Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

Лабораторная работа № 3

# По дисциплине: «Прикладные интеллектуальные системы и

# экспертные системы»

Классификация текстовых данных

Студент Коретников Н.И.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

к.т.н. доцент

Цель работы:

Получить практические навыки решения задачи классификации текстовых данных в среде Jupiter Notebook. Научиться проводить предварительную обработку текстовых данных, настраивать параметры методов классификации и обучать модели, оценивать точность полученных моделей.

Задание кафедры:

1) Загрузить выборки по варианту из лабораторной работы №2

2) Используя GridSearchCV произвести предварительную обработку данных и настройку методов классификации в соответствие с заданием, вывести оптимальные значения параметров и результаты классификации модели (полнота, точность, f1-мера и аккуратности) с данными параметрами. Настройку проводить как на данных со стеммингом, так и на данных, на которых стемминг не применялся.

3) По каждому пункту работы занести в отчет программный код и результат вывода.

4) Оформить сравнительную таблицу с результатами классификации различными методами с разными настройками. Сделать выводы о наиболее подходящем методе классификации ваших данных с указанием параметров метода и описанием предварительной обработки данных.

Вариант 6 Методы: (DT, KNN, LR).

Ход работы:

Исходные данные из 2 лабораторной работы импортируем.

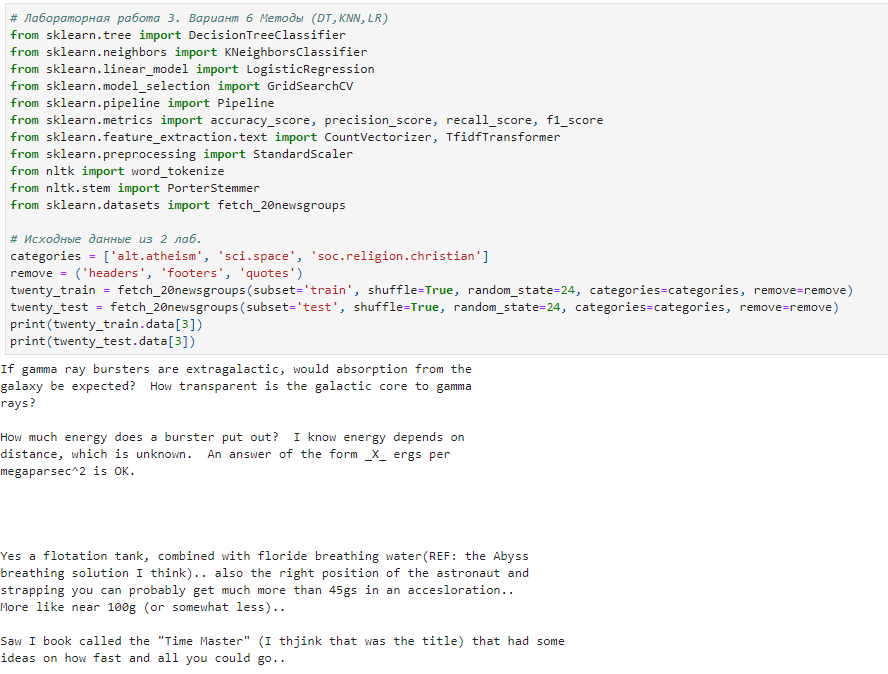


Рисунок 1 – Исходные данные (1)

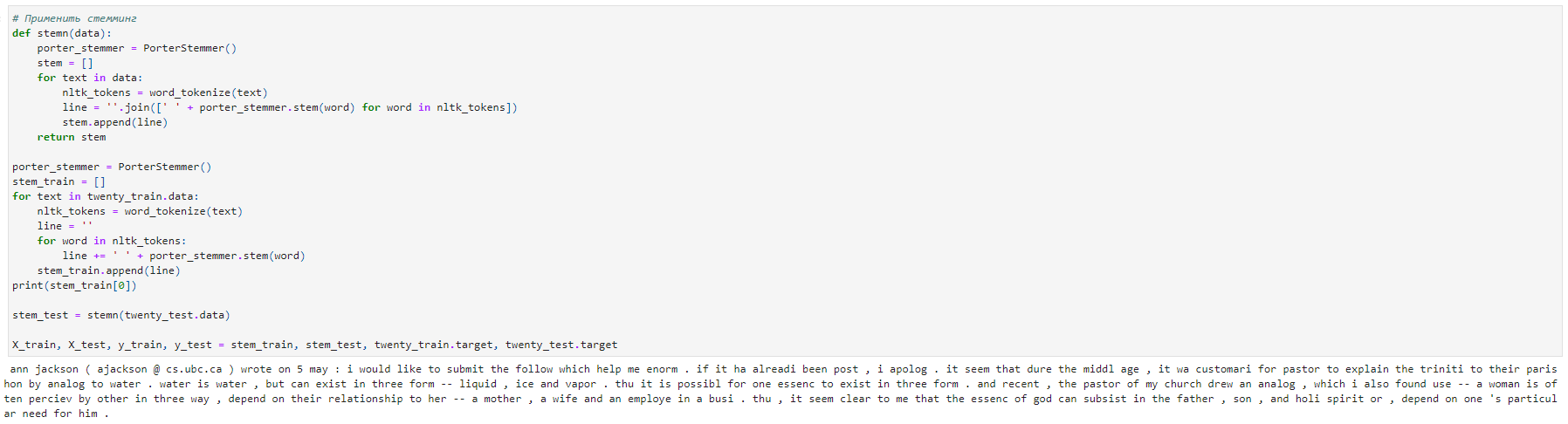


Рисунок 2 – Исходные данные (2)

Далее создаем модели с применением таких методов, как:

Дерево решений (DT): • критерий (параметр criterion: ‘gini’, ‘entropy’), • глубина дерева (параметр max\_depth от 1 до 5 с шагом 1, далее до 100 с шагом 20).

