Министерство образования Московской области

Государственное бюджетное профессиональноеобразовательное учреждение Московской области «Физико-технический колледж»

ЗАЩИЩЕНО

ОЦЕНКА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель специальности 09.02.07 Информационные

системы и программирование

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

Идентификация ненормативной лексики с применением машинного обучения

Пояснительная записка к курсовому проекту

**по МДК.05.02**

Руководитель КП Разработал

преподаватель студент группы ИСП 4-5

\_\_\_\_\_\_\_\_\_\_\_Н.М Абидинов

ГБПОУ МО «Физтех-колледж»

\_\_\_\_\_\_\_\_\_\_ Г.В.Базяк \_\_\_\_\_\_\_\_\_\_П.Д Михеева

Долгопрудный, 2024

**Министерство образования Московской области Государственное бюджетное профессиональное образовательное учреждение Московской области «Физико-технический колледж»**

Специальность 09.02.07 **УТВЕРЖДЕНО**

Информационные системы и цикловой комиссией специальности 09.02.07 программирование «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024 года Квалификация «Программист» Руководитель специальности

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

ЗАДАНИЕ

Для курсового проектирования по МДК 05.02 «Перевод с жестового языка» студенту 4 курса группы ИСП 4-5

Абидинова Никиты Манучехровича и Михеева Павла Дмитриевича

(фамилия, имя, отчество)

Тема задания: Идентификация ненормативной лексики

При выполнении курсового проекта на указанную тему должны быть представлены

1. Пояснительная записка

Введение

1 Анализ предметной области и описание технологии проектирования ИС

* 1. Описание предметной области (проблемы проекта)
  2. Определение целевых групп пользователей
  3. Определение ограничений проектного решения

2 Определение требований к системе

* 1. Определение требований к функциональности приложения
  2. Определение требований к графическому интерфейсу

2.3 Техническое задание

3 Этапы разработки

* 1. Сбор и анализ данных
  2. Разработка модели
  3. Разработка интерфейса
  4. Тестирование системы

Заключение

1. Приложение, размещенное на сервере.

Дата выдачи: « » 20 г. Срок окончания: « » 20 г

Руководитель курсового проектирования

Содержание

[1. Анализ предметной области и описание технологии проектирования ИС 7](#_Toc184083826)

[1.1 Описание предметной области (проблемы проекта) 7](#_Toc184083827)

[1.2 Определение целевых групп пользователей 8](#_Toc184083828)

[1.3 Определение ограничений проектного решения 8](#_Toc184083829)

[2. Определение требований к системе 9](#_Toc184083830)

[2.1 Определение требований к функциональности приложения 9](#_Toc184083831)

[2.2 Определение требований к графическому интерфейсу 10](#_Toc184083832)

[2.3 Техническое задание 10](#_Toc184083833)

[3 Этапы разработки 11](#_Toc184083834)

[3.1 Сбор и анализ данных 11](#_Toc184083835)

[3.2 Разработка модели 13](#_Toc184083836)

[3.3 Разработка интерфейса 14](#_Toc184083837)

[3.4 Тестирование системы 14](#_Toc184083838)

[Приложение 16](#_Toc184083839)

[Приложение Б 17](#_Toc184083840)

Введение

Актуальность

В современных цифровых средах, таких как социальные сети, онлайн-форумы, игровые платформы и различные мессенджеры, общение между пользователями происходит на постоянной основе. Однако, наряду с конструктивным взаимодействием, нередко возникает проблема использования некорректной и ненормативной лексики, что может привести к оскорблениям, распространению враждебных высказываний и формированию токсичной среды. Это негативно сказывается на опыте пользователей и может подрывать репутацию платформ, предоставляющих такие сервисы. Важность анализа и обработки текстовых данных становится всё более очевидной. С развитием социальных сетей и онлайн-платформ возрастает потребность в автоматическом выявлении и фильтрации нежелательных высказываний для создания безопасной среды общения. Наша цель — разработать модель, которая позволит автоматически выявлять и блокировать нежелательные высказывания. Такая модель будет анализировать текстовые данные и цензурировать некорректную и ненормативную лексику, что обеспечит безопасную и комфортную среду для пользователей.

