Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

Лабораторная работа № 2

# По дисциплине: «Прикладные интеллектуальные системы и

# экспертные системы»

Предварительная обработка текстовых данных

Студент Коретников Н.И.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

к.т.н. доцент

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 кафедры:

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:

a. Векторизовать обучающую и тестовую выборки простым подсчетом слов (CountVectorizer) и значеним max\_features = 10000

b. Вывести и проанализировать первые 20 наиболее частотных слов всей выборки и каждого класса по-отдельности.

c. Применить процедуру отсечения стопслов и повторить пункт b.

d. Провести пункты a – c для обучающей и тестовой выборки, для которой проведена процедура стемминга.

e. Векторизовать выборки с помощью TfidfTransformer (с использованием TF и TF-IDF взвешиваний) и повторить пункты b-d.

7) По результатам пункта 6 заполнить таблицы наиболее частотными терминами обучающей выборки и каждого класса по отдельности. Всего должно получиться по 4 таблицы для выборки, к которой применялась операция стемминга и 4 таблицы для выборки, к которой операция стемминга не применялась.





8) 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

9) По каждому пункту работы занести в отчет программный код и результат вывода.

10) По результатам классификации занести в отчет выводы о наиболее подходящей предварительной обработке данных (наличие стемминга, взвешивание терминов, стоп-слова, количество информативных терминов).

Вариант 6 (классы 1,15,16: alt.atheism, sci.space, soc.religion.christian).

Ход работы:

Импорт библиотек и обозначение варианта выполнения работы.

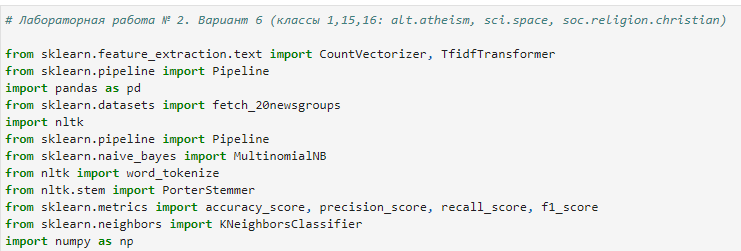


Рисунок 1 – Импорт библиотек

Загружаем данные dataset.

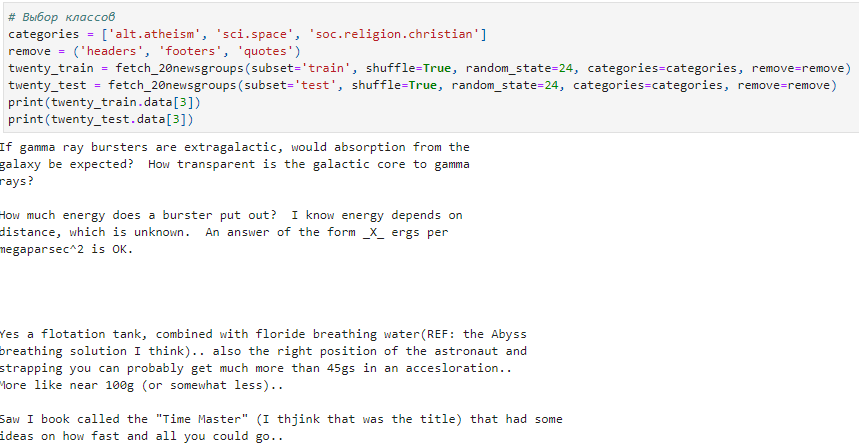


Рисунок 2 – Загрузка dataset

Далее с получением и загрузки dataset применяем к ним стемминг. Применение отображено на рисунке 3.