К-ближайших соседей (KNN): • количество ближайших соседей, • метрика (евклидова, городских кварталов).

Логистическая регрессия (LR): • метод нахождения экстремума (параметр solver: ‘newton-cg’, ‘lbfgs’, ‘sag’, ‘liblinear’), • регуляризация (параметр penalty: ‘L1’, ‘L2’) Обратить внимание, что разные виды регуляризации работают с разными методами нахождения экстремума.

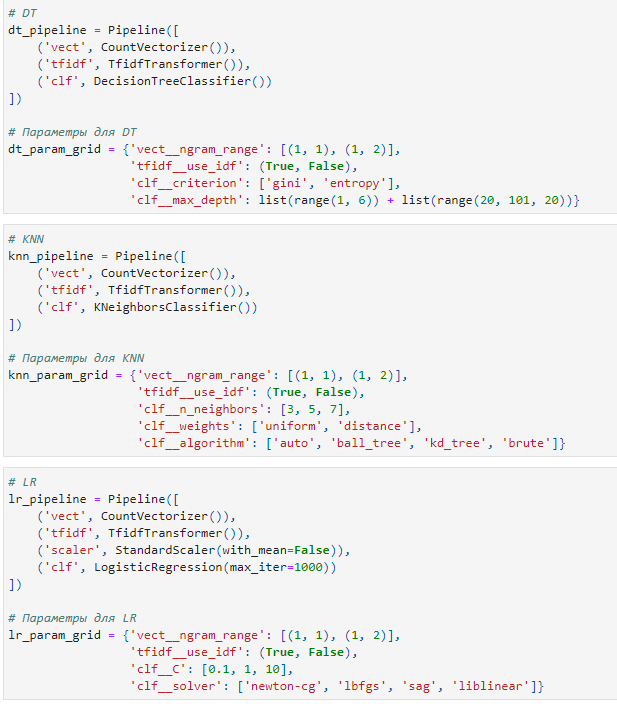


Рисунок 3 – Создание моделей методов DT, KNN, LR

Настроим параметры и опишим цикл обучения программы для используемых методов.



Рисунок 4 – Цикл обучения и настройка параметров

Результат программы описан на рисунке 5.

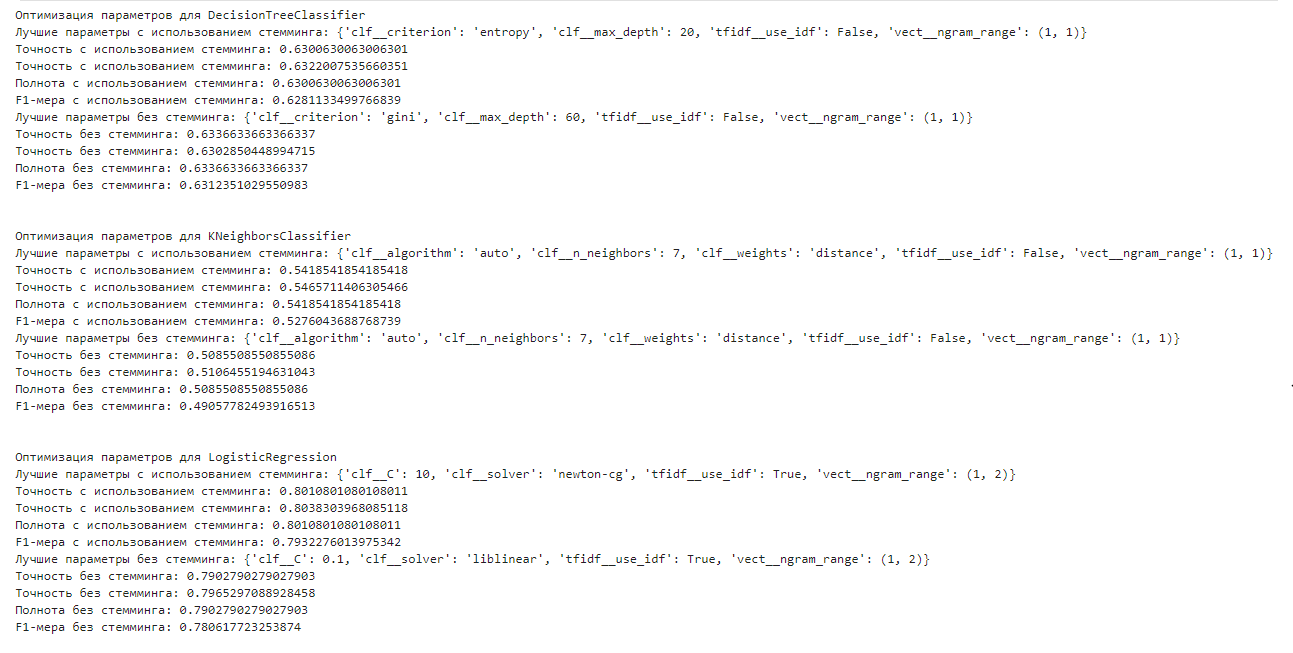


Рисунок 5 – Результат

Вывод:

В данной лабораторной работе получили практические навыки решения задачи классификации текстовых данных в среде Jupiter Notebook. Научились проводить предварительную обработку текстовых данных, настраивать параметры методов классификации и обучать модели, оценивать точность полученных моделей.

Также в данной работе были использованы методы DT, KNN, LR. В наилучшем по точности оказался метод логистической регрессии.