МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ «ВИТЕБСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ П. М. МАШЕРОВА»

Факультет математики и информационных технологий

Кафедра информационных технологий и управления бизнесом

КУРСОВАЯ РАБОТА

по дисциплине «Дополнительные главы информатики»

ПРИЛОЖЕНИЕ ДЛЯ АНАЛИЗА ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА

.

Царенко Антон Дмитриевич,

3 курс, 34 группа

Научный руководитель:

Новый Вадим Владимирович,

старший преподаватель кафедры прикладного и системного программирования

Витебск, 2023

**Реферат**

Курсовая работа 19 стр., 7 рис., 9 листингов, 5 источников

PYTHON, SPACY, DATA SCIENCE, NLP.

**Объект исследования** – технологии для разработки приложений в области NLP (Natural Language Processing) с использованием Python.

**Предмет исследования** – алгоритмы проектирования и разработки приложения для анализа эмоциональной окраски текста.

**Цель работы** – разработать приложение для анализа эмоциональной окраски текста на примере рецензий на фильмы.

**Задачи**:

1. Выполнить анализ открытых источников информации, литературы в сфере разработки приложений в области NLP.
2. Ознакомиться с особенностями реализации приложения с использованием языка программирования Python.
3. Подготовить план разработки приложения.
4. Разработать приложение.

**Методы исследования** – изучение документации, материалов научных и периодических изданий по теме исследования, существующих систем, общенаучные методы исследования (описание, анализ и т.д.).

**Теоретическая значимость и элементы новизны**: NLP имеет огромную значимость в современном мире, поскольку обработка естественного языка является неотъемлемой частью многих аспектов нашей жизни и большинства сфер деятельности.

**Практическая значимость**: в настоящее время NLP находит широкое применение в различных областях, включая компьютерную лингвистику, информационный поиск, анализ социальных медиа, обработку естественной речи и т.д.

Содержание

[Введение 4](#_Toc136851638)

[1. ОБОСНОВАНИЕ ВЫБОРА СРЕДСТВ РЕАЛИЗАЦИИ И ЕГО КРАТКАЯ ХАРАКТЕРИСТИКА 6](#_Toc136851639)

[2. РАЗРАБОТКА ПРИЛОЖЕНИЯ 8](#_Toc136851640)

[2.1 ПОДГОТОВКА ТЕКСТОВЫХ ДАННЫХ 8](#_Toc136851641)

[2.2 ОБУЧЕНИЕ МОДЕЛИ ДЛЯ АНАЛИЗА ЭМОЦИОНАЛЬНОЙ ОКРАСКИ 10](#_Toc136851642)

[Заключение 18](#_Toc136851643)

[Список использованных источников 19](#_Toc136851644)

Введение

NLP (Natural Language Processing) – это область искусственного интеллекта (ИИ), которая занимается изучением и разработкой алгоритмов и методов для обработки и анализа естественного человеческого языка. NLP позволяет компьютерам понимать, интерпретировать и генерировать текст на естественном языке, что делает возможным множество приложений, таких как перевод текста, анализ настроений, автоматическое составление резюме текстов и даже создание искусственных интеллектов, способных общаться с людьми на естественном языке.

Для решения задач в NLP используются методы машинного обучения, статистические модели, нейронные сети, а также лингвистические знания и правила. NLP находит широкое применение в различных областях, включая компьютерную лингвистику, информационный поиск, анализ социальных медиа, обработку естественной речи и т.д.

Анализ тональности – инструмент компьютерной лингвистики, оценивающий такую субъективную составляющую текста, как отношение пишущего.

**Цель работы** – разработать приложение для анализа эмоциональной окраски текста на примере рецензий на фильмы.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Выполнить анализ открытых источников информации, литературы в сфере создания приложений в области NLP.
2. Ознакомиться с особенностями разработки приложения с использованием языка Python.
3. Составить план разработки приложения и подготовить необходимые шаблоны и библиотеки.

**Методы исследования** – изучение документации, материалов научных и периодических изданий по теме исследования, существующих систем разработки редакторов, общенаучные методы исследования (описание, анализ, классификация, сравнение, аналогия).

!!!Работа включает введение, 2 главы, заключение, список использованных источников. В первой главе предоставлено обоснование выбора средства и платформы для разработки, а также языка программирования. Во второй главе приводится общая характеристика, функционал приложения и описание этапов разработки.

