**Отчет по заданию 2 «Наивный байесовский классификатор»**

Для выполнения задания использовался датасет сообщений следующих групп ВК:

* [ИТИС КФУ](https://vk.com/itis_kfu) – 2005 сообщений
* [Strategium](https://vk.com/strategium) – 2039 сообщений

Общий размер датасета – 4044 сообщений.

**Токенизация данных**

В начале предобработки данных была проведена токенизация по словам. Разделителями слов считались пробел, символ новой строки и нижнее подчеркивание (встречается, например, в тегах).

При токенизации символ новой строки и нижнее подчеркивание были заменены на пробелы. Далее сообщения были разделены на слова с помощью метода split().

**Предобработка данных**

Предобработка данных включает в себя следующие этапы:

**1. Приведение к нижнему регистру.**

**2. Удаление URL.** Были удалены все токены, включающие в себя подстроки "http://" или "https://".

**3. Извлечение слов из токенов**

Будем считать, что валидные слова обладают следующими свойствами:

* Состоят из букв русского и английского алфавитов.
* Могут содержать 1 дефис не на краю слова.
* Каждый токен может содержать только 1 валидное слово.

На основании этих свойств было составлено следующее регулярное выражение:

([a-zа-яё]+-?[a-zа-яё]+)|[a-zа-яё]

Далее из каждый токен был преобразован в валидное слово – подстроку токена, соответствующую этому регулярному выражению. Если токен не имеет подстрок, соответствующих регулярному выражению, токен заменяется на пустую строку.

Благодаря этой операции были достигнуты следующие результаты:

* Очистка токенов от знаков препинания.
* Удаление токенов, являющихся шумом (числа, тире, смайлики и др.)

**4. Лемматизация** с помощью pymorphy2.

**5. Удаление пустых токенов.**

**6. Удаление стоп-слов** с помощью nltk.

**Обучение модели**

Сначала выборки сообщений были перемешаны, чтобы как в обучающую, так и в тестовую выборки попали и старые, и новые сообщения

Далее данные были разделены на обучающую и тестовую выборки. В обучающую выборку вошло по 1500 сообщений из каждой группы, всего – 3000 сообщений. В тестовую выборку вошли оставшиеся 1044 сообщения.

Сообщения из группы ИТИС КФУ получили метку 0, из группы Strategium – 1.

После этого было проведено обучение наивного байесовского классификатора. Словари обученной модели были экспортированы в pkl-файл.

**Проверка качества модели**

После обучения модели был получен прогноз классификатора на тестовой выборке и были вычислены значения различных метрик классификации.

|  |  |
| --- | --- |
| **Метрика** | **Значение** |
| Accuracy | 0.971 |
| Accuracy (сообщения ИТИС КФУ) | 0.990 |
| Accuracy (сообщения Strategium) | 0.954 |
| Precision | 0.990 |
| Recall | 0.954 |
| AUC ROC | 0.972 |

Из 1044 сообщений тестовой выборки было неверно классифицировано всего 30 сообщений. Из 505 сообщений группы ИТИС КФУ было неверно классифицировано 5 сообщений, из 5039 сообщений группы Strategium – 25 сообщений.

Таким образом, модель имеет очень высокую точность.