

**Московский авиационный институт
(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Искусственный интеллект»

Лабораторная работа № 1
Тема: «Машинное обучение»

Студент: Петрин Сергей Александрович

Группа: М80-307Б-18

Преподаватель: Ахмед Самир Халид

Дата: _____

Оценка: _____

Москва, 2021

1. Постановка задачи

Найти себе набор данных (датасет), для следующей лабораторной работы, и проанализировать его. Выявить проблемы набора данных, устранить их. Визуализировать зависимости, показать распределения некоторых признаков. Реализовать алгоритмы К ближайших соседа с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки `sklearn`.

2. Описание программы

Для данной лабораторной работы я нашел базу данных плохих и хороших отзывов о фильмах на сайте Kaggle. Датасет представляет из себя таблицу с двумя колоннами: отзыв и его качество: позитивный или негативный.

Я обработал эту базу данных для ее использования в машинном обучении. Для начала я проверил, полна ли база данных, проанализировав ее на пустые значения. К счастью, она сразу оказалась полной. После этого я отфильтровал весь текст, избавившись от лишних ненужных символов и преобразовав некоторые слова в более удобную форму. Затем я токенизировал все слова, представив их натуральными числами в некотором указанном мной диапазоне. Данный числовой формат уже пригоден для использования в машинном обучении.

Далее я реализовал алгоритм К ближайшего соседа, используя стандартную документацию и язык Python. Чтобы проверить его корректность, я сравнил свою реализацию программы с реализацией модуля `sklearn`. Полученные результаты оказались схожи.

И, в заключение, я создал свой Байесовский классификатор на базе языка C++, аналогично, сравнив его работу с уже готовым Байесовским классификатором в библиотеке `sklearn`. Полученные результаты оказались схожи.

