Проект для студентов

«Персональный помощник для студентов»

“Автоматизация создания конспектов и краткого содержаний лекций и учебных материалов”

Выполнил:

Илья Берлов

1. Введение:

Цель проекта: Разработка системы автоматического создания кратких содержаний для лекций и учебных материалов с использованием методов машинного обучения и обработки естественного языка (NLP).

Основные задачи:

- Написание кода для создания краткого содержания, конспекта лекций
- Обучение модели на размеченном корпусе лекций, учебных материалов для выделения ключевых фраз и предложений.

2. Анализ проблемы:

Текущая ситуация: Студенты сталкиваются с проблемой неэффективного использования времени при подготовке и изучении лекций.

Проблемы и ограничения:

- Сложности студентов в обработке больших объемов информации.
- Необходимость вручную создавать краткие выжимки из лекций, учебных материалов.

3. Описание решения:

Концепция:

- Использование методов NLP для выделения ключевых фраз и предложений.
- Использование методов машинного обучения для выделения ключевых тем лекций

План реализации:

Основные моменты:

1. Сбор и подготовка размеченного корпуса лекций для обучения модели.

В качестве корпуса лекций используем материалы учебных пособий:

Лимановская, О.В. Основы машинного обучения : учебное пособие / О.В. Лимановская, Т.И. Алферьева ; Мин-во науки и высш. образования РФ.— Екатеринбург : Изд-во Урал. ун-та, 2020. — 88 с (Источник: elar.urfu.ru/bitstream/10995/88687/1/978-5-7996-3015-7_2020.pdf?ysclid=lpsf4sevqt806202918)

Канивец, Е.К. Л19 Лекции по информатике: учебное пособие / сост. Е.К. Канивец; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2018. –180 с.

Канивец.pdf (osu.ru)

Обучение модели с использованием библиотек, таких как NLTK и sickit-learn.

Неделя 1: Подготовка и Планирование

Исследование Требований Проекта:
- Подробное изучение требований проекта.
- Определение основных функциональных и технических требований.

Планирование Этапов:
- Разработка детального плана реализации, включая этапы работы и распределение задач.

Неделя 2: Сбор и Подготовка Данных

Поиск и Выбор Корпуса Лекций:

- Исследование доступных корпусов лекций.
- Выбор подходящего корпуса для обучения модели.

Предварительная Обработка Данных:

- Разработка скриптов для предварительной обработки текста лекций.
- Очистка данных от ненужных символов и форматирование.

Неделя 3: Разработка и Обучение Модели

Неделя 3: Разработка и Обучение Модели

3.1 Написание алгоритма для создания кратких содержаний лекций на базе библиотеки NLTK

Для создания кратких содержаний лекций используется библиотека Natural Language Toolkit (NLTK). NLTK предоставляет множество инструментов для обработки текста, включая токенизацию, определение частей речи и выделение ключевых слов.

Алгоритм включает следующие шаги:

Предварительная обработка текста: Лекции подвергаются предварительной обработке, включающей удаление ненужных символов и форматирование текста.

Токенизация: Используя NLTK, текст разбивается на отдельные слова (токены).

Удаление стоп-слов: Из текста удаляются стоп-слова (часто встречающиеся, но не несущие смысловой нагрузки).

Вычисление частот: Для каждого слова подсчитывается его частота в тексте.

Определение весов предложений: Каждому предложению присваивается вес на основе частот слов, входящих в него.

Пороговая фильтрация: Предложения, вес которых превышает установленный порог, выбираются для включения в краткое содержание.

3.2 Реализация Модели LDA

Для более глубокого анализа тематик лекций используется модель LDA (Latent Dirichlet Allocation). Этот метод машинного обучения помогает выделить ключевые темы в коллекции текстов.

Алгоритм включает следующие шаги:

Создание словаря и корпуса: Используя библиотеку NLTK, строится словарь слов и их частот в корпусе лекций. Каждая лекция представляется в виде вектора, где каждая компонента — это количество вхождений слова из словаря в данную лекцию.

Реализация LDA: С использованием библиотеки scikit-learn, модель LDA обучается на полученном корпусе. Этот процесс помогает выделить скрытые темы в тексте лекций.

Определение ключевых слов тем: Для каждой выделенной темы определяются ключевые слова, что обеспечивает более понятное представление содержания лекции.

Присвоение тем лекциям: Каждой лекции присваивается тема на основе ее содержания.

Таким образом, комбинированное использование NLTK для создания кратких содержаний и LDA для анализа тематик лекций обеспечивает более глубокое и информативное извлечение ключевой информации из учебных материалов.

Дополнительные Задачи:

- Оценка Эффективности:
    - Создание метрик и инструментов для оценки эффективности созданных кратких содержаний.
- Тестирование и Оптимизация:
    - Тестирование проекта на различных корпусах лекций.

- Оптимизация алгоритмов для повышения качества и скорости работы.

Распределение Задач:

- Я: Исследование, планирование, реализация LDA модели.
- Коллега 1: Поиск и подготовка данных, предварительная обработка текста лекций.
- Коллега 2: Интеграция Gensim для суммаризации, тестирование и оптимизация.

Ресурсы:

- Вычислительные Ресурсы: Google Cloud, Jupyter Lab или локальные вычислительные мощности для обучения модели.
- Корпусы Лекций: Ресурсы для доступа к корпусам лекций.
- Библиотеки: NLTK, sickit-learn и другие необходимые библиотеки для Python.

Технологии и инструменты:

- Python с использованием библиотек NLTK, sickit-learn.
- Методы машинного обучения для обучения модели.
- Алгоритмы кластеризации для группировки лекций.

4. Практическая ценность и применимость:

- Экономия времени: Автоматизация создания кратких содержаний ускорит процесс подготовки лекций и обучения.
- Легкость ориентации: Студенты смогут легче ориентироваться в материалах лекций благодаря кратким выжимкам.
- Повышение доступности образования: Создание кратких содержаний полезно для студентов с ограниченным временем или для повторения материалов.

5. Команда и план действий:

Члены команды:

- Проект менеджер и аналитик.

- Специалист по обработке текста и NLP.
- Инженер по машинному обучению.
- Разработчик для реализации алгоритмов и создания интерфейса.

План реализации:

Декомпозиция задач и определение ролей в команде.
Сбор данных и их предварительная обработка.
Обучение модели на подготовленных данных.
Реализация алгоритмов кластеризации.
Внедрение метрик и оценка качества созданных кратких содержаний.
Создание интерфейса для использования системы.

6. Заключение:

- Достижения и преимущества: Автоматическое создание кратких содержаний сэкономит время, облегчит процесс обучения и подготовки материалов для лекций.
- Практическая ценность: Проект обеспечит более эффективное использование времени преподавателей и обучаемых, улучшив качество образования.
- Потенциал для улучшения: Дальнейшее развитие может включать работу с мультимедийными данными и дополнительные алгоритмы NLP для улучшения точности анализа текста лекций.