв
- Автор — автор книги
- Ссылка на литрес — ссылка на книгу
- Рейтинг — общий рейтинг книги, рассчитанный на основе всех отзывов
- Количество оценок — суммарное число оценок книги в обзоре
- Количество отзывов — суммарное число письменных обзоров книги

- Имя читателя — имя пользователя, оставившего отзыв
- Оценка книги читателем (из 5 баллов) — оценка книги конкретным пользователем
- Отзыв — текст отзыва
- Лайки на отзыв — количество положительных оценок отзыва, оставленных другими пользователями
- Дислайки на отзыв — количество отрицательных оценок отзыва, оставленных другими пользователями

Предсказываемые значения:

- Релевантность — характеристика, отражающая степень соответствия контента теме исследования (Нерелевантно - 0, Релевантно - 1)
- Ценности — наличие необходимых ценностей в тексте (Нерелевантно - 0, Релевантно - 1)
- Таксономия релевантные — наличие в тексте отзыва слов, словосочетаний, связей между словами из словаря синонимических рядов каждой категории, обозначающих духовно-нравственную ценность. (не определено - 0, определено - 1)
- Таксономия нерелевантные — наличие в тексте отзыва слов, словосочетаний, связей между словами из словаря синонимических рядов каждой категории, обозначающих духовно-нравственную ценность, не подходящую к теме исследования. (не определено - 0, определено - 1)
- Длина отзыва — наличие у отзыва оптимального числа слов (не определено - 0, определено - 1)

На что стоит обратить внимание

Пояснение к «ценностям»

В Указе Президента РФ от 2 июля 2021 г. № 400 «О Стратегии национальной безопасности Российской Федерации», где к традиционным российским духовно-нравственным ценностям относятся «жизнь, достоинство, права и свободы человека, патриотизм, гражданственность, служение Отечеству и ответственность за его судьбу, высокие нравственные идеалы, крепкая семья, созидательный труд, приоритет духовного над материальным, гуманизм, милосердие, справедливость, коллективизм, взаимопомощь и взаимоуважение, историческая память и преемственность поколений, единство народов России» (IV, ст. 91) .

Использование этих понятий в качестве ключевых позволило выявить их в читательских отзывах. Для расширения зоны смыслового поиска к каждому понятию был подобран синонимический ряд, содержащий от 2-х до 22-х близких по смыслу слов.

Примеры таких синонимических рядов:

Милосердие – добросердечие, доброта, сочувствие, благожелательность, благость, благоволение, доброжелательность, благосклонность, любезность, великодушие, жалость, сострадание, ласка, любовь, снисхождение, добро, сердоболие, умение прощать.

Справедливость – праведность, честность, беспристрастность, беспристрастие, заслуженность, нелицеприятность, непредубежденность, непредвзятость, объективность, правда, правильность, верность, законность, достоверность, истинность, правосудие.

Единство (народов России) – целостность, цельность, монолитность, целость, общность, соборность, идентичность, слитность, тождественность, сплоченность, солидарность, единомыслие, единогласие, спаянность, гармония, гармоничность, мир, равенство, тождество, согласие, сходство.

Метрика

В качестве метрики выступает Recall по пяти столбцам:

$$Result = 0.2 * Recall_{\text{Релевантность}} + 0.2 * Recall_{\text{Ценности}} + 0.2 * Recall_{\text{Таксономия релевантные}} + 0.2 * Recall_{\text{Таксономия нерелевантные}} + 0.2 * Recall_{\text{Длина отзыва}}$$

Recall считается как:

$$recall = \frac{TP}{TP + FN}$$

где TP (True Positive) - количество верно угаданных значений одного класса
FN - False Negative - количество правильно угаданных значений класса

Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на региональном чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, в праве назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявление плагиата кода или несоблюдение Положения конкурса.

6. Участник, получивший 2 дисквалификации за сезон конкурса, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.