# ОБОСНОВАНИЕ ВЫБОРА СРЕДСТВ РЕАЛИЗАЦИИ И ЕГО КРАТКАЯ ХАРАКТЕРИСТИКА

Python – это высокоуровневый язык программирования, который широко используется в различных областях, включая веб-разработку, научные вычисления, анализ данных, искусственный интеллект и многое другое. Python был создан в 1991 году и с тех пор стал одним из самых популярных языков программирования в мире.

Python – это язык программирования с множеством преимуществ, вот некоторые из них:

1. Простота и читаемость кода: Python имеет простой и понятный синтаксис, что делает его легким для изучения и использования. Код на Python легко читать и понимать, даже для новичков.
2. Большое количество библиотек и модулей: Python имеет множество библиотек и модулей, которые позволяют решать различные задачи, такие как научные вычисления, анализ данных, искусственный интеллект и многое другое. Это позволяет программистам быстро и удобно решать задачи без необходимости писать все с нуля.
3. Переносимость кода: Python является переносимым языком, что означает, что код, написанный на одной операционной системе, может быть запущен на другой операционной системе без необходимости изменений.
4. Большое сообщество: Python имеет большое и активное сообщество разработчиков, которые создают новые библиотеки, модули и фреймворки, а также предоставляют поддержку и помощь другим разработчикам.
5. Широкое применение: Python используется в различных областях, включая веб-разработку, научные вычисления, анализ данных, искусственный интеллект и многое другое. Это делает его универсальным языком программирования, который может быть использован для решения различных задач.

Для решения задачи анализа тональности текста будет использована библиотека spaCy, предназначенная для обработки естественного языка. Она предоставляет эффективные и точные инструменты для обработки и анализа текстов на естественном языке.

Некоторые возможности spaCy:

1. Разбор текста на токены, предложения, сущности и другие элементы.
2. Построение зависимостей между словами в предложении.
3. Определение частей речи и синтаксических ролей слов в предложении.
4. Обработка текстов на разных языках.

# РАЗРАБОТКА ПРИЛОЖЕНИЯ

* 1. **ПОДГОТОВКА ТЕКСТОВЫХ ДАННЫХ**

Перед использованием текстовых данных, их нужно предварительно подготовить. Обработка включает следующие этапы:

* + - Токенизация текста

Токенизация текста – это процесс разбиения текста на отдельные слова или токены. Рассмотрим процесс токенизации на примере следующего текста:

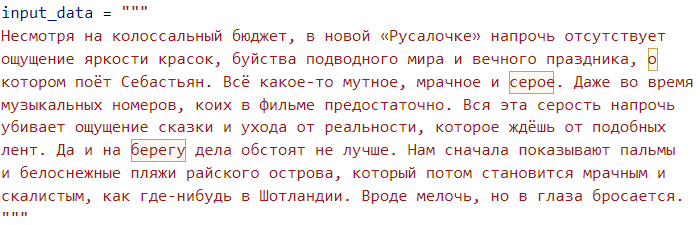


Рисунок 1 – текст до токенизации.

Текст будем разбивать по отдельным словам. В результате получим следующий массив:

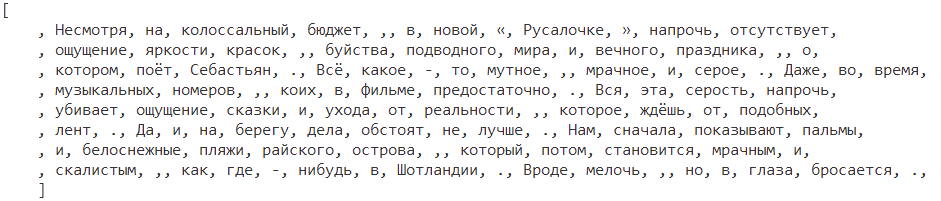


Рисунок 2 – текст после токенизации.

* + - Удаление стоп-слов

Стоп-слова – это слова, которые обычно не имеют смысловой нагрузки в тексте и, следовательно, могут быть исключены из анализа. К стоп-словам относятся артикли, союзы, предлоги и т.д.

Удалим стоп-слова из нашего списка токенов. Библиотека spyCy содержит встроенный список стоп-слов для различных языков, что можно использовать для их удаления.

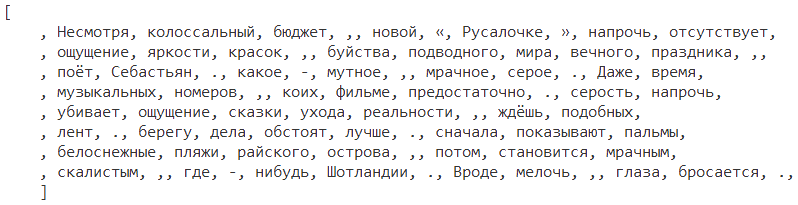


Рисунок 3 – список токенов после удаления стоп-слов.

