|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме \_\_\_\_\_\_\_ «Датасет токсичных комментариев с Youtube» \_\_\_ \_\_\_ \_\_ \_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-52Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Е. В. Козлов

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** К. Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2022 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме Анализ данных по базе данных «Токсичные комментарии с Youtube» \_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_\_\_\_ИУ5-52Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Козлов Егор Васильевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Производственная\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

Техническое задание Спроектировать систему анализа и визуализации по теме «Токсичные комментарии с Youtube», провести визуализацию полученных данных посредством языка программирования Python

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 24 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «10» сентября 2022 г.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc122019940)

[1. Определение данных для анализа 5](#_Toc122019941)

[2. Описание данных 5](#_Toc122019942)

[3. Формулирование гипотез 5](#_Toc122019943)

[4. Изучение общей информации 6](#_Toc122019944)

[5. Подготовка данных для работы 7](#_Toc122019945)

[6. Агрегирование данных 10](#_Toc122019946)

[7. Исследовательский анализ данных 12](#_Toc122019947)

[8 Общий вывод 17](#_Toc122019948)

[ЗАКЛЮЧЕНИЕ 18](#_Toc122019949)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 19](#_Toc122019950)

# ВВЕДЕНИЕ

Цель:

Необходимо исследовать базу данных и выявить закономерности между популярностью слова и его тематикой, токсичностью; отношение токсичных и остальных комментариев.

Задачи:

* 1. определение данных;
  2. формулирование гипотез;
  3. загрузка данных в Python;
  4. проверка данных;
  5. очистка данных;
  6. преобразование данных;
  7. выбор данных для анализа;
  8. агрегирование данных;
  9. визуализация данных;
  10. подтверждение или опровержение поставленных гипотез;
  11. формулирование ограничений и выводов.

# Определение данных для анализа

В качестве данных для анализа был взят датасет с сайта Youtube, позволяющий людям со всей планеты смотреть все жанры видео, а также предоставляющий огромный инструментарий для общения и комментирования. Для анализа было взято 1000 комментариев.

1. **Описание данных**

В наборе данных содержатся:

* + 1. id комментария;
    2. id видео;
    3. полный текст комментария;
    4. является ли комментарий токсичным;
    5. является ли комментарий оскорбительным;
    6. является ли комментарий угрозой;
    7. является ли комментарий провокационным;
    8. является ли комментарий нецензурным;
    9. является ли комментарий разжигающим ненависть;
    10. является ли комментарий расистским;
    11. является ли комментарий националистским;
    12. является ли комментарий сексистским;
    13. является ли комментарий гомофобным;
    14. является ли комментарий религиозно-ненанистническим;
    15. является ли комментарий радикалистским.

# Формулирование гипотез

Гипотеза 1: Самые популярные слова в датасете – касающиеся национальности или расы.

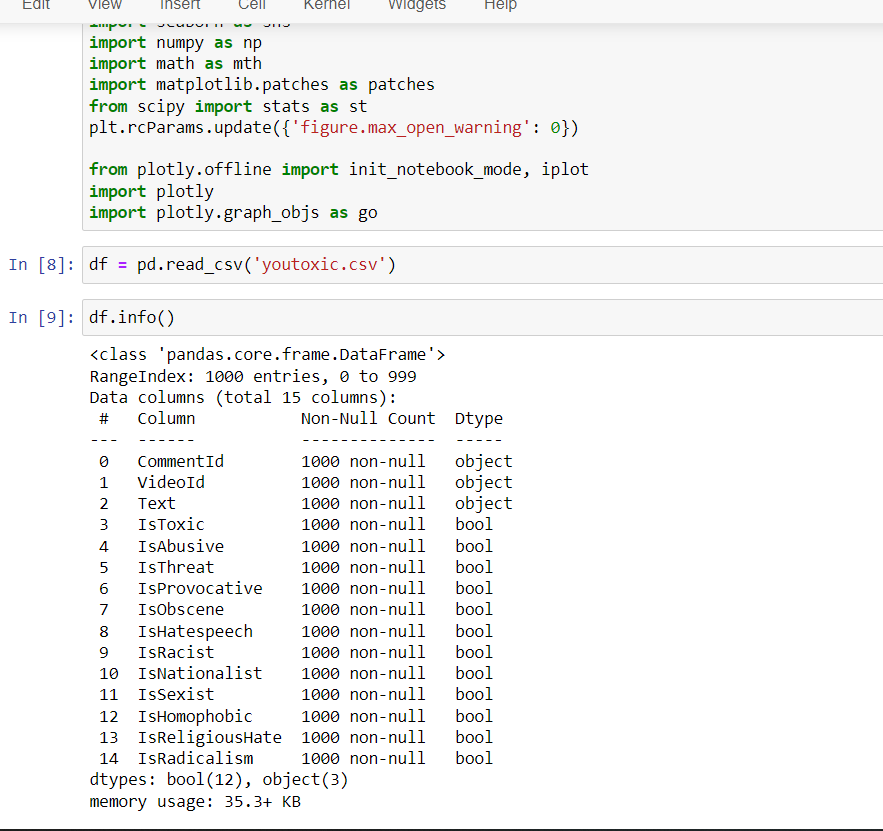
Гипотеза 2: Комментарии выражающие расизм или разжигающие ненависть популярнее прочих.

