**Лабораторная работа №3**

**Классификация текстовых данных**

*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:*

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.

**Варианты заданий:**

|  |  |
| --- | --- |
| **Вариант** | **Методы** |
| 1 | KNN, RF, LR |
| 2 | RF, MNB, SVM |
| 3 | KNN, DT, SVM |
| 4 | RF, MNB, KNN |
| 5 | LR, KNN, MNB |
| 6 | DT, KNN, LR |
| 7 | RF, LR, SVM |
| 8 | SVM, DT, LR |
| 9 | RF, LR, KNN |
| 10 | DT, SVM, LR |
| 11 | MNB, DT, KNN |
| 12 | RF, SVM, LR |

**Параметры, которые необходимо настроить**

Помимо параметров предварительной обработки, таких как: взвешивание, отсечение стоп-слов, количество информативных терминов, для каждого метода классификации необходимо настроить следующие параметры:

**К-ближайших соседей (*KNN*):**

* количество ближайших соседей,
* метрика (евклидова, городских кварталов)

**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).

**Случайный лес (*RF*):**

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

**Логистическая регрессия (*LR*):**

* метод нахождения экстремума (*параметр solver: ‘newton-cg’, ‘lbfgs’, ‘sag’, ‘liblinear’*),
* регуляризация (параметр penalty: ‘*L*1’, ‘*L*2’)

Обратить внимание, что разные виды регуляризации работают с разными методами нахождения экстремума.

**Метод опорных векторов (*SVM*):**

* функция потерь (параметр *loss: ‘hinge’, ‘squared\_hinge’*),
* регуляризация (параметр penalty: ‘*L*1’, ‘*L*2’)

Обратить внимание, что разные виды регуляризации работают с разными функциями потерь

**Мультиномиальный** **Наивный Байесовский метод (*MNB*)**

* параметр сглаживания α (параметр *alpha* {0,1;1;2}

**Контрольные вопросы**

1. Алгоритм и особенности метода опорных векторов.
2. Алгоритм и особенности метода логистической регрессии.
3. Алгоритм и особенности метода деревьев решений.
4. Что такое регуляризация?
5. Что такое метрика расстояния? Какие метрики вам известны?