* + - Приведение к нормальной форме

Процесс нормализации заключается в том, что все формы слова приводятся к единому представлению. В нашем случае все формы слова будут связываться с его простейшей формой – леммой.

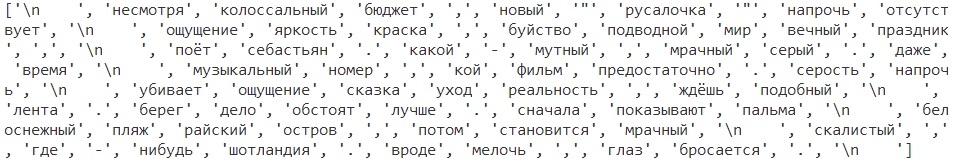


Рисунок 4 – список токенов после приведения к нормальной форме.

* + - Векторизация текста

Векторизация текста – это процесс преобразования текстовых данных в числовые векторы, которые могут быть использованы в алгоритмах машинного обучения. В spaCy используется алгоритм Word2Vec, который используется для создания векторных представлений слов на основе контекста, в котором они встречаются. Алгоритм Word2Vec использует нейронные сети для обучения эмбеддингов слов, которые позволяют представлять слова в виде числовых векторов, учитывая их семантическое значение. Векторы слов, полученные с помощью Word2Vec, обладают свойством семантической близости, то есть слова, близкие по значению, имеют близкие векторные представления.

* 1. **ОБУЧЕНИЕ МОДЕЛИ ДЛЯ АНАЛИЗА ЭМОЦИОНАЛЬНОЙ ОКРАСКИ**

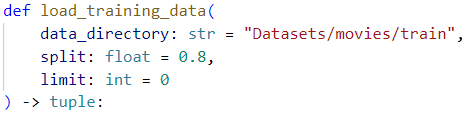
Разобьем задачу на следующие шаги:

* + - Загрузка данных

В качестве набора данных, на которых будет происходить обучение и проверка модели будет использован набор с положительными и отрицательными отзывами о фильмах. Данные поделены на обучающий набор (используется для обучения модели), и тестовый набор (включает данные, позволяющие судить о конечном качестве работы модели).

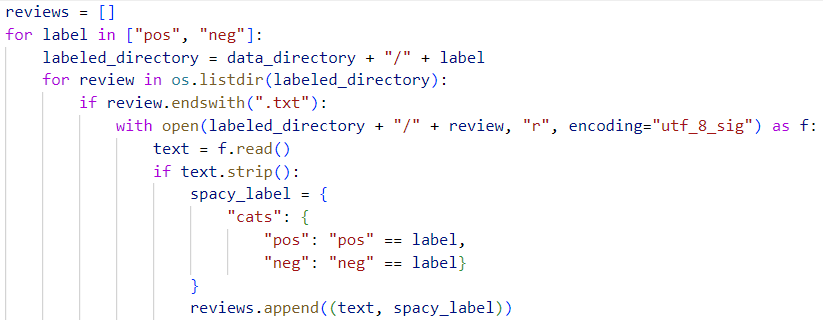
Для загрузки обучающих и тестовых данных создадим функцию load\_training\_data:

Листинг 2.1 – функция load\_training\_data



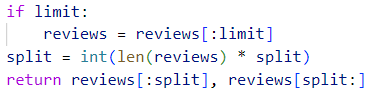
В нее передается название каталога с данными, соотношение между обучающими и тестовыми данными и количество отбираемых записей.

Листинг 2.2 – перебор всех файлов и загрузка их содержимого в список



Затем идет выбор определенного количества записей и их разбиение на 2 набора:

Листинг 2.3 – разбиение данных на 2 набора



* + - Обучение классификатора

Конвейер spaCy позволяет создать и обучить сверточную нейронную сеть (CNN) для классификации текстовых данных. Один из встроенных компонентов конвейера называется textcat (сокр. TextCategorizer), который позволяет назначать текстовым данным категории и использовать их в качестве обучающих данных нейронной сети.

