Молодёжный вестник НФ БГТУ. 2024. Том 04. № 03 (15)
https://rio-nb-bstu.science

ИНФОРМАТИКА

doi: 10.51639/2713-0576_2024_4_3_64
УДК 004.622
ГРНТИ 20.23.17
ВАК 05.13.01

Создание датасета для системы искусственного интеллекта «Твой путь»

М.В. Коломина, Р.Н. Ахатов, А.В. Марышев, Э.В. Нариманян, Д.Д. Самсонов

Астраханский государственный университет им В.Н.Татищева,
414056, Россия, г. Астрахань, ул. Татищева, 20а

email: mkolomina2014@gmail.com, , ahatovruslan02@mail.ru, maryshev3@gmail.com,
narimanyan_eleonora@mail.ru, danila.samsonov02@gmail.com

Аннотация

В статье описывается процесс сбора и обработки датасета для обучения нейронной сети. Датасет используется при разработке системы искусственного интеллекта «Твой путь». Система является рекомендательным сервисом и помогает абитуриентам определиться с выбором направления подготовки бакалавриата при поступлении в вуз. Особое внимание уделяется методам и алгоритмам, используемым при сборе и обработке данных. Статья будет полезна тем, кто хочет узнать больше о процессе сбора и обработки данных для обучения нейронных сетей.

Ключевые слова: датасет, VK, сообщества и подписки, абитуриент.

Введение

Датасет - это набор данных, обычно структурированных и организованных в определенном формате. Датасеты используются в машинном обучении, статистике, аналитике данных и других областях для обучения моделей, проведения исследований и анализа информации. Датасеты могут содержать различные типы данных, такие как числа, текст, изображения и другое [1].

Формирование датасетов происходит в несколько этапов:

1. Сбор данных осуществляется из различных источников, таких как базы данных, интернет, датчики, API и т.д. Важно убедиться в качестве, соответствии и достоверности данных, необходимых для решения поставленной задачи.
2. Соответствие критериям, удовлетворяющим конкретной задаче.
3. Очистка данных от ошибок, пропусков, выбросов и других аномалий [3].
4. Преобразование данных, например, нормализация, кодирование категориальных переменных или масштабирование [4].
5. Разделение на выборки: обучающую, тестовую и валидационную. Обычно используется соотношение 60-20-20 [2].
6. Балансировка классов.
7. Аугментация данных. Для увеличения разнообразия данных и повышения обобщающей способности модели используются методы аугментации данных, такие как повороты, отражения, изменение яркости и т.д.

8. Подготовка данных для загрузки в модель. Данные должны иметь формат, который может быть загружен и использован для обучения нейронной сети [4].
Правильная сборка и подготовка датасета играют ключевую роль в успешном обучении нейронных сетей и достижении высокой точности предсказаний [5].

Основная часть

Ранее абитуриенты вузов обладали системным мышлением, способностью к глубокому анализу и целостному восприятию информации. Однако современные абитуриенты чаще характеризуются клиповым мышлением, фрагментарным восприятием и низкой концентрацией внимания, сниженной способностью к анализу. Кроме того, вузы прошлого предлагали ограниченный набор образовательных программ, в то время как, современные имеют широкий выбор образовательных программ и направлений подготовки бакалавриата. Возникла проблема: абитуриент не готов к выбору направления подготовки бакалавриата. Для решения этой проблемы разработана система искусственного интеллекта «Твой путь». Она анализирует страницу абитуриента в VK, определяет его интересы, на основании которых

- быстро рекомендует абитуриенту направления бакалавриата;
- список профессий по рекомендуемому направлению подготовки бакалавриата;
- предполагаемый список вступительных испытаний.

Для разработки системы искусственного интеллекта, прежде всего, необходимо было сформировать датасет. Рассмотрим этапы его создания.