Гипотеза 3: Самые популярные слова в датасете с меткой токсичности совпадут с самыми популярными в принципе.

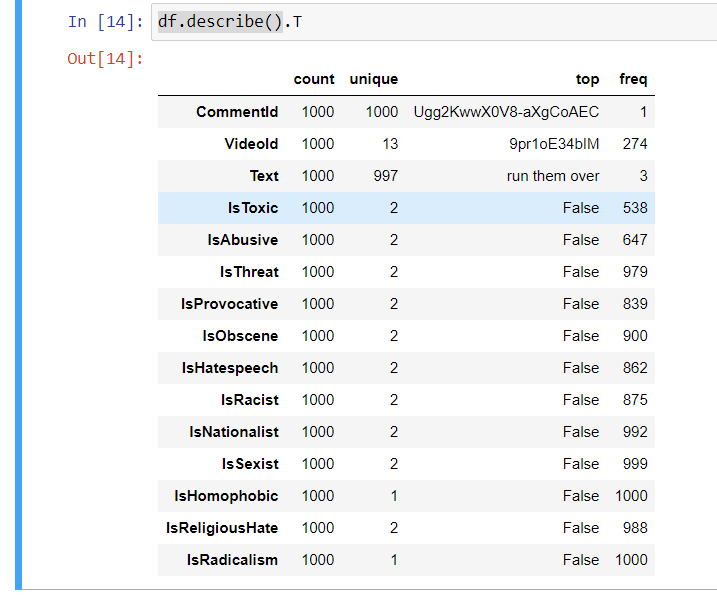
Гипотеза 4: Соотношение токсичных комментариев к нецензурным будет выше остальных.

# Изучение общей информации

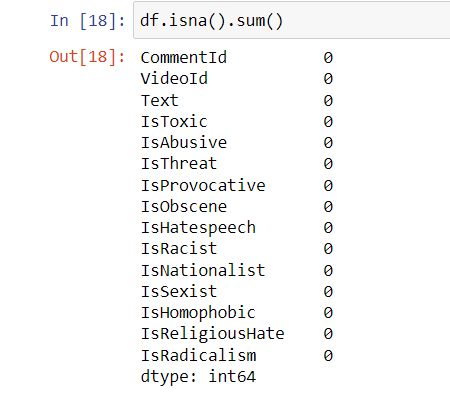
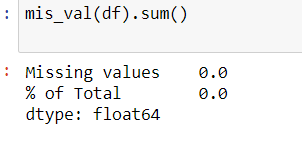
Загружаем датасет и подключаем необходимые библиотеки:



Получим информацию о датасете, на которой увидим использование памяти и количество переменных по их типам, заметим, что используется 35,3 КБ памяти, типов bool используется 12 раз, типов object используется 3 раза, в дальнейшем придется оптимизировать использование памяти для типов данных bool и object:



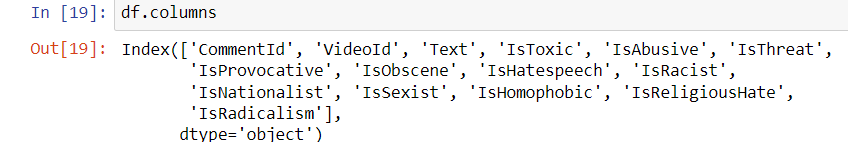
Посчитаем количество пустых значений:

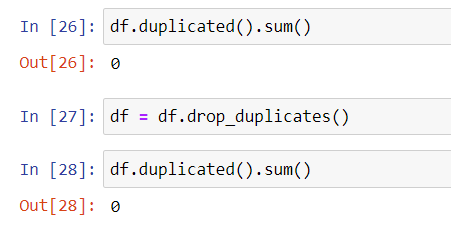
Пустых значений не было обнаружено.

