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 Фетисов В. Д.

Группа М-ИАП-23

Руководитель Кургасов В. В.

Доцент

Липецк 2023 г.

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 №17

Классы 4, 14, 18 ('comp.os.ms-windows.misc', 'sci.med','talk.politics.mideast') Методы KNN, LR, MNB

Ход работы

На рисунках 1-16 представлен код и процесс работы программы.



Рисунок 1 – Подключение библиотек и категорий по варианту

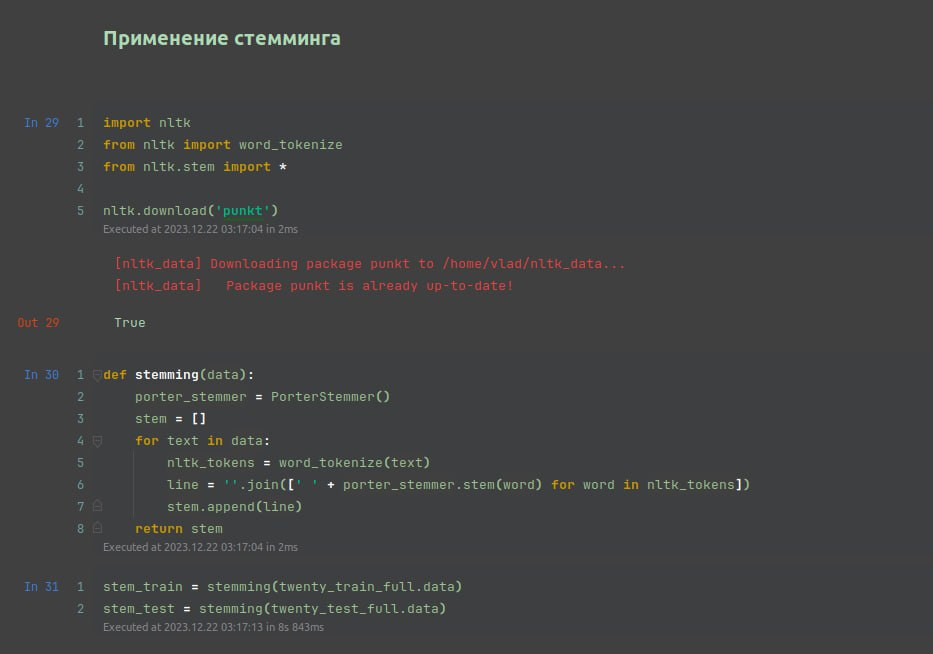


Рисунок 2 – Применение стемминга

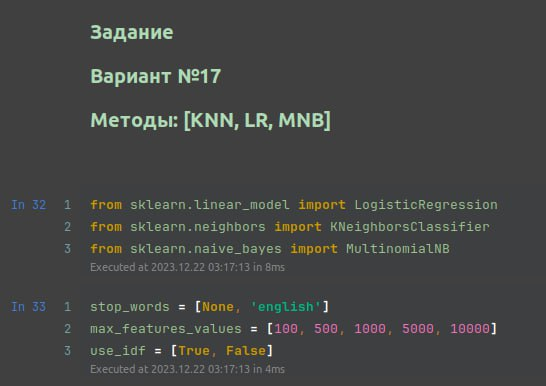