3. Результаты работы программы

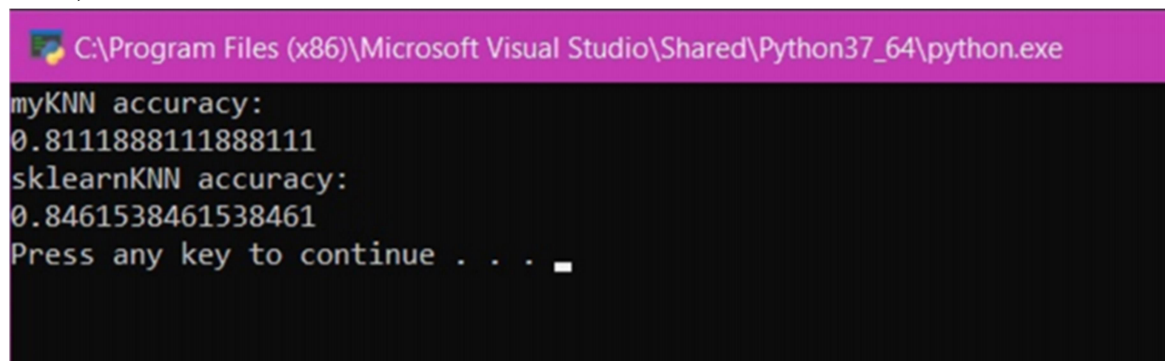
1) Программа анализирующая данные

```

0 One of the other reviewers has mentioned that ...
1 A wonderful little production. cbr /><br />The...
2 I thought this was a wonderful way to spend ti...
3 Basically there's a family where a little boy ...
4 Petter Mattei's "Love in the Time of Money" is...
...
995 Nothing is sacred. Just ask Ennis Fossellius. I...
996 I hated it. I hate self-aware pretentious inan...
997 I usually try to be professional and construct...
998 If you like me is going to see this in a film ...
999 This is like a zoology textbook, given that it...
Name: review, Length: 1000, dtype: object
count of bigrams: 242988
freq (M = count of stop word):
('M': 49381, 'in': 3694, 'i': 3086, 'thi': 2861, '': 2333, 'br': 2330, 'wa': 1835, 'with': 1758, 'for': 1704, 'film': 1679, 'but': 1601, 'movi': 1514, 'on': 1337, 'you': 1276, 'be': 1272, 'not': 1204, 'have': 1177, 'hi': 1130, 'he': 1003, 'one': 983, 'all': 958, 'at': 952, 'by': 849, 'an': 839, 'like': 828, 'from': 818, 'so': 804, 'who': 795, 'they': 769, 'just': 748, 'about': 712, 'or': 655, 'out': 651, 'if': 647, 'it': 631, 'there': 618, 'ha': 611, 'get': 601, 'what': 599, 'see': 594, 'her': 592, 'some': 575, 'make': 573, 'watch': 540, 'veri': 535, 'more': 532, 'when': 531, 'good': 525, 'no': 523, 'even': 517, 'their': 509, 'which': 501, 'would': 491, 'time': 488, 'my': 482, 'me': 445, 'were': 435, 'realli': 432, 'do': 431, 'it': 427, 'she': 418, 'charact': 417, 'can': 413, 'movie': 412, 'other': 405, 'onli': 400, 'well': 383, 'i': 383, 'been': 381, 'scene': 380, 'go': 378, 'will': 377, 'than': 369, 'think': 363, 'look': 360, 'stor': 359, 'most': 346, 'way': 345, 'how': 344, 'first': 343, 'great': 341, 'made': 339, 'also': 336, 'bad': 333, 'show': 328, 'then': 324, 'play': 321, 'the': 307, 'him': 306, 'don't': 305, 'peopl': 304, 'them': 300, 'know': 297, 'end': 294, 'your': 292, 'want': 290, 'too': 283, 'after': 279, 'mani': 277, 'thing': 276, 'ani': 275, 'never': 273, 'two': 271, 'work': 270, 'could': 268, 'seem': 267, 'come': 265, 'act': 264, 'say': 262, 'best': 255, 'where': 251, 'seen': 248, 'off': 246, 'life': 245, 'did': 242, 'doe': 241, 't': 238, 'lover': 231, 'these': 224, 'actor': 219, 'year': 218, 'ever': 217, 'give': 216, 'find': 215, 'back': 214, 'man': 206, 'still': 205, 'actual': 203, 'better': 202, 'use': 201, 'here': 195, 'such': 194, 'part': 189, 'old': 188, 'real': 187, 'lot': 184, 'those': 179, 'whi': 174, 'now': 173, 'i'm': 172, 'down': 171, 'someth': 170, 'director': 168, 'though': 167, 'world': 166, 'pretti': 163, 'doesn't': 161, 'noth': 157, 'ever': 156, 'befor': 155, 'turn': 154, 'us': 153, 'young': 152, 'star': 151, 'interest': 150, 'set': 148, 'around': 146, 'must': 144, 'kill': 142, 'guy': 141, 'least': 138, 'may': 136, 'thought': 134, 's': 133, 'anoth': 132, 'long': 131, 'whole': 130, 'got': 129, 'both': 128, 'believ': 127, 'own': 125, 'there': 124, 'fact': 122, 'read': 121, 'done': 120, 'i've': 119, 'becom': 118, 'saw': 117, 'tv': 116, 'bi': 115, 'laugh': 114, 'screen': 113, 'war': 112, 'rather': 111, 'right': 109, 'shot': 108, 'far': 107, 'funni': 106, 'away': 105, 'idea': 104, 'differ': 103, 'sinc': 102, 'that': 101, 'friend': 100, 'worth': 99, 'expect': 98, 'wonder': 97, 'woman': 96, 'happen': 94, 'mean': 93, 'call': 92, 'comedi': 91, 'goe': 90, 'total': 89, 'home': 88, 'main': 87, 'attempt': 86, 'instead': 85, 'word': 84, 'review': 83, 'given': 82, 'entir': 81, 'and': 79, 'mention': 78, 'until': 77, 'classic': 76, 'leav': 75, 'high': 74, 'talent': 73, 'sens': 72, 'episod': 71, 'lack': 70, 'trull': 69, 'let': 68, 'extrem': 67, 'case': 66, '': 65, 'sex': 64, 'written': 63, 'death': 62, 'less': 61, 'heart': 60, 'pictur': 59, 'manag': 58, 'characters': 57, 'basic': 56, 'couldn't': 55, 'er': 54, 'disappoint': 53, 'terribl': 52, 'sit': 51, 'you'll': 50, 'power': 49, 'style': 48, 'class': 47, 'touch': 46, 'joke': 45, 'order': 44, 'michael': 43, 'due': 42, 'develop': 41, 'deal': 40, 'voic': 39, 'pull': 38, 'theater': 37, 'exactli': 36, 'wouldn't': 35, 'sometim': 34, 'regard': 33, 'street': 32, 'fashion': 31, 'state': 30, 'forget': 29, 'front': 28, 'middl': 27, 'prison': 26, 'violence': 25, 'paint': 24, 'are': 23, 'match': 22, 'readi': 21, 'refer': 20, 'dare': 19, 'brutal': 18, 'foc': 17, 'guard': 16, 'production': 15, 'mainli': 14, 'trust': 13, 'surreal': 12, 'struck': 11, 'mainstream': 10, 'glass': 9, 'drugs': 8, 'mebr': 7, 'oz': 6, 'section': 5, 'cell': 4, 'oswald': 3, 'hooked': 2, 't': 1, 'mid': 1)
[[ 53 384 1 ... 100 3262 437]
[ 90 1 89 ... 512 64 220]
[ 270 2 2 ... 38 7 1024]
...
[ 1 714 12 ... 43 759 904]
[ 57 2278 452 ... 738 326 8]
[ 186 374 100 ... 1 184 4266]]
0 positive
1 positive
2 positive
3 negative
4 positive
...
995 positive
996 negative
997 negative