1. **Подготовка данных для работы**

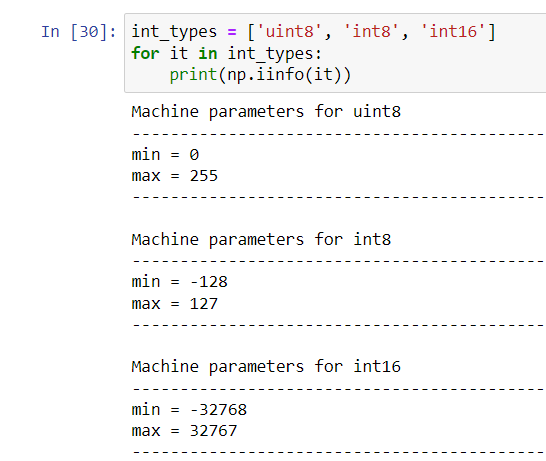
Проверим названия столбцов:

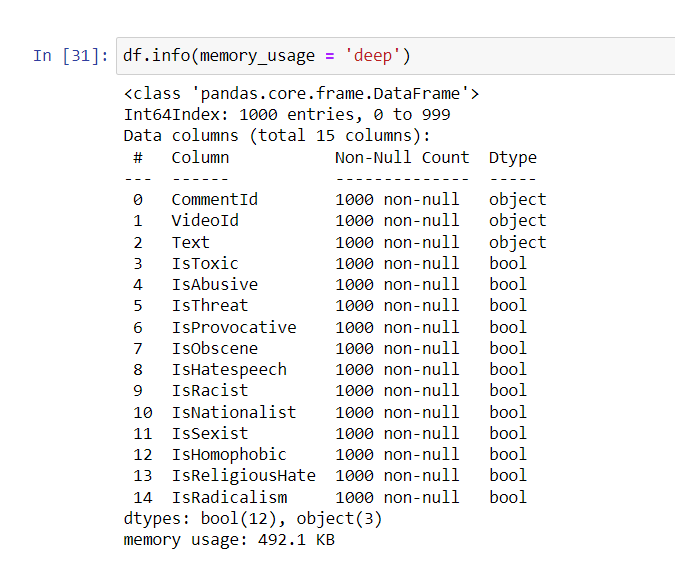


Посчитаем количество дубликатов и удалим их из выборки:

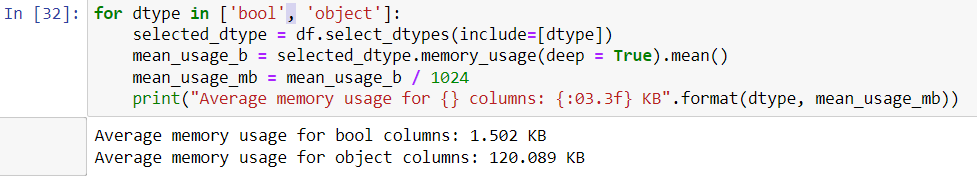
****

Оптимизируем память и изменим типы данных на более корректные:

****

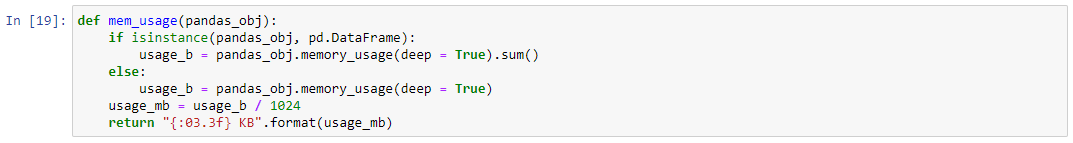
****

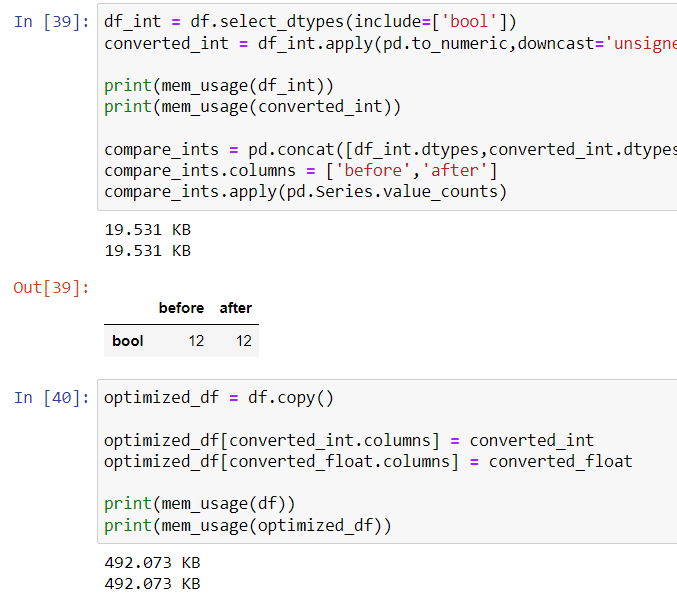
Так как данные разных типов хранятся раздельно, мы исследуем использование памяти разными типами данных. Начнём со среднего показателя использования памяти по разным типам данных:



Эти сведения дают нам понять то, что большая часть памяти уходит на столбцы, хранящих объектные значения.

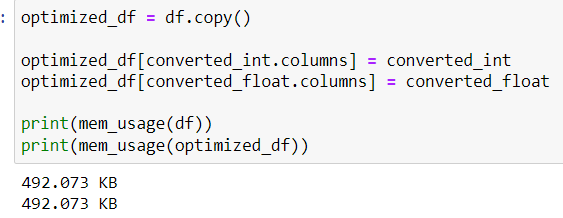
Создадим функцию, которая поможет нам сэкономить немного времени:





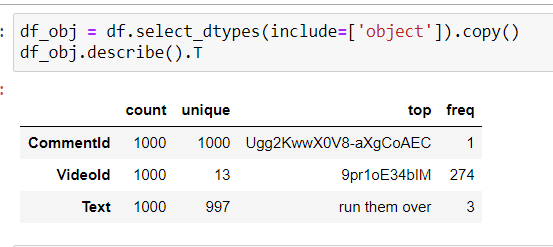
В результате можно видеть, что экономии памяти на полях типа bool добиться не получилось.

Создадим копию исходного объекта Data Frame, используем эти оптимизированные числовые столбцы вместо тех, что присутствовали в нём изначально, и посмотрим на общий показатель использования памяти после оптимизации:

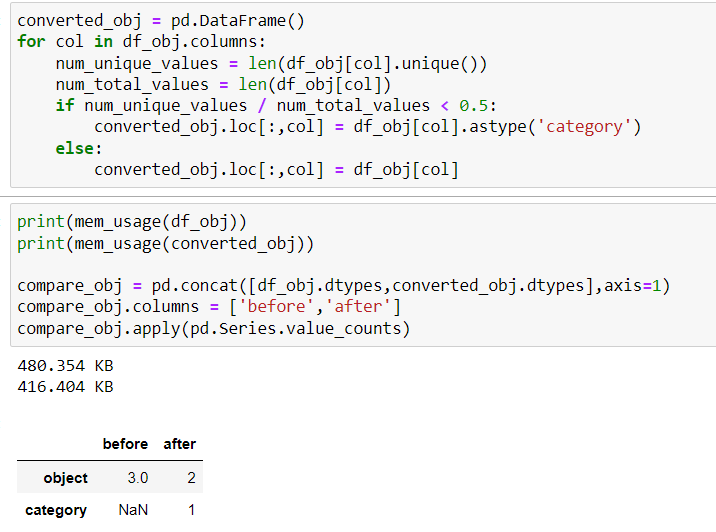


После уменьшения потребления памяти столбцами, хранящими числовые данные, нам удалось уменьшить потребление памяти. Для достижения еще большего результата оптимизируем хранение объектных типов.