Рисунок 3 – Подключение методов



Рисунок 4 – Настройка параметров



Рисунок 5 – KNN без стемминга

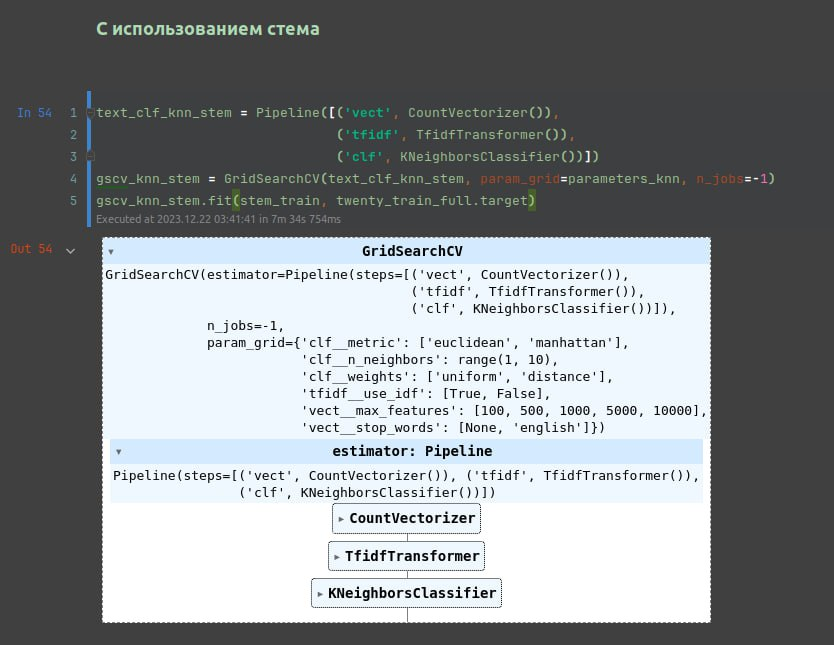


Рисунок 6 - KNN со стеммингом

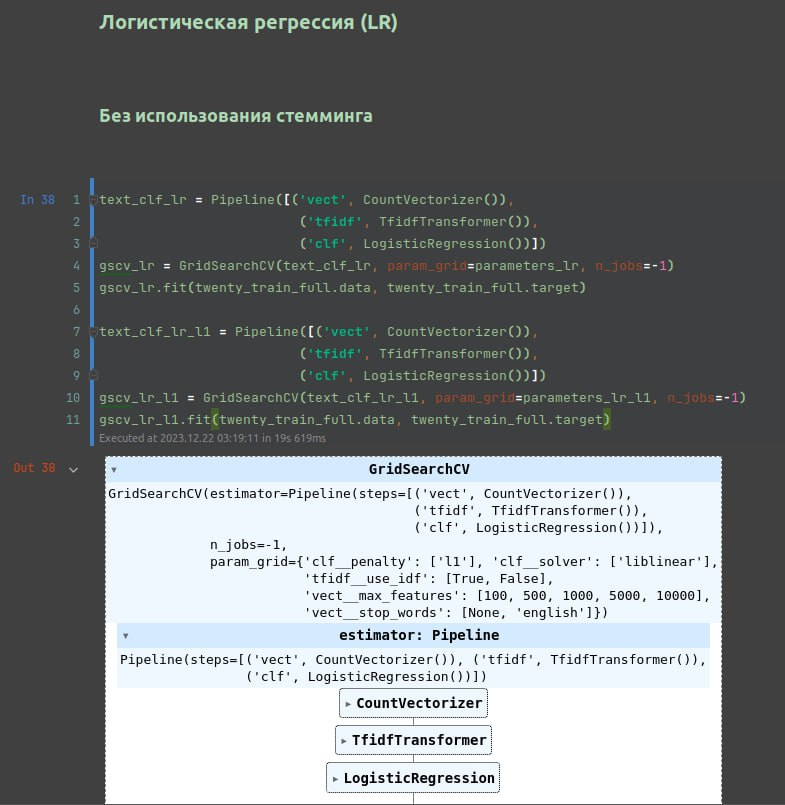


Рисунок 7 - LR без стемминга

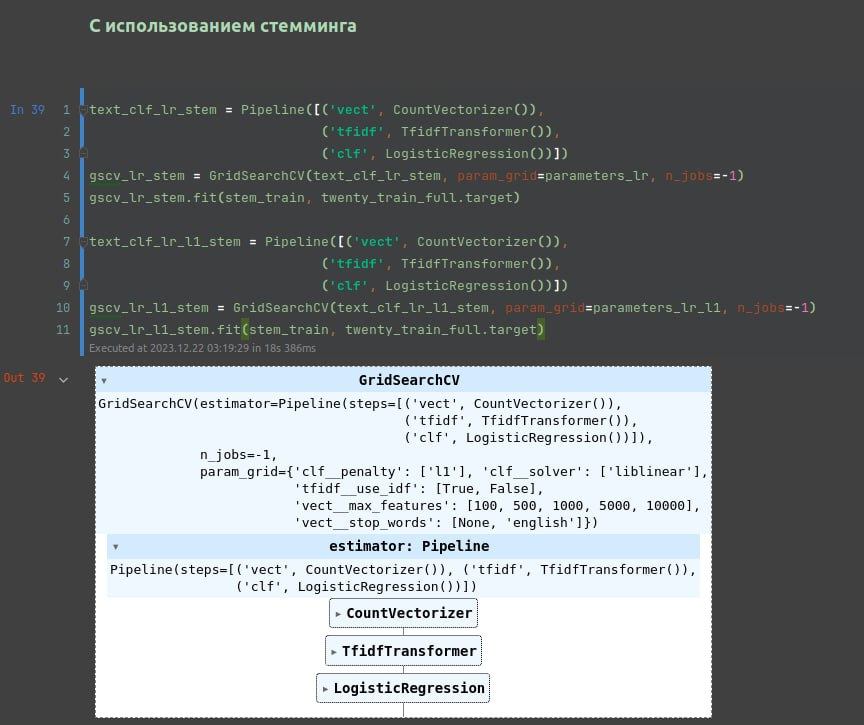