Объект и предмет исследования в теме:

Объектом нашего исследования является текстовая информация, содержащая сообщения, комментарии и другие текстовые данные, генерируемые пользователями на различных цифровых платформах (социальные сети, форумы, мессенджеры). Эта информация может включать как корректные, так и некорректные выражения, в том числе ненормативную лексику, оскорбления и другие формы токсичного общения. Предметом исследования являются методы и алгоритмы машинного обучения, используемые для автоматического распознавания и фильтрации некорректной и ненормативной лексики в текстах.

Цель работы: Цель нашей работы — разработка модели на основе машинного обучения для автоматического обнаружения некорректной и ненормативной лексики в текстах, что позволит эффективно фильтровать контент и обеспечивать безопасную среду общения.

Задачи:

1. Анализ требований и сбор данных
2. Предварительная обработка данных
3. Разработка и выбор модели
4. Обучение и тестирование моделей
5. Оценки и оптимизация:
6. Анализ результатов и документация

### Анализ предметной области и описание технологии проектирования ИС

Идентификация ненормативной лексики — это процесс автоматического обнаружения и фильтрации слов или выражений, которые являются неприемлемыми в определенном контексте (например, оскорбительные или грубые слова, нецензурная лексика). Проблема распространения ненормативной лексики является актуальной для множества платформ, включая социальные сети, форумы, мессенджеры и корпоративные коммуникационные системы.

Задача состоит в том, чтобы система автоматически выявляла нежелательные выражения в текстах, помогала блокировать их или уведомлять модераторов. Внедрение машинного обучения в решение этой задачи позволяет значительно повысить точность фильтрации, улучшая способности системы к адаптации к новым словам, фразам или контекстам. Модели машинного обучения могут быть обучены на больших объемах данных, что позволяет эффективно выявлять даже неочевидные случаи ненормативной лексики, такие как оскорбления, завуалированные выражения или нецензурные слова в различных формах.

# 1.1 Описание предметной области (проблемы проекта)

Некорректная лексика является одной из ключевых проблем современных онлайн-платформ, социальных сетей и корпоративных коммуникаций. Она включает в себя не только устаревшие, неуместные и оскорбительные слова, но и другие формы неподобающего выражения, такие как жаргон, ругательства, неуместные сокращения и лексические ошибки. Эти ошибки могут возникать в результате сложных ассоциаций с культурными или социальными контекстами. Некорректная лексика может выражаться через стереотипы, предвзятые взгляды или оскорбления, направленные на определенные социальные или этнические группы. Такие выражения способны ухудшить качество коммуникации, вызвать конфликты и негативно повлиять на имидж платформы или компании. Важно учитывать, что смысл слов может варьироваться в зависимости от контекста, и одно и то же слово может быть интерпретировано по-разному в различных ситуациях. Эта проблема особенно актуальна для сфер, связанных с модерацией контента, такими как социальные сети, платформы пользовательских обзоров, системы корпоративных коммуникаций и даже судебные разбирательства, где корректность языка играет важную роль. Также немаловажное значение имеет отслеживание некорректной лексики в сфере маркетинга, где имидж бренда может пострадать от неправильного выбора слов или высказываний.

# 1.2 Определение целевых групп пользователей

Основными пользователями систем для выявления некорректной лексики могут быть следующие категории: Модераторы контента на онлайн-платформах и в социальных сетях (таких как ВКонтакте, RUTUBE и т.д.), где необходимо автоматически отслеживать токсичные и неподобающие комментарии. Корпоративные коммуникационные платформы, где важно поддерживать профессиональный и уважительный тон общения среди сотрудников. Образовательные платформы, где некорректная лексика может нарушать этические нормы и культуру общения среди студентов и преподавателей. Маркетинговые компании, которые отслеживают обратную связь от пользователей, чтобы предотвратить использование оскорбительных или неуместных выражений. Юридические системы, где важно правильно интерпретировать текстовые сообщения и учитывать все возможные культурные и социальные контексты для принятия справедливых решений.

# 1.3 Определение ограничений проектного решения

Решение по выявлению некорректной лексики сталкивается с рядом ограничений:

Контекст: Одно и то же слово может иметь разные коннотации в зависимости от контекста, и алгоритмы должны уметь различать эти различия.

Многоязычность: При использовании на глобальных платформах необходимо учитывать не только разные языки, но и их культурные и лингвистические особенности.