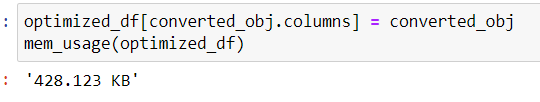
Для того чтобы понять, где именно мы сможем воспользоваться категориальными данными для снижения потребления памяти, выясним количество уникальных значений в столбцах, хранящих значения объектных типов:



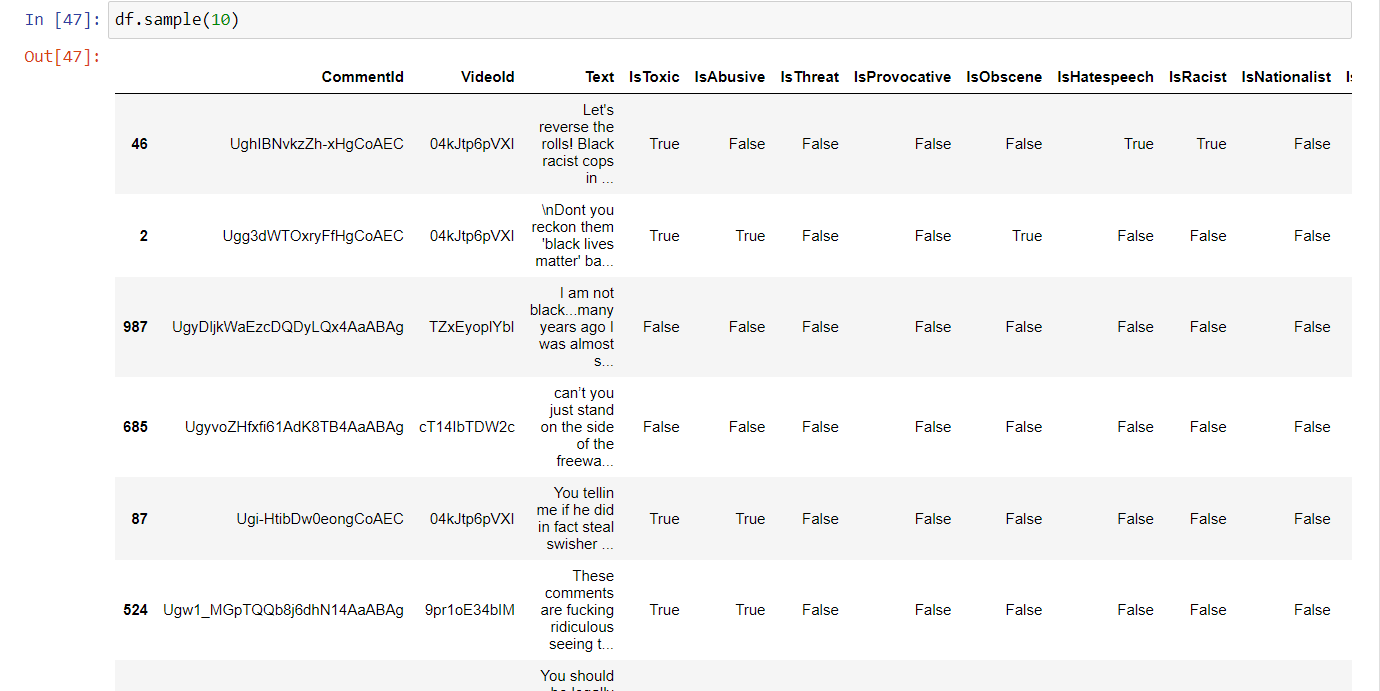
Создадим цикл, который перебирает все столбцы, хранящие данные типа object, выясняет, не превышает ли число уникальных значений в столбцах 50% и, если это так, преобразует их в тип category.:



Теперь сравним то, что получилось после оптимизации, с тем, что было раньше:

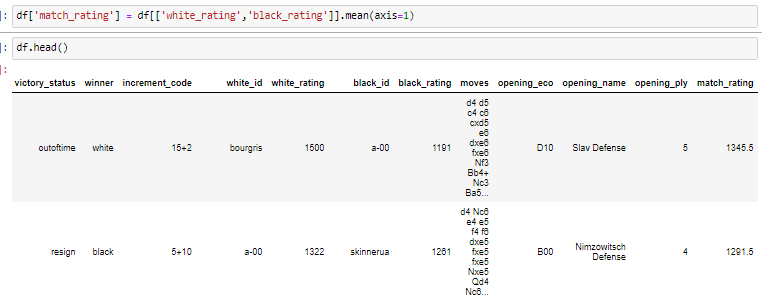


Размер уменьшился почти в два раза (с 480 КБ до 428 КБ).



1. **Агрегирование данных**

Добавим новый столбец: рейтинг партии, который получается вычислением среднего между рейтингом «черных» и «белых».