Рисунок 8 - LR со стеммингом

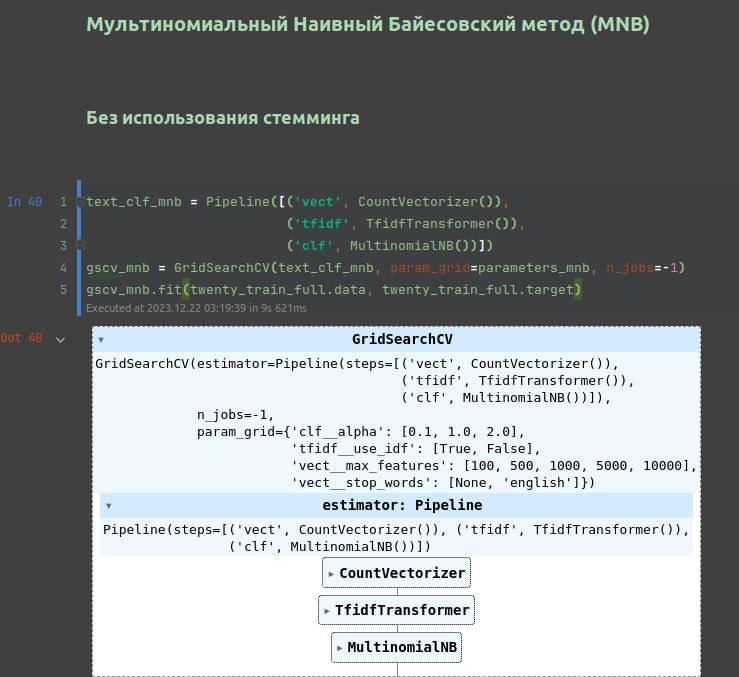


Рисунок 9 - MNB без стемминга

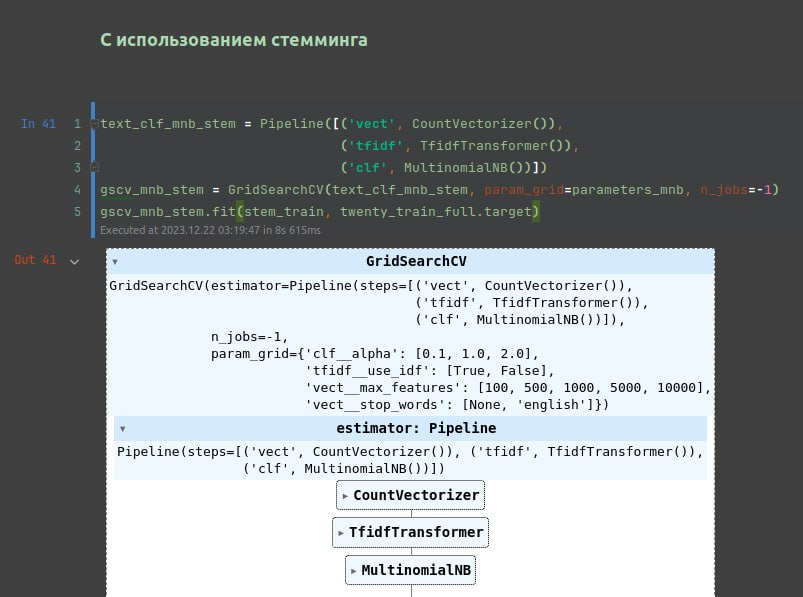


Рисунок 10 - MNB со стеммингом

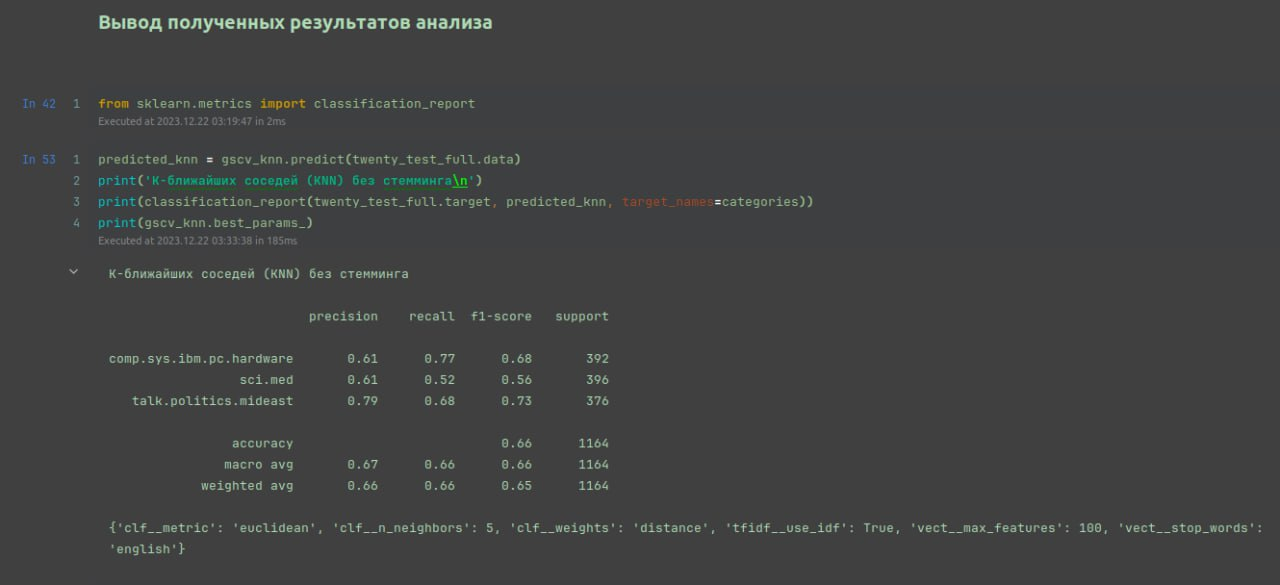


Рисунок 11 – Результаты KNN без стемминга

