**Лабораторная работа №2**

**Предварительная обработка текстовых данных**

*Цель работы:*

Получить практические навыки обработки текстовых данных в среде Jupiter Notebook. Научиться проводить предварительную обработку текстовых данных и выявлять параметры обработки, позволяющие добиться наилучшей точности классификации.

*Задание:*

1. В среде Jupiter Notebook создать новый ноутбук (Notebook)
2. Импортировать необходимые для работы библиотеки и модули
3. Загрузить обучающую и экзаменационную выборку в соответствие с вариантом
4. Вывести на экран по одному-два документа каждого класса.
5. Применить стемминг, записав обработанные выборки (тестовую и обучающую) в новые переменные.
6. Провести векторизацию выборки:
7. Векторизовать обучающую и тестовую выборки простым подсчетом слов (CountVectorizer) и значеним max\_features = 10000
8. Вывести и проанализировать первые 20 наиболее частотных слов всей выборки и каждого класса по-отдельности.
9. Применить процедуру отсечения стоп-слов и повторить пункт b.
10. Провести пункты a – c для обучающей и тестовой выборки, для которой проведена процедура стемминга.
11. Векторизовать выборки с помощью TfidfTransformer (с использованием TF и TF-IDF взвешиваний) и повторить пункты b-d.
12. По результатам пункта 6 заполнить таблицы наиболее частотными терминами обучающей выборки:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Без стемминга | | | | | |
|  | Count | | TF | | TF-IDF | |
| № | Без стоп-слов | С стоп-словами | Без стоп-слов | С стоп-словами | Без стоп-слов | С стоп-словами |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| 20 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Со стеммингом | | | | | |
|  | Count | | TF | | TF-IDF | |
| № | Без стоп-слов | С стоп-словами | Без стоп-слов | С стоп-словами | Без стоп-слов | С стоп-словами |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| 20 |  |  |  |  |  |  |

1. Используя конвейер (Pipeline) реализовать модель Наивного Байесовского классификатора и выявить на основе показателей качества (значения полноты, точности, f1-меры и аккуратности), какая предварительная обработка данных обеспечит наилучшие результаты классификации. Должны быть исследованы следующие характеристики:

* Наличие - отсутствие стемминга
* Отсечение – не отсечение стоп-слов
* Количество информативных терминов (max\_features)
* Взвешивание: Count, TF, TF-IDF

1. По каждому пункту работы занести в отчет программный код и результат вывода.
2. По результатам классификации занести в отчет выводы о наиболее подходящей предварительной обработке данных (наличие стемминга, взвешивание терминов, стоп-слова, количество информативных терминов).

**Варианты заданий:**

|  |  |
| --- | --- |
| **Вариант** | **Классы** |
| 1 | 2, 3, 8 |
| 2 | 6, 10, 11 |
| 3 | 1, 9, 17 |
| 4 | 7, 12, 18 |
| 5 | 4, 14, 18 |
| 6 | 1, 15, 16 |
| 7 | 3, 7, 13 |
| 8 | 5, 16, 20 |
| 9 | 6, 17, 19 |
| 10 | 3, 5, 15 |
| 11 | 7, 14, 20 |
| 12 | 2, 12, 13 |

**Названия классов**

|  |  |
| --- | --- |
| **№ класса** | **Название класса** |
| 1 | 'alt.atheism' |
| 2 | 'comp.graphics' |
| 3 | 'comp.os.ms-windows.misc' |
| 4 | 'comp.sys.ibm.pc.hardware' |
| 5 | 'comp.sys.mac.hardware' |
| 6 | 'comp.windows.x' |
| 7 | 'misc.forsale' |
| 8 | 'rec.autos' |
| 9 | 'rec.motorcycles' |
| 10 | 'rec.sport.baseball' |
| 11 | 'rec.sport.hockey' |
| 12 | 'sci.crypt' |
| 13 | 'sci.electronics' |
| 14 | 'sci.med' |
| 15 | 'sci.space' |
| 16 | 'soc.religion.christian' |
| 17 | 'talk.politics.guns' |
| 18 | 'talk.politics.mideast' |
| 19 | 'talk.politics.misc' |
| 20 | 'talk.religion.misc' |

**Контрольные вопросы**

1. Особенности задачи классификации текстовых данных.
2. Этапы предварительной обработки данных.
3. Алгоритм и особенности Наивного Байесовского метода.
4. Как влияет размер словаря терминов на точность классификации?
5. Какие способы выявления информативных терминов вам известны?
6. Как влияет способ взвешивания терминов на точность классификации?