**Вывод:**

1. Была произведена оптимизация по памяти и адаптация части данных для дальнейшего анализа;
2. Пропущенных значений не имеется;
3. Дубликаты были отсеяны.
4. **Исследовательский анализ данных**
   1. **Самые популярные слова**

****

****

Рисунок 7.1 – Самые популярные слова в датасете

**Вывод:**

По графику на рисунке 7.1 можно заметить, что самыми популярными словами являются слова, касающиеся расовой принадлежности, полиции, а также ругательные выражения, скорее всего данный факт является следствием тем видео, комментарии из под которых попали в датасет.

Гипотеза 1 – подтверждена.

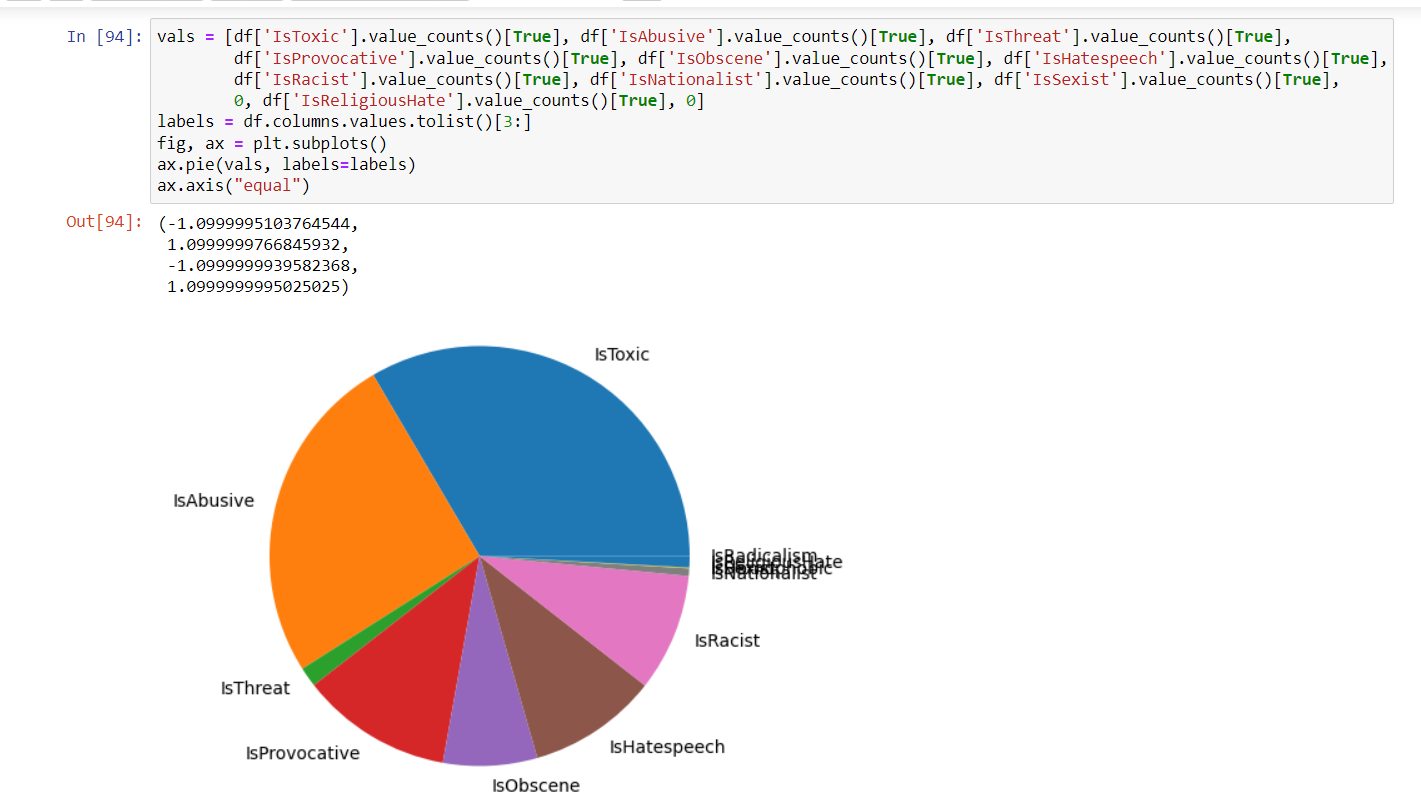
* 1. **Соотношение всех меток к общему числу**

Рисунок 7.2 – Соотношение всех меток

**Вывод:**

Из графиков на рисунке 7.2 видно, что оскорбительные комментарии или выражающие токсичность, популярнее чем выражающие расизм или разжигающие ненависть, хотя можно заметить, что в целом, расистские комментарии далеко не самые малочисленные, также как и матерные.