Этап 1. Сбор данных

На страницах различных вузов, в сети Internet, отыскивались приказы прошлых лет о зачислении абитуриентов на программы бакалавриата. Из приказов выбиралась информация: Фамилия, Имя, Отчество, Номер группы направления подготовки бакалавриата.
Затем, у этих абитуриентов собиралась информация со страниц в сети VK о сообществах и подписках. Данная информация была необходима для установления связи между интересами абитуриента и направлением подготовки бакалавриата на которое он был зачислен.
Для сбора информации в сети VK разработана платформа «VK API Groups Parser» (VK_API_GP). Сервис предоставляет возможность получить информацию с серверов VK, посредством взаимодействия с их API, по заданным Фамилии и Имени пользователя или поего ID. Критерии качества данных: открытая страница VK, наличие целевых сообществ, наличие открытых подписок и сообществ.
При разработке платформы использовались следующие технологии:
- фреймворк для разработки современных веб-приложений Vue.JS;
- фреймворк Nuxt.JS для более удобной разработки, настройки, сборки приложений на Vue.JS;
- язык препроцессора SASS — SCSS. Метаязык CSS, дающий дополнительный функционал стилизации поверх стандартного функционала CSS;
- библиотека XLSX, позволяющая сгенерировать из HTML таблицы таблицу формата «.xlsx».
Сервис VK_API_GP имеет два варианта сбора данных.
Вариант 1. Используется ID зачисленного абитуриента в сети VK для этого:
а) по известным фамилии и имени абитуриента отыскивалась его страница в VK;
б) на странице абитуриента проверялись данные на соответствие критериям качества;

в) ID VK копировалось в VK_API_GP. Если ID несколько, то они перечисляются через запятую (рис. 1);
г) кнопка «Получить» выдает сообщества и подписки со страницы абитуриента. В ходе работы кнопки выполняются методы VK API: users.get, users.getSubscriptions, groups.get. В случае возникновения ошибки, приложение выводит о ней сообщение;
д) действие кнопки «Добавить» осуществляет заполнение таблицы с сырыми данными (рис. 2).

