МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Костромской государственный университет» (КГУ)
Институт Высшая IT-школа
Кафедра информационных систем и технологий
Направление подготовки/Специальность* 09.03.02
Информационные системы и технологии
Профиль Разработка программного обеспечения информационных систем
Дисциплина Моделирование процессов и систем

МЕЖДИСЦИПЛИНАРНЫЙ ПРОЕКТ

Тема:

«Портрет» абитуриента

Выполнили студенты Смирнов Кирилл Андреевич
Шкунов Владимир Викторович
Группа 22-ИСбо-2а
Проверила Кириллова Екатерина Сергеевна
Оценка _____
Подпись преподавателя _____

Кострома 2024

Содержание

Анализ задачи

Наша задачаразработать ПО, которое позволит получить “портрет” абитуриента на основе его страницы в социальной сети ВКонтакте, то естьузнать его интересы и хобби, составить приблизительный круг общения и вывести его личностные характеристики. В результате анализа данной задачи, удалось составить план работы:

1. Сбор данных

- Идентификация абитуриента: Определяем, какие данные будут собираться (например, профиль абитуриента, его посты, лайки, комментарии, подписки на группы и т.д.).
- Получение данных: Использовать API ВКонтакте для сбора необходимых данных.

2. Обработка данных

- Очистка данных: Удаляем дубликаты, нерелевантные данные и шум.
- Форматирование данных: Приводим данные к удобному для анализа формату (например, CSV, JSON).

3. Анализ данных

Текстовый анализ:

- Топик-моделирование: Определяем основные темы, которые интересуют абитуриента.
- Сентимент-анализ: Оцениваем эмоциональную окраску постов и комментариев.
- Частотный анализ: Выявляем наиболее часто упоминаемые слова и фразы.

Социальный анализ:

- Анализ друзей и подписок: Определите круг общения абитуриента и его интересы.
- Анализ активности: Оцените, насколько активен абитуриент в социальной сети (частота постов, лайков, комментариев).

4. Интерпретация результатов

Создание портрета:

- Личностные характеристики: Определите возможные личностные черты абитуриента на основе анализа его активности и интересов.
- Интересы и хобби: Составьте список интересов и хобби абитуриента.
- Социальные связи: Определите круг общения и возможные влияния.

Выводы и рекомендации:

- Совместимость с учебным заведением: Оцените, насколько интересы и личностные характеристики абитуриента соответствуют требованиям и ценностям учебного заведения.
- Рекомендации: Предложите возможные направления для дальнейшего развития абитуриента или его участия в различных программах и мероприятиях.

Анализ предметной области

1. Цели анализа

- Определение интересов и предпочтений: Понять, какие темы и деятельности интересуют абитуриента.
- Оценка личностных характеристик: Определить основные черты характера и поведения абитуриента.
- Анализ социальных связей: Изучить круг общения абитуриента и его влияние.
- Прогнозирование поведения: Предсказать возможное поведение абитуриента в учебном заведении.

2. Методы и инструменты

Сбор данных:

- API ВКонтакте: Использование API для получения данных о профиле, постах, комментариях, лайках и подписках.
- Веб-скрапинг: В случае необходимости, использование веб-скрапинга для дополнительного сбора данных.

Обработка данных:

- Очистка данных: Удаление дубликатов, нерелевантных данных и шума.
- Форматирование данных: Приведение данных к удобному для анализа формату (CSV, JSON).

Анализ данных:

- Текстовый анализ: Топик-моделирование, сентимент-анализ, частотный анализ.
- Социальный анализ: Анализ друзей и подписок, активности в социальной сети.

- Машинное обучение: Использование алгоритмов машинного обучения для прогнозирования поведения.

3. Потенциальные вызовы и ограничения

- Конфиденциальность данных: Необходимо соблюдать все правила конфиденциальности и получить согласие абитуриента на использование его данных.
- Качество данных: Данные могут быть неполными или неточными, что может повлиять на результаты анализа.
- Этические аспекты: Важно учитывать этические нормы при сборе и анализе данных, чтобы не нарушать права абитуриента.
- Технические ограничения: Ограничения API ВКонтакте могут повлиять на объем и качество собираемых данных.

4. Примеры использования

- Приемная комиссия: Анализ данных может помочь приемной комиссии лучше понять абитуриента и принять обоснованное решение о его зачислении.
- Персонализированное обучение: На основе анализа данных можно разработать персонализированные образовательные программы, соответствующие интересам и потребностям абитуриента.
- Маркетинг и реклама: Анализ данных может быть использован для целевого маркетинга и рекламы, направленной на привлечение абитуриентов.

Инструментарий

Для реализации данного проекта нам понадобятся следующие инструменты:

- VK API - для сбора необходимых данных
- Язык программирования Python - для написания самого ПО

Функционал системы

- Анализ страниц пользователей с указанными ID на основании постов на его странице и названий групп, на которые он подписан.
- Составление “среднего” пользователя, на основе проанализированных данных.
- Сравнение конкретного пользователя с ранее сформированным “средним”.

Структура системы

В прототипе системы было принято решение использовать алгоритм TF-IDF. Алгоритм TF-IDF (Term Frequency-Inverse Document Frequency) — это статистический метод, используемый для оценки важности слова в контексте документа, который является частью коллекции документов или корпуса. Этот алгоритм широко применяется в области обработки естественного языка (NLP) и информационного поиска для задач, таких как классификация текстов, кластеризация и поиск информации.

Основные компоненты TF-IDF:

1. Term Frequency (TF):

Определение: Частота термина (слова) в документе.

Формула: $\mathrm{TF}(t, d) = \frac{\text{Число вхождений термина } t \text{ в документе } d}{\text{Общее число терминов в документе } d}$

Интерпретация: Чем чаще слово встречается в документе, тем выше его TF.

2. Inverse Document Frequency (IDF):

Определение: Мера того, насколько редко термин встречается во всей коллекции документов.

Формула: $\mathrm{IDF}(t, D) = log(\frac{\text{Общее число документов в коллекции } D}{\text{Число документов, содержащих термин } t})$

Интерпретация: Чем реже слово встречается в коллекции документов, тем выше его IDF.

3. TF-IDF:

Определение: Произведение TF и IDF для каждого термина в документе.

Формула: TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)

Интерпретация: TF-IDF увеличивается пропорционально частоте термина в документе, но уменьшается с увеличением частоты термина в коллекции документов. Это помогает выделить слова, которые являются важными для конкретного документа, но не слишком распространенными в коллекции.

Преимущества и недостатки:

Преимущества:

- Простота реализации.
- Эффективность для больших коллекций документов.
- Способность выделять важные термины, игнорируя часто встречающиеся, но малоинформативные слова.

Недостатки:

- Не учитывает семантику и контекст слов.
- Может быть чувствителен к длине документа.
- Не всегда эффективен для коротких текстов или текстов с большим количеством редких слов.

TF-IDF является мощным инструментом для анализа текстовых данных и широко используется в различных приложениях, связанных с обработкой естественного языка и информационным поиском, но возможно в процессе дальнейшей работы над проектом мы заменил данный алгоритм на более подходящий для нашей задачи.