Гипотеза 2 – опровергнута.

* 1. **Самые популярные слова токсичных комментариев**

****

****

Рисунок 7.3 – Самые популярные слова токсичных комментариев

**Вывод:**

Из графика на рисунке 7.3 видно повторение многих слов приведенных на рисунке 7.1, что говорит о том, что вне зависимости от метки токсичности слова самые популярные слова, в целом, не меняются и касаются все тех же тем, хотя можно заметить небольшое увеличение нецензурных слов.

Гипотеза 3 – подтверждена.

* 1. **Отношение различных меток к метке токсичности**

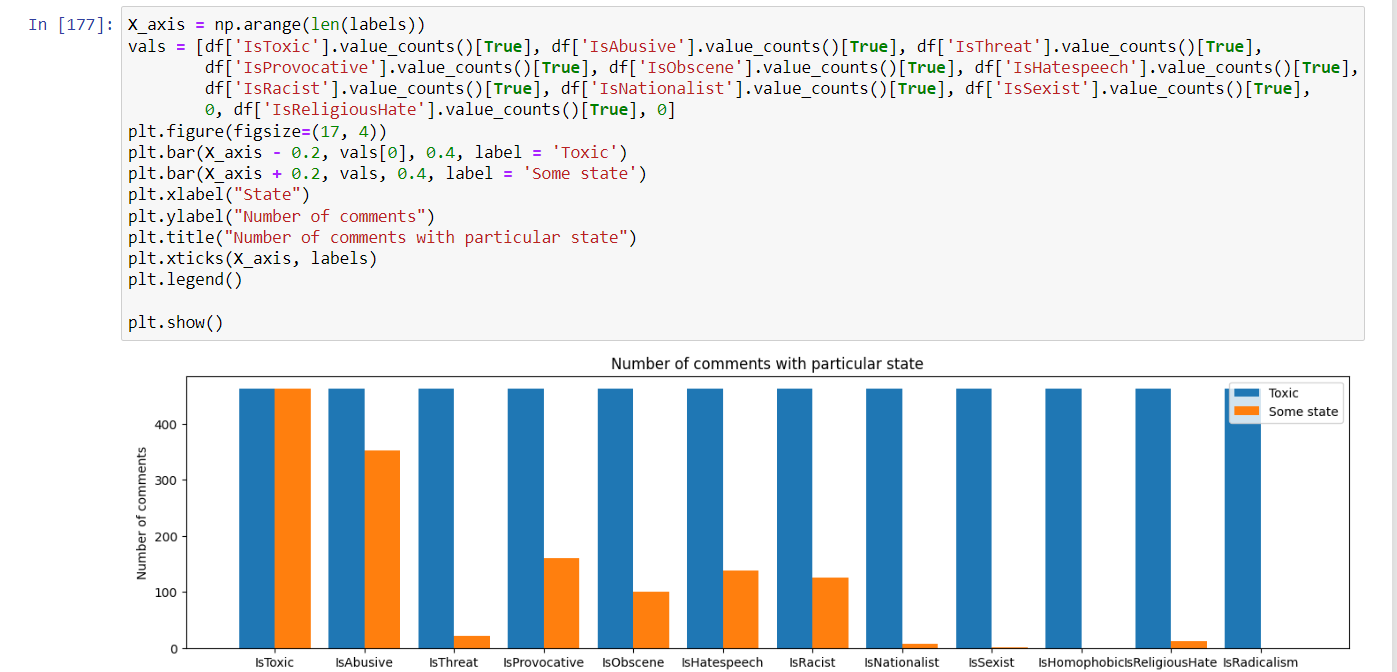


Рисунок 7.4 – Отношение всех меток к метке токсичности

**Вывод:**

Из графика на рисунке 7.4 видно, что наибольшим по сравнению с другими является соотношение токсичных к оскорбительным комментариям, причем разрыв между данным соотношением и прочими многократный.

Гипотеза 4 – опровергнута.