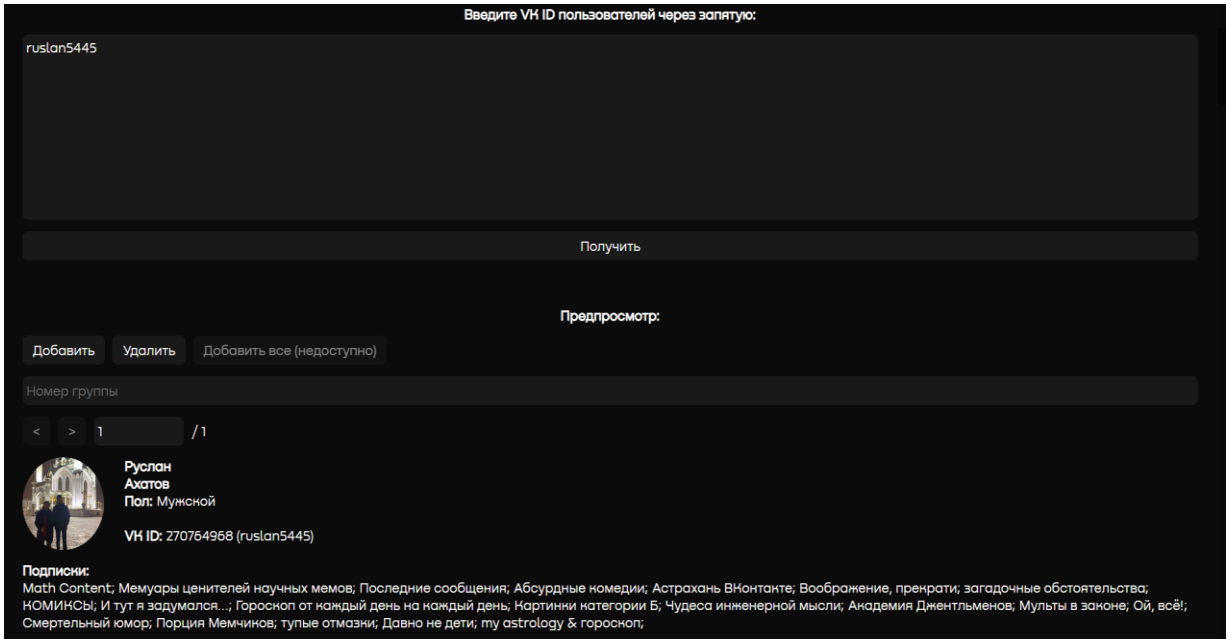

Рис. 1. Работа программы по первому варианту

Имя	Фамилия	Отчество	Шифр	VK ID (Number)	VK ID	Город	Школы, колледжи и университеты	Любимые книги	Любимая музыка	Любимые игры	Любимые телешоу	Любимые цитаты	Деятельность	Интересы	Сообщества и подписки
Руслан	Ахатов			270764968	ruslan5445										Math Content, Мемуары ценителей научных мемов, Последние сообщения, Абсурдные комедии, Астрахань ВКонтакте, Воображение, прекрати, загадочные обстоятельства, КОМИКСЫ, И тут я задумался..., Гороскоп от каждый день на каждый день, Картинки категории Б, Чудеса инженерной мысли, Академия Джентльменов, Мульты в законе, Ой, всё!, Смертельный
Никита	Козлюк			555240590	nekit7144		АГУ								Антисоциальная эстетика, АГУ Медиа, Empire of Астрахань [PVA], МХН, Астрахань online, Единый деканат АГУ, Рифмы и Панчи, АГУ |Астраханский государственный университет|, Леонардо Дрейвинчик, ПРИКОЛЫ | Смешно, МДК, Киномания - Лучшие фильмы, Литература | Великие поэты, Сарказм, Чётные приколы, МЕМЫ, 5 умных мыслей, Я хочу..., НЕНОРМАЛЬНО, 4ch,
Василя	Абубекерова			164022131	yellowboe										фен, tranquility, На все случаи жизни, искусство психологии, Рыбка, mur, Смотри и Думай, saudade, заходи на чай, moe's, it's OK, potrachenó, отрывок из твоего любимого фильма, Натальная карта, the nbhd's soundcheck, мёд в голове, Эстетика | Wildberries, песня, под которую реву, Кусочек Фильма, CLIQUE, точно, где сопротивляется всё, в горчичном лесу, какие-то

Рис. 2. Таблица с сырыми данными

Вариант 2. Сбор данных по фамилии и/или имени зачисленных абитуриентов (студентов), и ID группы ВУЗа в VK для этого:
а) осуществлялся поиск страницы ВУЗа в сети VK;
б) определялся ID VK ВУЗа;
в) в программе VK_API_GP, на форме для ввода, заполнялись поля Фамилия, Имя, Отчество, шифр направления подготовки зачисленного абитуриента и ID группы ВУЗа в сети VK (рис. 3);
г) кнопка «Получить» выдает сообщества и подписки со страницы студента. В ходе работы кнопки выполняются методы VK API: users.search, users.getSubscriptions, groups.get. В случае возникновения ошибки, приложение выводит о ней сообщение;
д) полученные данные проверялись на соответствие критериям качества;
е) действие кнопки «Добавить» осуществляет заполнение таблицы с сырыми данными (рис. 3).