Технические ограничения: Не все алгоритмы машинного обучения способны эффективно обрабатывать большие объемы данных в реальном времени, особенно если речь идет о сложных задачах синтаксического и семантического анализа.

Этика и цензура: Необходимо учитывать социальные и этические нормы, чтобы избежать чрезмерной цензуры или неуместного подавления контента, что может привести к правовым или репутационным рискам.

# Определение требований к системе

Требования к системе для идентификации ненормативной лексики с применением машинного обучения направлены на создание высокоэффективной, надежной и безопасной системы, которая будет обеспечивать точное и своевременное обнаружение нежелательных выражений в текстах. Внедрение таких решений способствует поддержанию культуры общения и улучшению качества взаимодействия на платформах, где пользователи генерируют контент.

Функциональные требования

Автоматическое обнаружение некорректной и ненормативной лексики: Наша система должна быть способна автоматически анализировать текстовые данные в режиме реального времени, выявляя нежелательные высказывания, включая оскорбления, ненормативную лексику, токсичные комментарии и другие виды агрессивного общения. Мы должны интегрировать механизмы классификации текстов, которые будут оценивать каждый вводимый пользователем текст и определять его корректность.

# 2.1 Определение требований к функциональности приложения

 Обнаружение ненормативной лексики: Система должна автоматически распознавать слова, выражения и фразы, которые считаются ненормативной лексикой. Это может включать:

* Прямые оскорбления и нецензурные слова.
* Завуалированные формы ненормативной лексики.
* Контекстные выражения, которые могут быть оскорбительными в определенных ситуациях.

 Контекстный анализ: Система должна учитывать контекст, в котором используется слово, чтобы исключить ложные срабатывания. Например, слово, которое в одном контексте является нецензурным, может быть нейтральным в другом.

 Реальный анализ в реальном времени: Конфигурация должна обеспечивать обработку текста в реальном времени, анализируя и фильтруя текстовые данные по мере их поступления в систему (например, комментарии, сообщения или чаты).

 Отчеты и уведомления: Система должна генерировать отчеты о найденной ненормативной лексике, а также отправлять уведомления администраторам или модераторам о заблокированных сообщениях.

 Обратная связь и улучшение: Система должна поддерживать механизмы для сбора обратной связи от пользователей и модераторов для улучшения точности модели и минимизации ложных срабатываний.

# 2.2 Определение требований к графическому интерфейсу

Простота и интуитивность: Одной из ключевых задач является разработка удобного и интуитивно понятного интерфейса для администраторов и модераторов, которые будут использовать нашу систему. Мы планируем сделать интерфейс минималистичным, чтобы пользователи могли легко находить необходимые функции и работать с системой без необходимости проходить длительное обучение.

# 2.3 Техническое задание

Целью этой курсовой работы является разработка и анализ системы идентификации ненормативной лексики . Для достижения данной цели необходимо решить следующие задачи:

1. Изучить существующие методы и технологии идентификации ненормативной лексики.
2. Провести анализ существующих систем распознавания и их эффективности.
3. Разработать собственную систему распознавания ненормативной лексики.
4. Оценить эффективность разработанной системы через экспериментальные методы исследования.

# 3 Этапы разработки

## Сбор и анализ данных

Для создания системы, автоматически идентифицирующей ненормативную лексику в текстах, важнейшим этапом является сбор и анализ данных. На основе собранных данных строится модель машинного обучения, способная классифицировать текстовые фрагменты как содержащие или не содержащие ненормативные выражения. Качественные данные являются основой для эффективной работы системы, так как они обеспечивают точность и надежность модели.

Источники данных

Сбор данных для задачи идентификации ненормативной лексики может происходить из различных источников:

1. Открытые текстовые корпуса:
   * Использование доступных наборов данных, содержащих различные типы текстов, включая форумы, социальные сети, комментарии и отзывы. Эти данные могут быть использованы как примеры для обучения, если они включают ненормативную лексику.
2. Ручной сбор данных:
   * В случае, если открытые данные не удовлетворяют требованиям, можно собирать тексты вручную. Например, из форумов, комментариев пользователей, чатов, где встречается ненормативная лексика.
3. Генерация синтетических данных:
   * Использование программных инструментов для создания текстов с ненормативной лексикой, таких как генераторы фраз или шаблонные подходы. Это поможет дополнить тренировочные данные для моделей машинного обучения.

