Регрессия

Беляков Дмитрий

April 2, 2022

1 Описание задачи

Используем датасет YoutubeVideoDataset с базы данных Kaggle. Датасет содержит названия видео, его описание и одну из 6 категорий к которым это видео относится. Возможные категории: Путешествия, наука и технологии, еда, исскуство и музыка, производство. Будем пытаться отгадать категорию по названию видео. Для этого используем структуру RandomForest из библиотеки sklearn. Будем использовать 4 способа анализа текста.

- Для каждого названия найдем сколько раз в нем встречается каждая буква алфавита.
- Для каждого названия найдем плотность количества каждой буквы алфавита
- Составим слоарь из 100 самых частых слов в датасете и найдем количесто каждого из 100 слов в каждом названии
- Составим слоарь из 100 самых частых слов в датасете и найдем плотность каждого из 100 слов в каждом названии

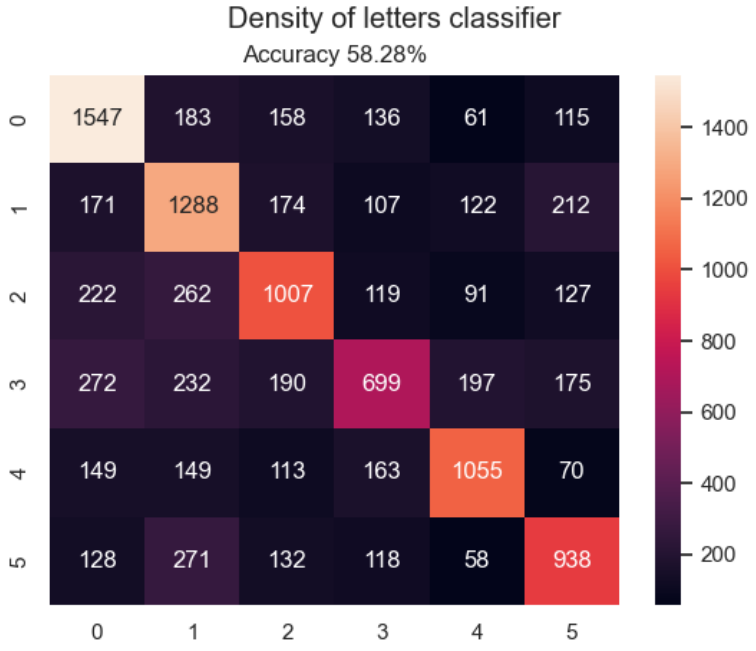
Под плотностью будем подразумевать количество каждой буквы/слова поделить на количетсво букв/слов в тексте. Используем 20 процентов д

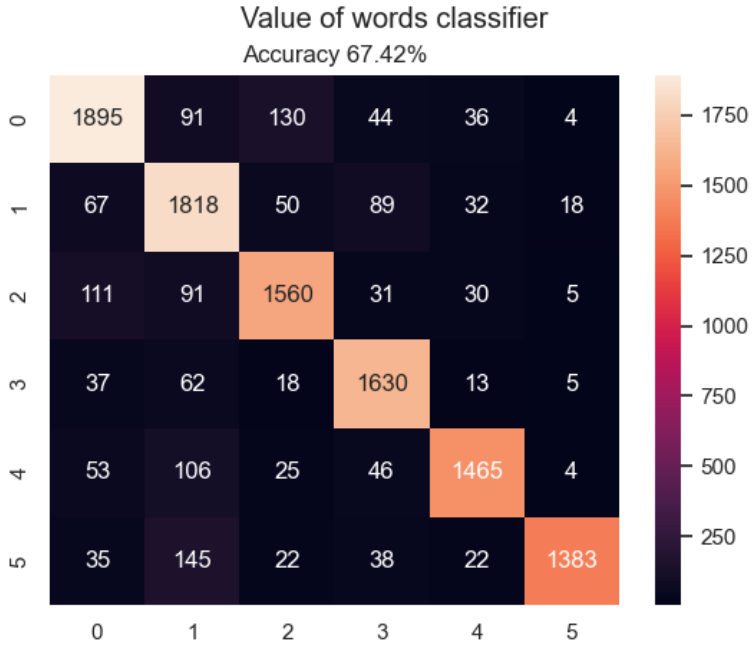
2 Описание решения

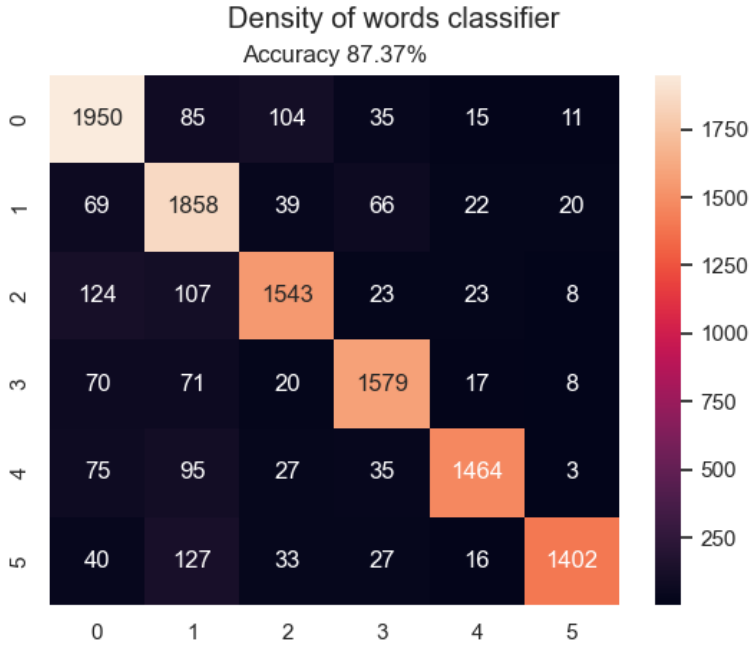
С помощью библиотеки WordCloud проиллюстрируем 100 самых частых слов во всем датасете

Выведем результаты с помощью тепловой карты


Value of letters classifier
Accuracy 67.42%
1711 111 139 141 43 55
168 1432 154 126 88 106
159 201 1196 124 64 84
225 185 138 951 184 82
95 136 58 134 1251 25
120 265 86 110 46 1018
0 1 2 3 4 5
1600 1400 1200 1000 800 600 400 200


Density of letters classifier
Accuracy 58.28%
1547 183 158 136 61 115
171 1288 174 107 122 212
222 262 1007 119 91 127
272 232 190 699 197 175
149 149 113 163 1055 70
128 271 132 118 58 938
0 1 2 3 4 5
1400 1200 1000 800 600 400 200


Value of words classifier
Accuracy 67.42%
1895 91 130 44 36 4
67 1818 50 89 32 18
111 91 1560 31 30 5
37 62 18 1630 13 5
53 106 25 46 1465 4
35 145 22 38 22 1383
0 1 2 3 4 5
1750 1500 1250 1000 750 500 250


Density of words classifier
Accuracy 87.37%
1950 85 104 35 15 11
69 1858 39 66 22 20
124 107 1543 23 23 8
70 71 20 1579 17 8
75 95 27 35 1464 3
40 127 33 27 16 1402
0 1 2 3 4 5
1750 1500 1250 1000 750 500 250

3 Вывод

Классификация по частоте слов показала впечатляющий результат почти в 90 процентов. Однако не менее впечатляющим оказалась классификация по количеству и плотности букв, так как при предполгаемой случайности распределения букв, показала аккуратность в 50 процентов, что выше чем 16 процентов если бы распределение было бы лучайным