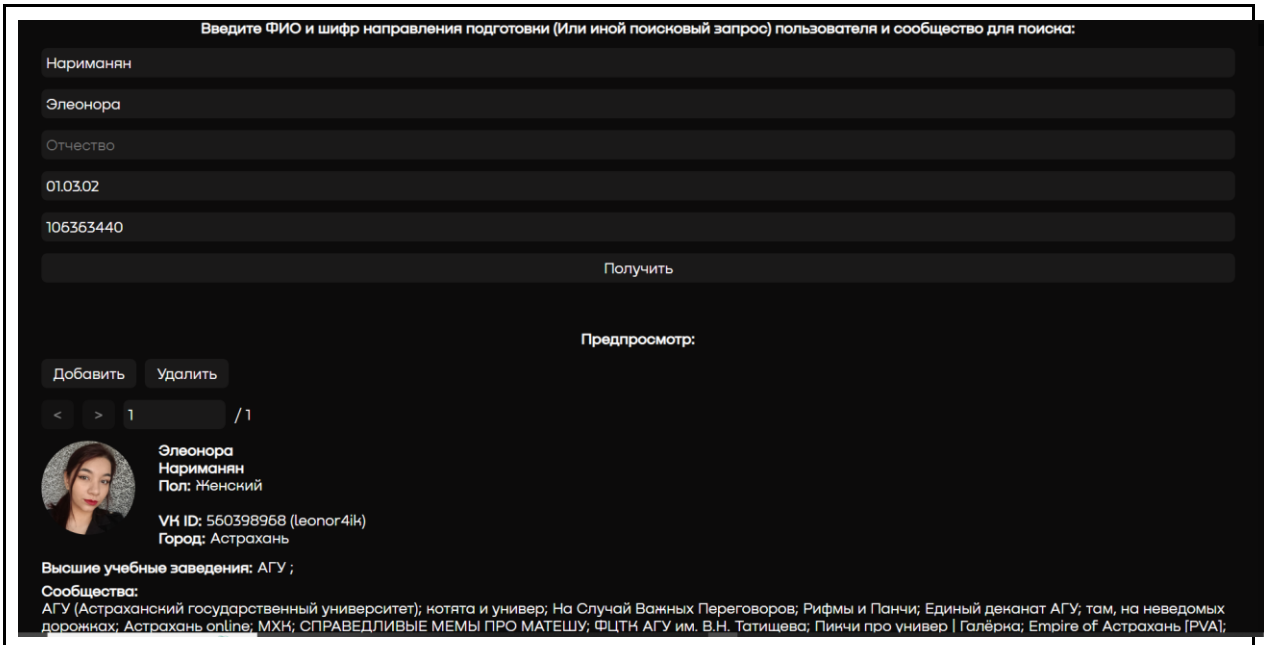

Рис. 3. Работа программы по второму варианту

В сервисе VK_API_GP реализована возможность сохранить таблицу с сырыми данными в нативном формате для excel (.xlsx), а также в общепринятом формате (.csv).
Собранные данные структурированы, допускают пропуски в поле «Предыдущее место обучения».

Этап 2. Обработка данных

Для работы с базой данных реализован класс DBService, содержащий методы:

- __init__ - конструктор, который реализует подключение к базе данных.
- create_popular_publics - метод, который заполняет таблицу popular_publics переданным списком сообществ и подписок.

- create_groups - метод, который заполняет таблицу groups переданным словарём значений “Порядковый номер группы в датасете - реальный номер группы направлений”.
- is_popular_public - метод, который определяет, является ли сообщество или подписка популярной.

Сырые данные обрабатываются в файле формата csv. Для работы с ними в системе искусственного интеллекта реализован модуль DataModule, который содержит

1. del_punctuation (str, delimeters) – функцию очищения строки str от лишних символов, например, !, @, №, $, %, ^ и т.д., а также удаления от лишних пробелов. Функция возвращает очищенную строку.

```
def del_punctuation(str, delimeters):
    result_str = (str + '.')[:-1]
for delimeter in delimeters:
while delimeter in result_str:
        result_str = result_str.replace(delimeter, '')
while '  ' in result_str:
    result_str = result_str.replace('  ', ' ')
while result_str[0] == ' ':
    result_str = result_str[1:]
while result_str[len(result_str) - 1] == ' ':
    result_str = result_str[:-1]
return result_str
```

2. clear (str_list, db_service) – функцию очищения данных от популярных названий сообществ и подписок. Они хранятся в базе данных. Определение, является ли сообщество или подписка популярной, осуществляется методом is_popular_public.

```
def clear(str_list, db_service):
    i = 0
while i < len(str_list):
if db_service.is_popular_public(str_list[i]):
        str_list.pop(i)
else:
i += 1
    return str_list
```

3. CSVService – класс, необходимый для обработки датасета, и подготовки данных для обучения нейронной сети. Он содержит:

- __db_service__ – поле, с помощью которого происходит обращение к таблице популярных пабликов popular_publics в базе данных;
- __init__(self, db_service) – конструктор, создающий объект класса CSVService и инициализирующий поле __db_service__ типа DBService, для реализации возможности обращения к таблице популярных пабликов popular_publics и заполнения таблицы groups в базе данных;
- create_popular_publics(self, dataset_path) – метод формирующий таблицу самых популярных сообществ и подписок.

```
def create_popular_publics(self, dataset_path):
    csv_groups = pandas.read_csv(dataset_path, delimiter=';')['groups']
    frequence_dict = dict()
for groups_str in csv_groups:
    groups = del_punctuation(str(groups_str).lower(), './\\!@#$%^&*()-+_?;\"\':`|<>[]').split(',')
        i = 0
while i < len(groups):
if groups[i] == '' or groups[i] == ' ':
```

```
groups.pop(i)
else:
                i += 1
for group in groups:
group = del_punctuation(group, '')
if group in frequence_dict:
                frequence_dict[group] += 1
else:
                frequence_dict[group] = 1
        count_clearing_groups = int(len(frequence_dict) / 20)
        clearing_groups = list()
while len(clearing_groups) < count_clearing_groups:
            current_max_frequence = max(frequence_dict.values())
            extended_groups = [group for group in frequence_dict if frequence_dict[group] == current_max_frequence]
            clearing_groups.extend(extended_groups)
for group in extended_groups:
del frequence_dict[group]
while len(clearing_groups) > count_clearing_groups:
            clearing_groups.pop()
        self.__db_service__.create_popular_publics(clearing_groups)
```