Разметка данных

Для того чтобы модель машинного обучения могла правильно классифицировать текст, необходимо провести разметку данных. Этот этап включает в себя следующие шаги:

1. Типы разметки:
   * Ненормативная лексика: Все фразы или слова, содержащие оскорбительные, грубые или нецензурные выражения, должны быть помечены как "ненормативная лексика".
   * Нормативная лексика: Тексты, не содержащие ненормативных выражений, помечаются как "нормативная лексика".
2. Метод разметки:
   * Для разметки данных можно использовать как ручной, так и автоматический подход. Ручная разметка включает в себя аннотирование текста экспертами, которые определяют, является ли текст ненормативным. Автоматическая разметка может быть выполнена с использованием предварительно обученных моделей, но она потребует дополнительной проверки и корректировки.
3. Использование стандартов:
   * Важно установить единые стандарты для разметки, чтобы избежать субъективности. Например, необходимо четко определить, какие выражения считаются ненормативными, а какие могут быть допустимыми в определенных контекстах.

Предобработка данных

Для того чтобы текст был пригоден для машинного обучения, нужно выполнить несколько шагов по его предобработке:

1. Токенизация:
   * Разбиение текста на отдельные элементы (слова, символы или фразы). Это важно для дальнейшего анализа и обработки текста.
2. Лемматизация и стемминг:
   * Лемматизация и стемминг помогают привести слова к их исходной форме (например, "бегал" → "бегать"), что позволяет уменьшить количество уникальных слов и сделать данные более структурированными.
3. Удаление стоп-слов:
   * Стоп-слова — это часто встречающиеся в языке слова (например, "и", "на", "для"), которые не несут значимой информации для анализа и могут быть удалены.
4. Нормализация текста:
   * Приведение текста к стандартному виду (удаление лишних символов, корректировка ошибок в правописании и т.д.).
5. Обработка сложных случаев:
   * Важно учесть синонимы, эвфемизмы и жаргон, которые могут быть использованы для замены ненормативной лексики (например, "порой" вместо "матерных слов").

Анализ данных

После сбора и предобработки данных необходимо провести их анализ, чтобы извлечь ключевые особенности, которые помогут в обучении модели:

1. Статистический анализ:
   * Оценка частоты появления ненормативной лексики в текстах.
   * Определение наиболее часто используемых ненормативных слов или фраз, а также контекста их использования.
2. Контекстуальный анализ:
   * Изучение контекста, в котором встречаются ненормативные выражения. Это поможет улучшить модель, так как одно и то же слово может быть нецензурным в одном контексте, но вполне безобидным в другом.
3. Понимание структуры текста:
   * Анализ структуры текста (например, наличие диалогов, комментариев или сленговых выражений), что может помочь в более точной классификации.

## 3.2 Разработка модели

**Подготовка обучающих и тестовых данных**:

* После того как данные подготовлены и векторизованы, их следует разделить на обучающую и тестовую выборки. Обучающая выборка используется для тренировки модели, а тестовая — для проверки её качества.
* Обычно данные делятся в пропорции 70-80% для обучения и 20-30% для тестирования.

**Обучение модели**:

* Для каждого выбранного алгоритма машинного обучения проводится обучение на обучающих данных. Это включает настройку гиперпараметров модели, таких как количество деревьев в случайном лесе или количество слоев в нейронной сети.
* Во время обучения модель пытается минимизировать ошибку на обучающих данных, подбирая параметры, которые лучше всего соответствуют классификации.

**Кросс-валидация**:

* Для получения более стабильных результатов рекомендуется использовать кросс-валидацию, чтобы убедиться, что модель не переобучается на обучающих данных и будет хорошо работать на новых данных.

## 3.3 Разработка интерфейса

Использование Streamlit для создания интерфейса для идентификации ненормативной лексики позволяет быстро разработать удобное и интерактивное приложение. Интеграция с моделью машинного обучения даёт возможность пользователям проверить текст на наличие нецензурных выражений в реальном времени. Streamlit предоставляет все необходимые инструменты для создания веб-интерфейсов, и с его помощью можно легко создать интуитивно понятное приложение для различных задач машинного обучения, включая фильтрацию ненормативной лексики. Интуитивно понятный дизайн: Интерфейс спроектирован таким образом, чтобы пользователи могли легко вводить текст для обработки

# 3.4 Тестирование системы

На этапе тестирования системы идентифицирования ненормативной лексики были достигнуты следующие результаты по основным метрикам:

Точность (Accuracy): 0.64

Эти результаты демонстрируют высокую точность и надежность системы

Заключение

В ходе выполнения курсовой работы была разработана система для идентификации ненормативной лексики в текстах с применением методов машинного обучения. Задача заключалась в создании модели, способной автоматически обнаруживать оскорбительные или нецензурные выражения в текстах, что является актуальной проблемой для многих приложений, таких как модерация контента в социальных сетях, форумах и чатах.