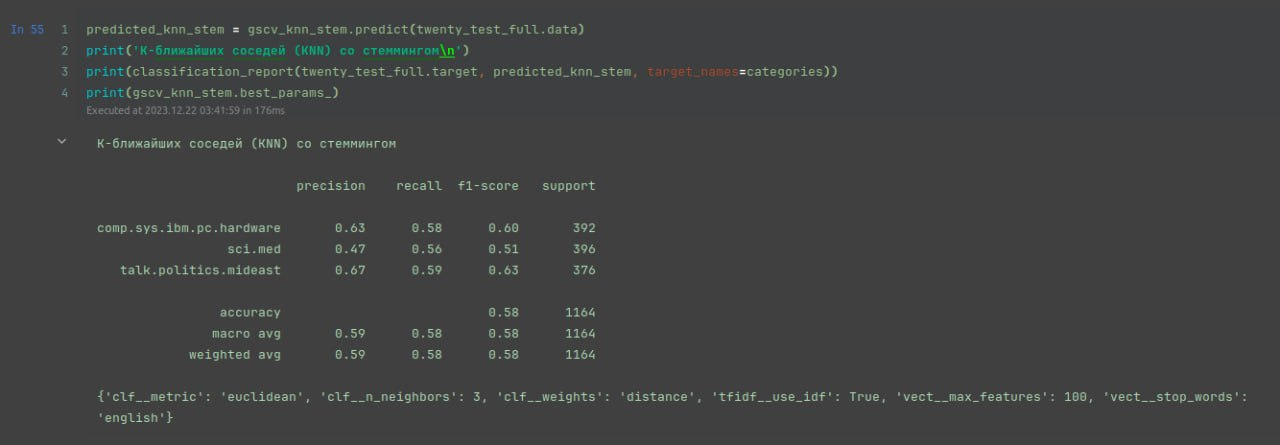


Рисунок 12 - Результаты KNN со стеммингом

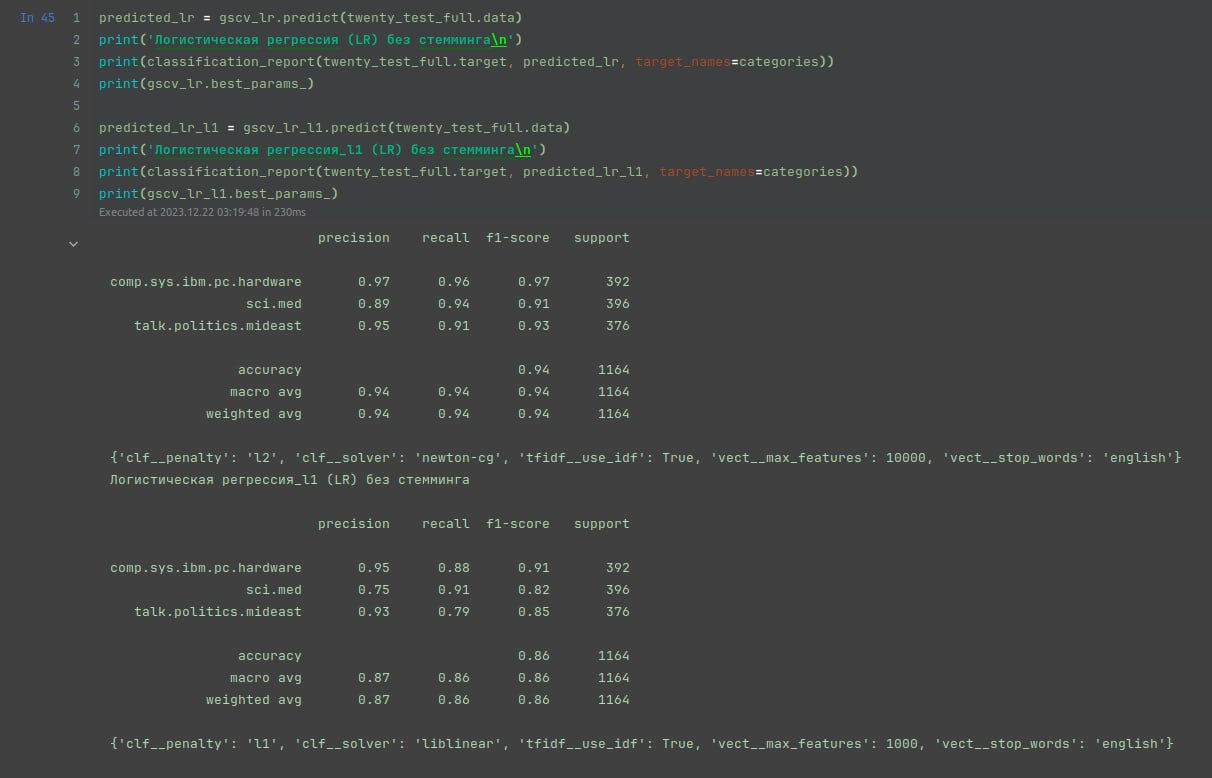


Рисунок 13 - Результаты LR без стемминга

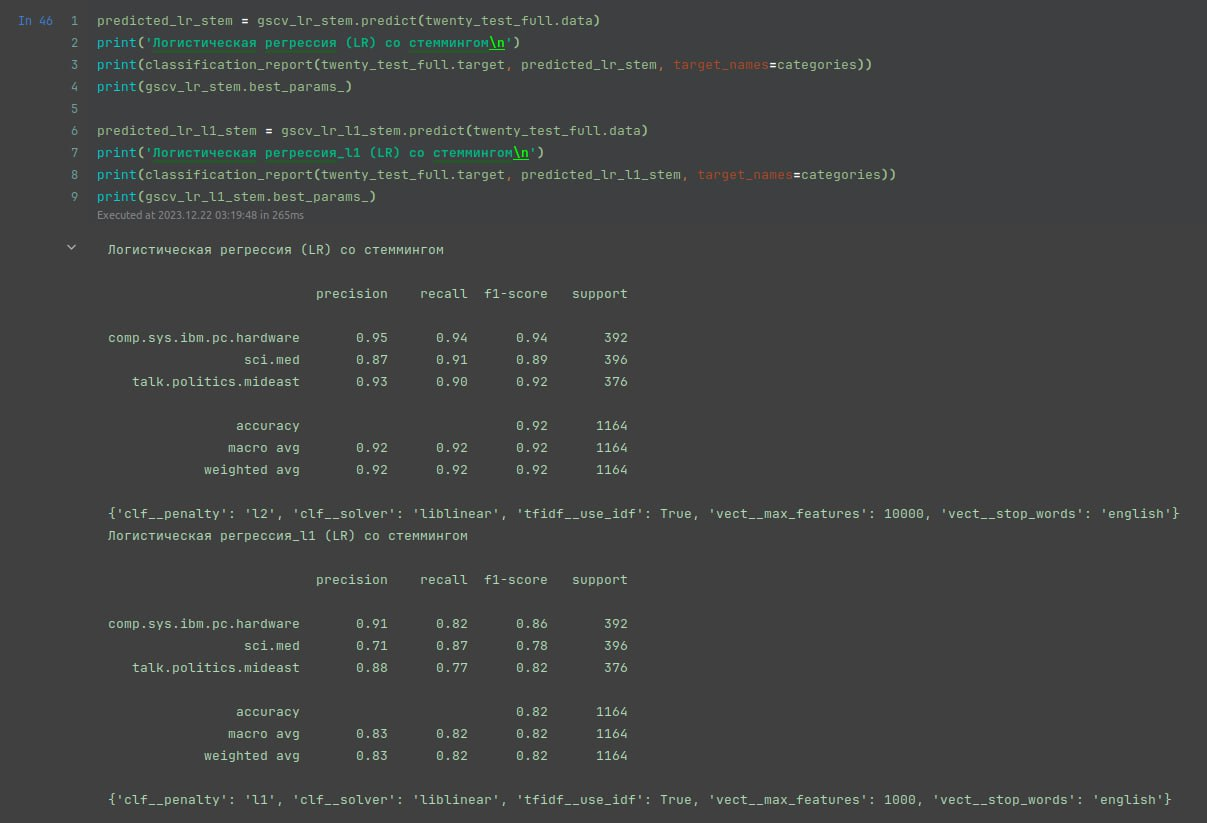


Рисунок 14 - Результаты LR со стеммингом

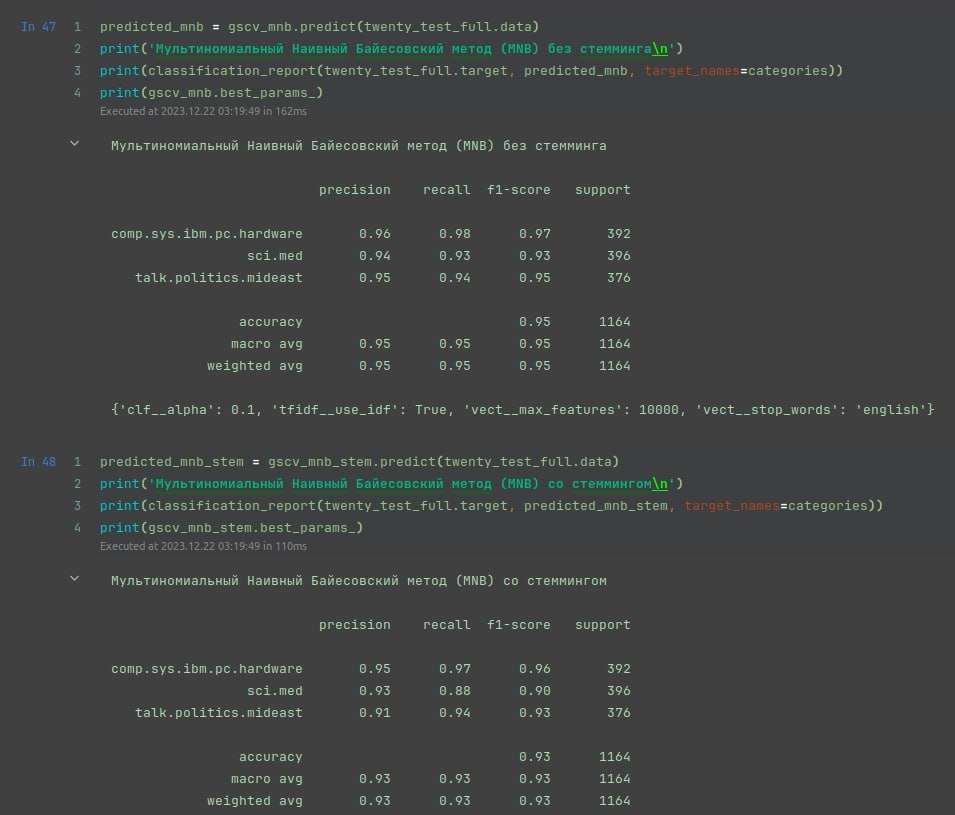