```

2) KNN



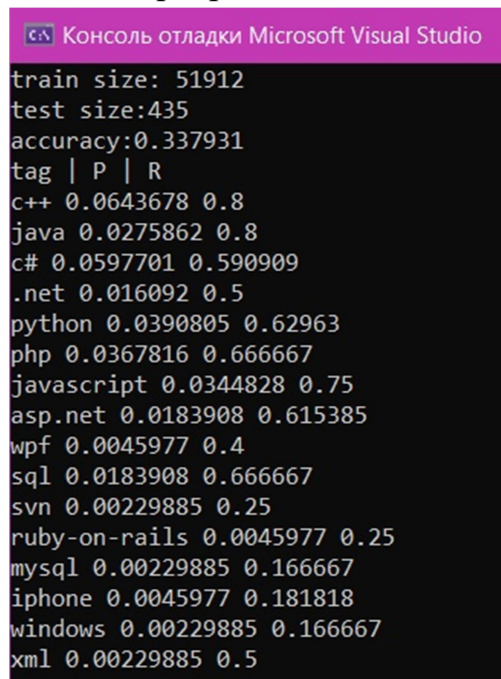
```

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python37_64\python.exe
myKNN accuracy:
0.8111888111888111
sklearnKNN accuracy:
0.8461538461538461
Press any key to continue . . .

```

3) NB

Моя программа:



```

Консоль отладки Microsoft Visual Studio
train size: 51912
test size:435
accuracy:0.337931
tag | P | R
c++ 0.0643678 0.8
java 0.0275862 0.8
c# 0.0597701 0.590909
.net 0.016092 0.5
python 0.0390805 0.62963
php 0.0367816 0.666667
javascript 0.0344828 0.75
asp.net 0.0183908 0.615385
wpf 0.0045977 0.4
sql 0.0183908 0.666667
svn 0.00229885 0.25
ruby-on-rails 0.0045977 0.25
mysql 0.00229885 0.166667
iphone 0.0045977 0.181818
windows 0.00229885 0.166667
xml 0.00229885 0.5

```

Программа sklearn:

```
train size:
51912
test size: 500
sklearnNB accuracy:
0.268
Press any key to continue . . .
```

4. Вывод

В данной лабораторной работе я отыскал некоторую базу данных и, проанализировав ее, подготовил к алгоритмам машинного обучения. Также я реализовал и сами алгоритмы машинного обучения, такие как: К ближайших соседа и Байесовский классификатор. Сравнив мою реализацию с реализацией sklearn, я убедился в достоверной реализации своих программ.

В заключение хочу сказать, что полученные знания и опыт в данной лабораторной работе мне, несомненно, пригодятся, как в будущих студенческих проектах, так и далеко за пределами института.