Для обучения модели создадим функцию train\_model:

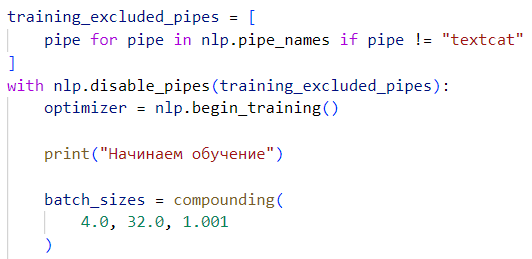
Листинг 2.4 – функция train\_model



Для начала загружается встроенный конвейер, затем в нем проверяется наличие компонента textcat. В случае его отсутствия он создается и добавляется в конвейер. Затем в textcat добавляются имена категорий – pos для положительных отзывов и neg для отрицательных.

Следующим шагом идет написание цикла обучения textcat. Чтобы начать цикл обучения, настраиваем конвейер на обучение компонента textcat, генерируем для него пакеты данных с помощью функций из пакета spacy.util – minibatch() и compounding(). Под пакетом данных или батчем понимается просто та небольшая часть данных, которая участвует в обучении. Пакетная обработка данных позволяет сократить объем памяти, используемый во время обучения и быстрее обновлять гиперпараметры.

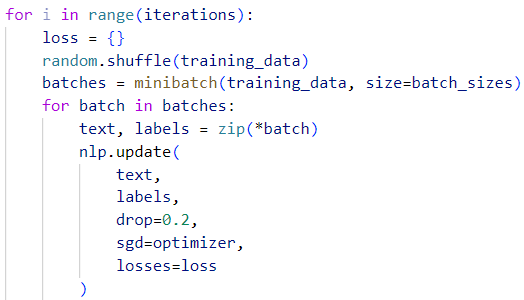
Листинг 2.5 – начало обучения



Для начала отключаем в конвейере все компоненты, кроме textcat. Затем с помощью функции begin\_training получаем начальную функцию оптимизатора, которая будет использована для обновления весов базовой модели. Функция batch\_sizes возвращает последовательность размеров батчей.

Затем идет непосредственно цикл обучения:

Листинг 2.6 – начало цикла обучения



Сначала обучающие данные перемешиваются с помощью функции shuffle. Затем данные разбиваются на батчи с использованием функции minibatch. В каждом батче текст отделяется от меток (категорий), затем идет вызов оптимизатора update, что фактически запускает обучение.

Для наблюдения прогресса обучения модели напишем отдельную функцию evaluate\_model. В этой функции мы будем классифицировать тексты из валидационного набора данных на недообученной модели и сравнивать результаты модели с метками исходных данных.

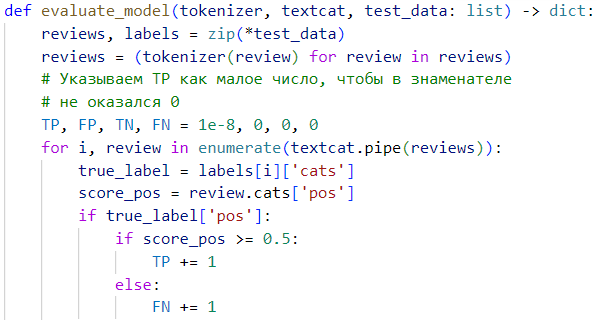
Будем высчитывать следующие данные:

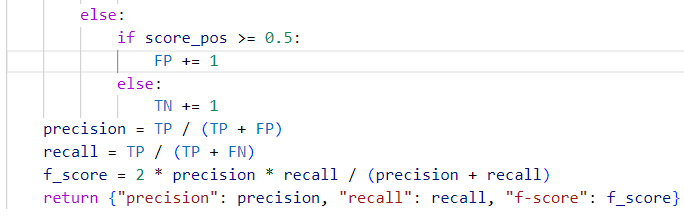
* + - Истинно положительные – число отзывов, которые модель правильно предсказала как положительные.
    - Ложно положительные – число отзывов, которые модель неверно предсказала как положительные.
    - Истинно отрицательные – число отзывов, которые модель правильно предсказала как отрицательные.
    - Ложно отрицательные - число отзывов, которые модель неверно предсказала как отрицательные.