В рамках работы была изучена и применена технология машинного обучения, включая выбор подходящих алгоритмов классификации, таких как логистическая регрессия, SVM, и нейронные сети. Для обработки текстовых данных были использованы методы токенизации и векторизации, такие как Bag-of-Words и TF-IDF. В качестве модели для классификации использована предварительно обученная модель, которая была адаптирована под задачу фильтрации ненормативной лексики.

Также был разработан интерфейс для взаимодействия с системой с использованием библиотеки Streamlit. Интерфейс позволил пользователям вводить текст и получать результаты анализа о наличии ненормативной лексики в реальном времени.

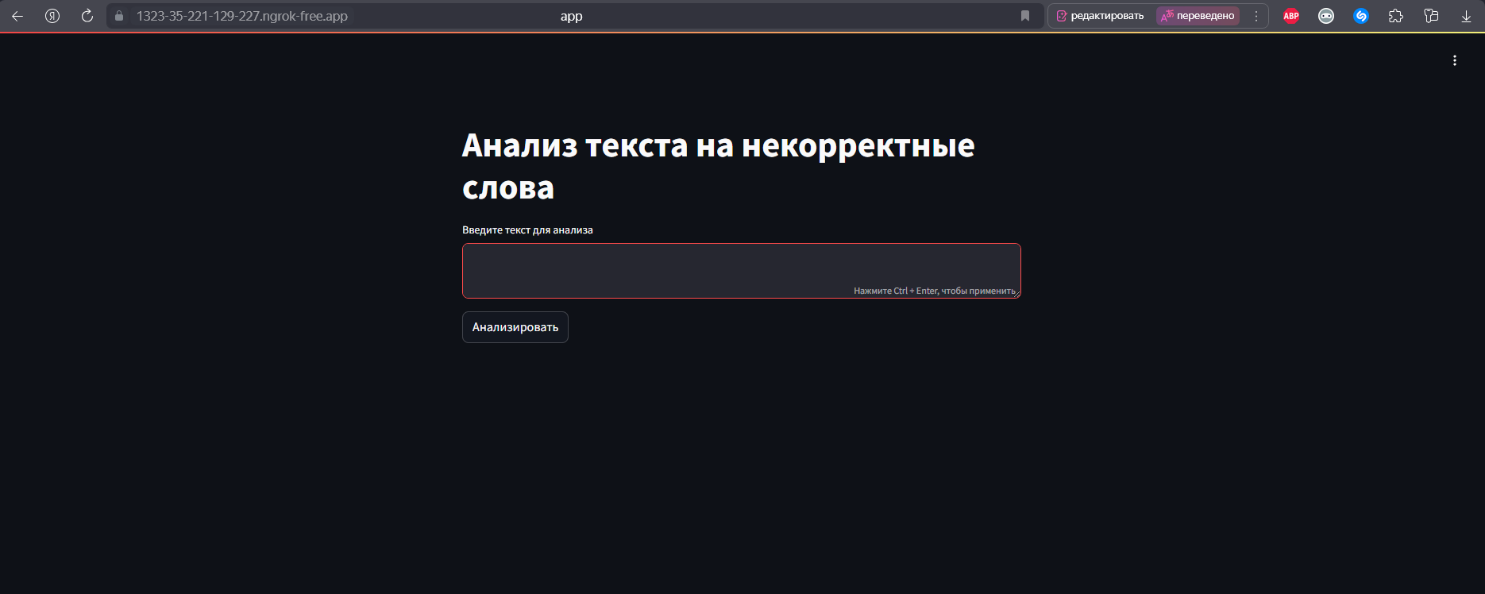
Результаты тестирования показали высокую точность модели в задаче классификации, а также удобство использования интерфейса. Таким образом, система может быть использована для автоматической модерации текста в различных приложениях, что позволяет значительно сократить время, затрачиваемое на ручную проверку контента.

