1. Классы:
    - Происшествия;
    - Здоровье;
    - Общество;
    - Политика;
    - Культура;
    - Экономика;
    - Технологии;
    - Экология.
2. Вектор признаков — сумка слов (bag of words), представляющая из себя набор слов с указанной частотой их повторения в тексте (TF-IDF) и класс (категория), к которому они принадлежат.
3. Способ векторизации — Метод опорных векторов
4. Обучающая выборка представлена в файле формата csv в директории: `app/modules/classifier/data/stemmed_text.csv`
   Модель классификатора представлена в директории: `app/modules/classifier/model/model.pkl`

Для классификации текста производится следующая обработка текста:

1. Предобработка – удаление знаков препинания и пунктуации, замена заглавных букв на прописные.
2. Стемминг – выделение основы слов без их окончаний.
3. Лемматизация – приведение глаголов к начальной форме (в инфинитив)

Данная обработка выполняется для текста, используемого в обучающей выборке и для текста, поступающего в классификатор для определения его категории.

Для работы классификатора используется метод опорных векторов, с помощью которого строится гиперплоскость, разделяющая объекты выборки оптимальным способом. Алгоритм измеряет расстояние между разделяющей гиперплоскостью и объектами, тем самым выполняет классификацию объектов.

Для тренировки классификатора было сформировано суммарно 3360 текстов, равномерно распределенных по 8 категориям (см п.1).

Точность классификации текста составила 74,6%. Результаты теста приведены на рисунке 1.

Точность работы классификатора: 0.7465208747514911

	precision	recall	f1-score	support
Происшествия	0.85	0.88	0.87	135
Здоровье	0.74	0.72	0.73	120
Общество	0.83	0.71	0.77	133
Политика	0.63	0.56	0.60	119
Культура	0.77	0.80	0.78	139
Экономика	0.74	0.82	0.78	119
Технологии	0.75	0.61	0.67	127
Экология	0.66	0.86	0.75	114
accuracy			0.75	1006
macro avg	0.75	0.75	0.74	1006
weighted avg	0.75	0.75	0.74	1006

Рисунок 1 – Результаты работы классификатора.