На основе четырех описанных статистических данных вычисляются две метрики: точность и полнота. Эти метрики являются показателями эффективности модели классификации. Точность (precision) – отношение истинно положительных результатов ко всем элементам, отмеченным моделью как положительные. Полнота (recall) – это отношение истинно положительных отзывов ко всем фактическим положительным отзывам:

Также популярной метрикой является F1-мера – среднее гармоническое точности и полноты. Максимизация F1-меры приводит к одновременной максимизации этих двух критериев:

Листинг 2.7 – функция evaluate\_model

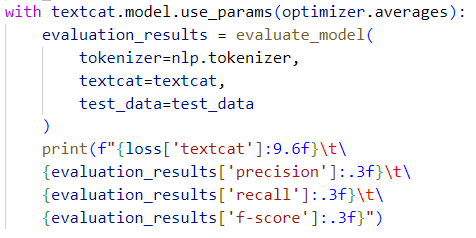




В этой функции сначала разделяются обзоры и их метки, затем с помощью функции tokenizer каждый обзор токенизируется для передачи в textcat. Затем идет подсчет ложных и истинных срабатываний и модели и расчет точности, полноты и F-меры.

Организуем вывод данных метрик в цикле обучения для оценки модели:

Листинг 2.8 – вывод метрик



После обучения модели сохраним ее в каталоге model:

Листинг 2.9 – сохранение модели



Наконец, можно произвести обучение модели:

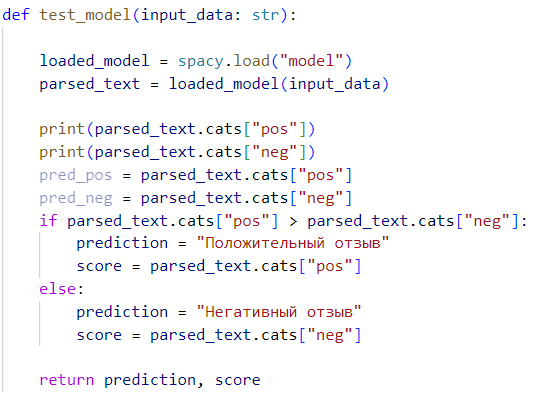


Рисунок 5 – результаты обучения.

* + - Тест модели

Создадим функцию test\_model для оценки заданного текста:

Листинг 2.9 – функция test\_model



Сначала загружаем модель, затем передаем в нее текст. Затем сравниваем оценки каждого настроения и в зависимости от этого выдаем ответ.

С помощью библиотеки tkinter создадим простое приложение с полем для ввода текста:

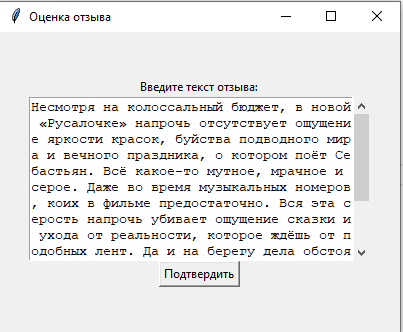


Рисунок 6 – окно приложения.

После нажатия на кнопку получаем ответ:

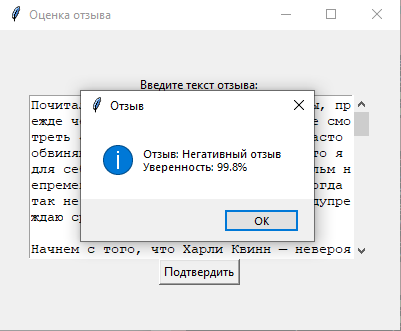


Рисунок 7 – вывод результата.

Заключение

В данной курсовой работе выполнена разработка приложения для оценки эмоциональной окраски текста на примере отзывов о фильмах.

В результате проделанной работы были решены следующие задачи:

1. Проанализированы открытые источники информации, литература по теме курсовой работы.
2. Изучены особенности создания приложений с использованием языка Python.
3. Подготовлен план разработки приложения.
4. Реализовано приложение.

Созданное приложение с заданным набором функций свидетельствует о выполнении поставленных задач и о том, что цель работы достигнута.

Список использованных источников

1. Projects – [электронный ресурс] – Mode of access: <https://spacy.io/usage/projects#textcat> – Date of access: 30.05.2023
2. Семантический анализ для автоматической обработки естественного языка – [электронный ресурс] – Mode of access: https://rdc.grfc.ru/2021/09/semantic\_analysis/#post-1707-\_Toc69397631 – Date of access: 30.05.2023
3. Анализ тональности текстов с помощью сверточных нейронный сетей – [электронный ресурс] – Mode of access: https://habr.com/ru/companies/vk/articles/417767/ – Date of access: 30.05.2023
4. Анализ тональности в русскоязычных текстах – [электронный ресурс] – Mode of access: https://habr.com/ru/companies/vk/articles/516214/ – Date of access: 30.05.2023
5. Что такое анализ эмоциональной окраски? – [электронный ресурс] – Mode of access: https://aws.amazon.com/ru/what-is/sentiment-analysis/ – Date of access: 30.05.2023