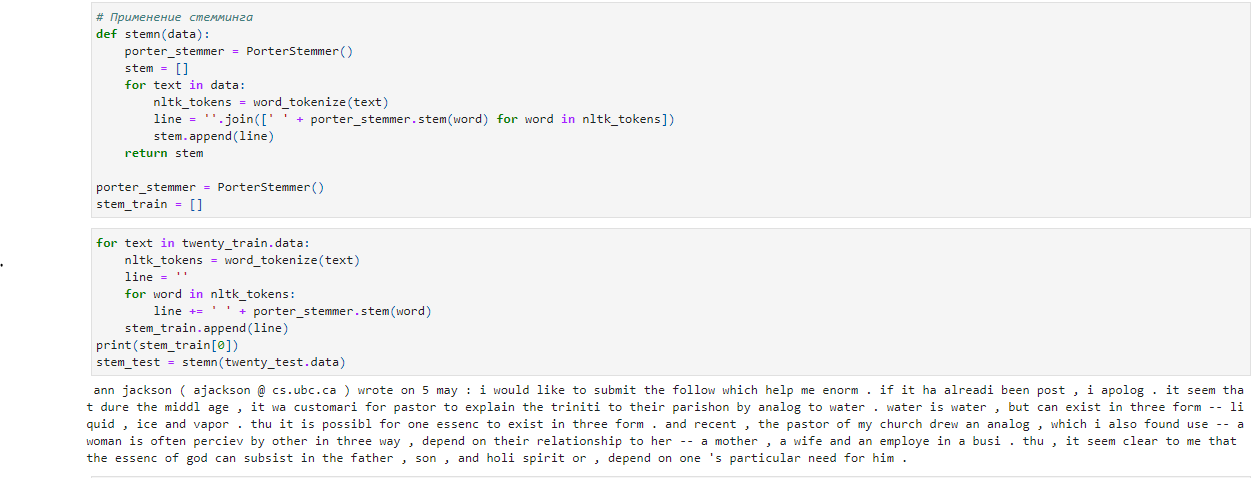


Рисунок 3 – Применение стемминга

Проведем векторизацию выборки со значением max\_features = 1000.



Рисунок 4 – Векторизация выборки

После отсекаем стоп слова повторяя тот же цикл действий.

Выполним векторизацию с помощью TfdgTransformer с использованием (TF и TF-IDF взвешиваний) со стоп-слово и без стоп-слово.

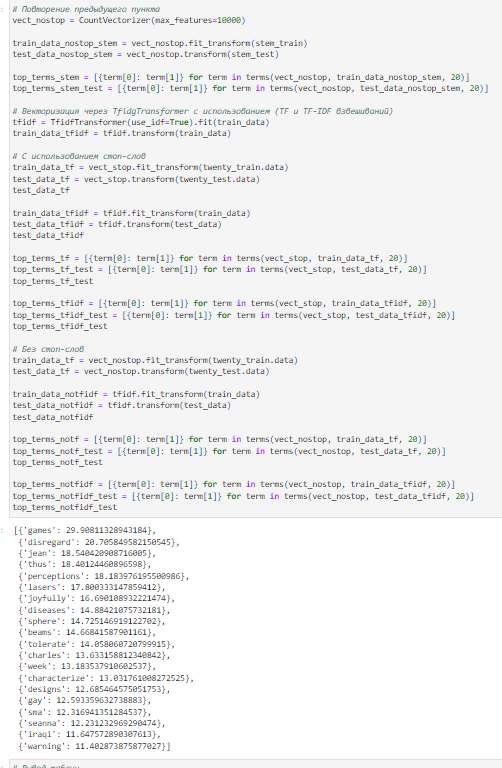


Рисунок 5 – Код векторизации с отсечением стоп-слов и с применением TfdgTransformer

Используем конвейер Pipeline для реализации модель Наивного Байесовского классификатора, чтобы определить точность классификации и других характеристик с различными вариантами.



Рисунок 6 – Код модели Наивного Байесовского классификатора

Выведем полученные результаты.

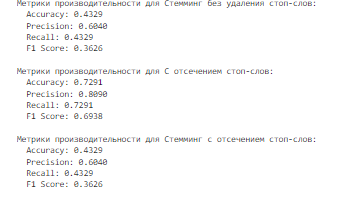


Рисунок 7 – Результат работы модели

Далее изобразим таблицы для 6 пункта по выборки на рисунках 8,9,10,11.