* 1. **Корреляционный анализ**



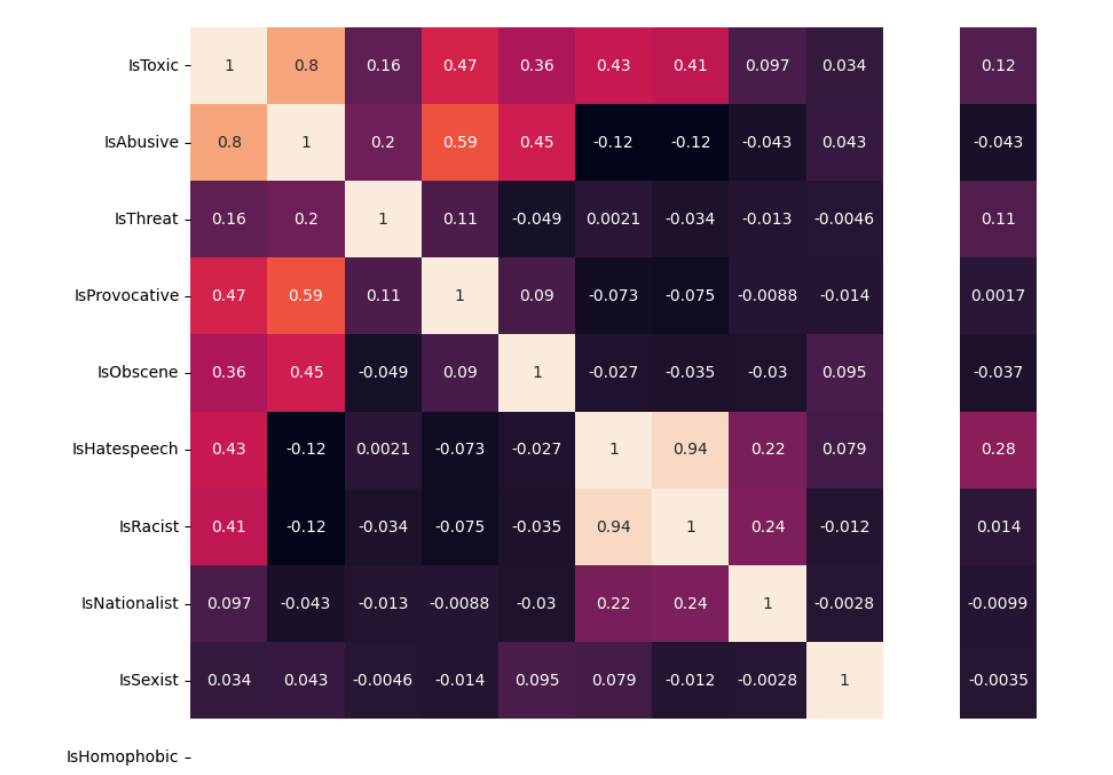


Рисунок 7.5 – Корреляционный анализ

По данной матрице корреляций на рисунке 7.5 можно проследить следующие зависимости:

1. Существенную положительную корреляцию IsToxic и IsAbusive.

# Общий вывод

Перед анализом было оптимизированно использование памяти путем преобразования данных к более корректным типам. Также данные были подвержены предобработке и добавлены необходимые столбцы при помощи агрегирования данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания о природе токсичности на площадке Youtube.

В ходе анализа было выявлено:

1. в связи с выборкой комментариев – самые популярные слова в них касаются национальности или расы;
2. самые популярная метка для комментариев - токсичный;
3. в целом, самые популярные слова в комментариях не меняются вне зависимости от метки токсичности.
4. если комментарией токсичный, то скорее он всего он еще и оскорбительный.

# ЗАКЛЮЧЕНИЕ

В ходе научно-исследовательской работы был проведен анализ датасета токсичных комментариев с Youtube, выявлены важные статистические данные, которые могут улучшить бизнес составляющую контент-мейкинга и найти подходы к уменьшению ненависти в сообществе Youtube.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas» и «Seaborn» на базе языка программирования Python.

# СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Методические указания по программному обеспечению «Pandas»;
2. Методические указания по программному обеспечению «Seaborn»;
3. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.;
4. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»;
5. <https://devpractice.ru>;
6. <https://coderoad.wiki>;
7. <https://www.delftstack.com>;
8. <https://www.machinelearningmastery.ru>;
9. <https://pythonru.com>.