В будущем можно дополнительно улучшить систему, расширив её возможностями для работы с более сложными текстами и многими языками, а также оптимизировав модель для обработки больших объемов данных.

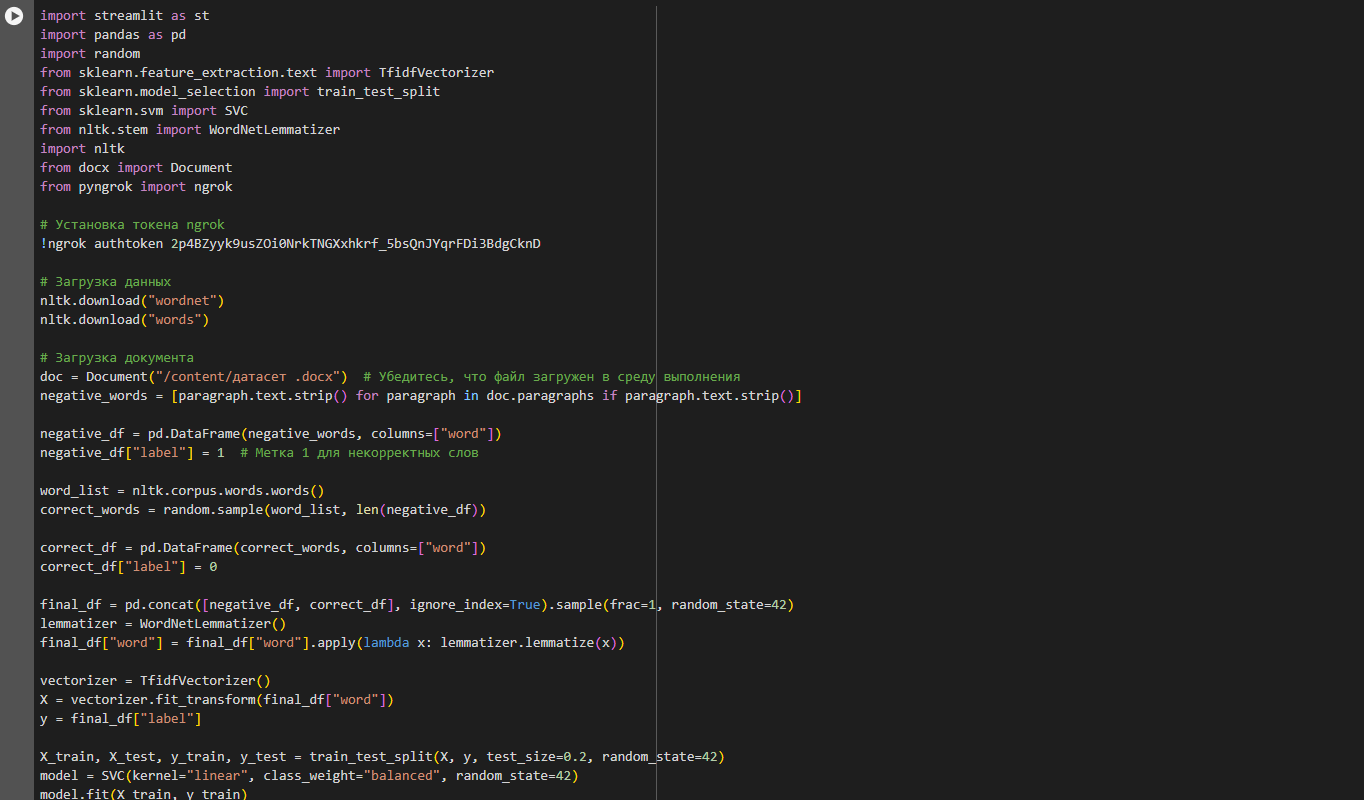
# Список используемых источников

* 1. [Документация Python](https://docs.python.org/3/index.html)
  2. [Документация Streamlit](https://docs.streamlit.io/)
  3. [Документация SVC](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)
  4. [Kaggle](https://www.kaggle.com/datasets)
  5. [Github](https://github.com/)
  6. [Документация NGROK](https://ngrok.com/docs/getting-started/)

#### Приложение А



### Приложение Б



Начало формы