Рисунок 15 - Результаты MNB без стемминга и со стеммингом

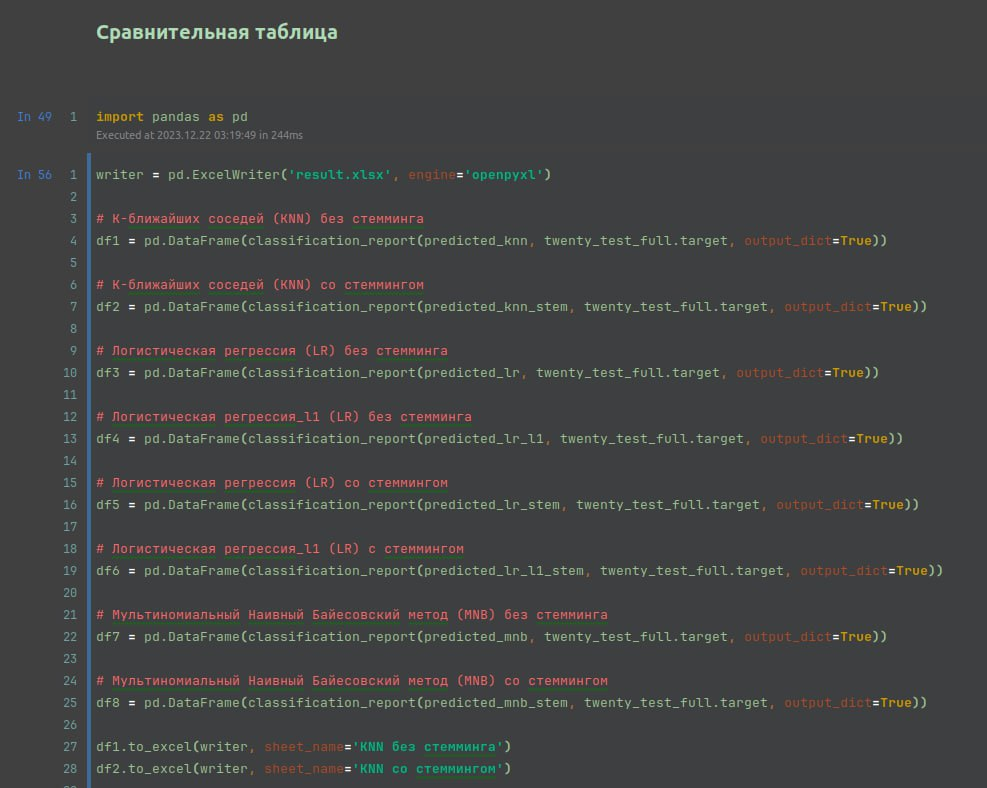


Рисунок 16 - Итог

Вывод

В результате выполнения данной лабораторной работы я получил практические навыки решения задачи классификации текстовых данных в среде Jupiter Notebook. Мною были применены следующие методы: k-ближайших соседей (KNN), логистической регрессии (LR) и наивный байесовский (MNB).

Наилучшей точностью классификации для данного набора данных обладает модель наивный байесовский метод (MNB) без применения стемминга. Точность которого составила 95%.