

**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: Купцов Илья Владимирович

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:

```
[nltk_data] Downloading package wordnet to
[nltk_data] c:\Users\kupts\AppData\Roaming\nltk_data...
[nltk_data] Package wordnet is already up-to-date!
0 One of the other reviewers has mentioned that ...
1 A wonderful little production. <br /><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 Mattel's "Love in the Time of Money" is...
...
995 Nothing is sacred. Just ask Ernie Fossellus. T...
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
('w': 1, 'br': 2, 'in': 3, 'i': 4, 'this': 5, 'wa': 6, 'with': 7, 'fil': 8, 'for': 9, 'but': 10, 'movi': 11, 'on': 12, 'you': 13, 'be': 14, 'not': 15, 'have': 16, 'hi': 17, 'he': 18, 'one': 19, 'all': 20, 'at':
21, 'by': 22, 'an': 23, 'like': 24, 'from': 25, 'who': 26, 'so': 27, 'they': 28, 'just': 29, 'about': 30, 'on': 31, 'out': 32, 'if': 33, 'it': 34, 'there': 35, 'ha': 36, 'what': 37, 'see': 38, 'her': 39, 'get':
40, 'make': 41, 'some': 42, 'it': 43, 'watch': 44, 'good': 45, 'more': 46, 'no': 47, 'when': 48, 'which': 49, 'ver': 50, 'even': 51, 'time': 52, 'their': 53, 'would': 54, 'movie': 55, 'me': 56, 'my': 57, 'the'
: 58, 'do': 59, 'were': 60, 'reall': 61, 'had': 62, 'she': 63, 'well': 64, 'can': 65, 'charact': 66, 'other': 67, 'onli': 68, 'much': 69, 'scene': 70, 'go': 71, 'been': 72, 'into': 73, 'will': 74, 'than': 75, '
think': 76, 'way': 77, 'look': 78, 'becaus': 79, 'stor': 80, 'how': 81, 'first': 82, 'bad': 83, 'most': 84, 'great': 85, 'made': 86, 'also': 87, 'show': 88, 'play': 89, 'then': 90, 'his': 91, 'them': 92, 'don't'
: 93, 'and': 94, 'love': 95, 'too': 96, 'know': 97, 'peopl': 98, 'plot': 99, 'you': 100, 'want': 101, 'thing': 102, 'after': 103, 'week': 104, 'man': 105, 'take': 106, 'ani': 107, 'never': 108, 'too': 109, 'c
ome': 110, 'could': 111, 'act': 112, 'say': 113, 'seem': 114, 'littl': 115, 'best': 116, 'life': 117, 'where': 118, 'seen': 119, 'off': 120, 'and': 121, 'did': 122, 'doe': 123, 'tri': 124, 'over': 125, 'ever': 1
26, 'man': 127, 'here': 128, 'these': 129, 'actor': 130, 'year': 131, 'back': 132, 'give': 133, 'better': 134, 'find': 135, 'still': 136, 'actual': 137, 'while': 138, 'through': 139, 'use': 140, 'such': 141, 'fe
el': 142, 'part': 143, 'real': 144, 'old': 145, 'new': 146, 'perform': 147, 'lot': 148, 'now': 149, 'those': 150, 'world': 151, 'i'm': 152, 'director': 153, 'whi': 154, 'though': 155, 'down': 156, 'a': 157, 'som
eth': 158, 'should': 159, 'enjoy': 160, 'some': 161, 'pretti': 162, 'quit': 163, 'day': 164, 'doesn't': 165, 'can't': 166, 'befor': 167, 'ever': 168, 'star': 169, 'noth': 170, 'turn': 171, 'young': 172, 'set':
173, 'interest': 174, 'cast': 175, 'point': 176, 'few': 177, 'big': 178, 'guy': 179, 'again': 180, 'am': 181, 'around': 182, 'didn't': 183, 'us': 184, 'must': 185, 'girl': 186, 'kill': 187, 'need': 188, 'without'
: 189, 'role': 190, 'least': 191, 'both': 192, 'may': 193, 'thought': 194, 'this': 195, 'enough': 196, 'long': 197, 'start': 198, 'anoth': 199, 'whole': 200, 'book': 201, 'got': 202, 'plum': 203, 'almost': 204
ed': 16883, 'sacred': 16884, 'fossellus': 16885, 'hardwar': 16886, 'fluke': 16887, 'starbuck': 16888, 'flashlight': 16889, 'lightsab': 16890, 'chewchilla': 16891, 'wooki': 16892, 'auggi': 16893, 'doggie': 1689
4, 'nah': 16895, 'fastfoward': 16896, 'shennanigan': 16897, 'clap': 16898, 'cliche': 16899, 'arrives': 16900, 'survivors': 16901, 'redund': 16902, 'esqu': 16903, 'rotations': 16904, 'acid': 16905, 'spat': 16906
, 'disabled': 16907, 'constructed': 16908, 'griny': 16909, 'intension': 16910, 'robbed': 16911, 'zoolog': 16912, 'metr': 16913, 'testing': 16914, 'cameraman': 16915, 'microphon': 16916, 'faintest': 16917, 'expo
sing': 16918, 'journalism': 16919, 'mmo': 16920, 'unsafe': 16921, 'deadlines': 16922, 'creatures': 16923, 'intuit': 16924, 'stubborn': 16925, 'swearing': 16926, 'volume': 16927)
[[ 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
998 negative
999 negative
Name: sentiment, length: 1000, dtype: object
Press any key to continue . . .
```

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
```

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.