Санкт-Петербургский государственный университет
Факультет прикладной математики - процессов управления

Проект по курсу "Искусственный интеллект"
Предсказание жанра фильма по диалогу (Multilabel classification)

Ермоленко Александр
19.Б05-пу
Last update: 30 ноября 2020 г.

Санкт-Петербург 2020

Оглавление

1. Постановка задачи

Для индивидуального проекта по курсу была выбрана следующая задача: необходимо предсказать жанр фильма по диалогу из него. Всего в нашем распоряжении находится 20 жанров. Каждый фильм может иметь несколько жанров (Multilabel classification). В данных возможно встретить несколько разных диалогов из одного и того же фильма. Названия фильмов, год выпуска, режиссер и актеры для классификации не доступны, так что в рамках задания будет произведена работа только с текстами.

1.1. Описание данных

В качестве данных используются датасеты с соревнования Kaggle [1].
В работе используются два набора данных:

- train.csv – тренировочный датасет
- test.csv – тестовый датасет

Откроем тренировочное множество и посмотрим, что оно из себя представляет:

- id – индивидуальный идентификаторы для каждого диалога
- movie – идентификатор фильма
- dialogue – представленный диалог из фильма
- genres – размеченные жанры для каждого диалога

	id	movie	dialogue	genres
0	0	0	I thought you were in a meeting--? I am. ...	[u'drama', u'romance']
1	1	1	Are you sure you're okay? You're pale. I...	[u'drama']
2	2	2	Go on! Get out! Mom look don't say anythi...	[u'comedy']
3	3	3	I could have lost my fucking hands. That ...	[u'mystery', u'thriller']
4	4	4	Stick with me on this Gloria. I need you... <...	[u'crime', u'thriller']
...	...	...	...	...
36986	36986	246	There's a man downstairs. He brought us eggs....	[u'drama', u'war']
36987	36987	43	Hi. I'd prefer it if you didn't speak to ...	[u'comedy', u'drama']
36988	36988	459	I tried to call you I'm running a little late ...	[u'drama']
36989	36989	174	What are you crazy? I just thought we sho...	[u'drama', u'romance']
36990	36990	255	I wouldn't have uh killed you Father. Dominus...	[u'crime', u'drama']

Рис. 1. train.csv

1.2. Цель

Так как данные были импортированы с соревнования на Kaggle, то в качестве метрики выберем ту же, по которой оценивались работы участников. В данном случае такой была выбрана f1-score. По той же логике поставим себе цель по качеству, которое хотим получить по выполнении задачи предсказания: открыв leaderboard, увидим score первого места: 0.65816.
Итого, для оценки своей деятельности были выбраны метрика f1-score и значение по ней в районе 0.65816.

2. Подход к выполнению задачи

После ознакомления с данными и постановки цели необходимо определиться с тем, с помощью каких методов можно реализовать выбранную задачу.
Так как исходные данные у нас это текст – диалоги из фильмов, то очевидно, что для обучения любой модели сначала будет необходимо преобразовать текст в численный вид. После изучения литературы

на эту тему оптимальным вариантом оказался TF-IDF [2].
Для классификации диалогов использовалась логистическая регрессия [3].

3. Ход работы

Достаточно подробно процесс был описан в ноутбуке [4].

4. Итог

Для решения поставленной задачи прогнозирования жанра фильма по диалогу из него мною была поставлена задача подобраться к первому месту в соревновании на Kaggle, откуда и были взяты тестовые данные, и достичь оценки с помощью метрики f1-score в районе 0.65816. Также в ходе изучения подходов к решению подобных задач был установлен следующий порядок действий: векторизовать диалоги с помощью TF-IDF, вектор ответов привести к виду multilabel [6] и классифицировать диалоги с помощью логистической регрессии.
С помощью приобретенных знаний удалось достичь отметки в 0.60053, что не является идеальным результатом, но в рамках подобного соревнования уверенно закрепило бы в верхней части таблицы.

Литература

[1] ДАННЫЕ. *Тестовый и тренировочный датасеты.*
URL: https://www.kaggle.com/c/made-hw-2/data

[2] TF-IDF. *Документация по векторизации текста с помощью библиотеки sklearn.*
URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[3] LogisticRegression. *Документация по логистической регрессии.*
URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] Итоговый код. *Файл ipynb с реализацией проекта.*
URL: https://github.com/Xelanid/Artificial_intelligence/blob/master/movie_genres_classification.ipynb

[5] Репозиторий проекта.
URL: https://github.com/Xelanid/Artificial_intelligence

[6] MultiLabelBinarizer *Документация.*
URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html