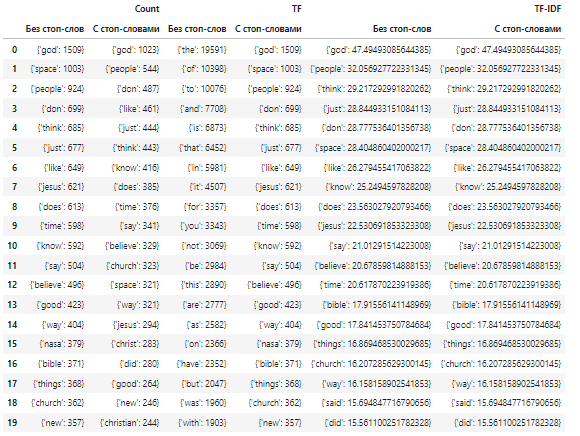


Рисунок 8 – Результат векторизации данных обучающего множества

без стемминга

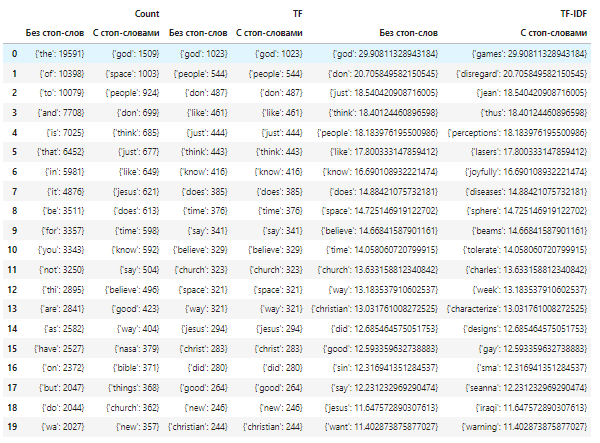


Рисунок 9 – Результат векторизации данных тестового множества без стемминга

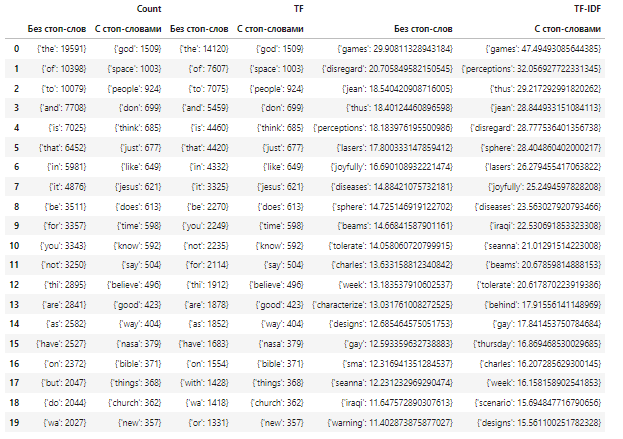


Рисунок 10 – Результат векторизации данных обучающего множества со стеммигом

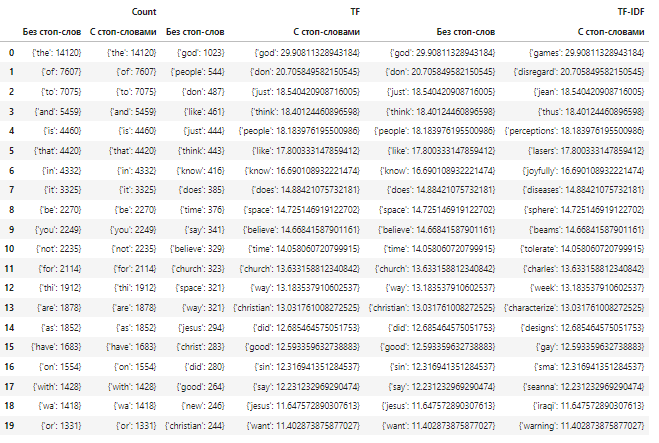


Рисунок 11 – Результат векторизации данных тестового множества со стеммигом

Вывод:

В процессе выполнения лабораторной работы получили практические навыки обработки текстовых данных в среде Jupiter Notebook. Научились проводить предварительную обработку текстовых данных и выявлять параметры обработки, позволяющие добиться наилучшей точности классификации.