- clear_dataset(self, dataset_path, cleared_dataset_path) – метод очистки каждой записи в датасете от самых популярных названий сообществ и подписок.

```
def clear_dataset(self, dataset_path, cleared_dataset_path):
        csv_file = pandas.read_csv(dataset_path, delimiter=';')
for index, record in csv_file.iterrows()
record_groups = del_punctuation(str(record['groups']).lower(), './\\!@#$%^&*()-+_?;\"\':`|<>[]').split(',')
print(index)
            i = 0
while i < len(record_groups):
if record_groups[i] == '' or record_groups[i] == ' ':
                record_groups.pop(i)
else:
                record_groups[i] = del_punctuation(record_groups[i], '')
                i += 1
            record_groups = clear(record_groups, self.__db_service__)
            group_str = ' '.join(record_groups)
while '  ' in group_str:
                group_str = group_str.replace('  ', ' ')
            csv_file.at[index, 'groups'] = group_str
        csv_file.to_csv(cleared_dataset_path, index = False, sep=';')
```

- synthesize (self, dataset_path, dataset_with_synthesize_path, tematics_publics_path, group, count=1) – методдобавляющийсинтезированные (искусственные) записивдатасет. Генерация синтетических данных происходит путем смешивания тематических пабликов для каждой группы направлений с сообществами, не относящимися к этой группе направлений.

```
def synthesize(this, dataset_path, dataset_with_synthesize_path, tematics_publics_path, notematics_publics_path, group, count_data = 1):
        tematics_csv_file = pandas.read_csv(tematics_publics_path, delimiter=';')
        no_tematics_csv_file = pandas.read_csv(notematics_publics_path, delimiter=';')
```

```
        tematics_csv_file = tematics_csv_file[tematics_csv_file.old_code == group]
        csv_file = pandas.read_csv(dataset_path, delimiter=';')
for i in range(0, count_data):
            id_direction_code = 0
for index, record in csv_file.iterrows():
if int(record['old_code']) == group:
                    id_direction_code = int(record['code'])
break
if id_direction_code == 0:
continue
            TematicsCSVFileRand = tematics_csv_file.sample(n = 3)
            NoTematicsCSVFileRand = no_tematics_csv_file.sample(n = 10)
            GroupsList = TematicsCSVFileRand['groups'].to_list()
GroupsList.extend(NoTematicsCSVFileRand['groups'].to_list())
            j = 0
while j < len(GroupsList):
                GroupsList[j] = del_punctuation(GroupsList[j], './\\!@#$%^&*()-+_?;\"\':`|<>[]')
                j += 1
            csv_file.loc[len(csv_file.index)] = [len(csv_file) + 1, group, id_direction_code, '', del_punctuation(' '.join(GroupsList).lower(), './\\!@#$%^&*()-+_?;\"\':`|<>[]')]
        csv_file.to_csv(dataset_with_synthesize_path, index = False, sep=';')
```

- create_groups (self, dataset_path) – метод создающий таблицу groups в базе данных для связи результатов работы нейронной сети с группой направлений подготовки бакалавриата.

```
def create_groups(self, dataset_path):
groups = pandas.read_csv(dataset_path, delimiter=';')[['old_code', 'code']]
        groups_dict = dict()
for index, record in groups.iterrows():
            groups_dict[int(record['code'])] = int(record['old_code'])
        self.__db_service__.create_groups(groups_dict)
```

В итоге был сформирован датасет, состоящий из 3200 записей.
Для каждой группы направлений бакалавриата число записей, в среднем, составило 100. Произведено разделение датасета на обучающий, состоящий из 60 записей, тестовый - число записей 20, валидационный - число записей 20.
На рис. 4 представлено распределение записей в обучающем датасете по группам направлений бакалавриата.

Заключение

Подготовленный датасет направляется в нейронную сеть для ее обучения. Сбор и подготовка датасета – это ключевой шаг при обучении нейронных сетей, который влияет на эффективность и качество предсказаний. Важно тщательно подходить к подбору данных, обеспечивать их разнообразие и качество, проводить необходимую обработку данных для минимизации шума и искажений.

Рис. 4. Распределение записей в обучающем датасете по группам направлений

Конфликт интересов

У авторов нет конфликта интересов по материалам данной статьи с третьими лицами на момент подачи статьи в редакцию журнала, им ничего не известно о возможных конфликтах интересов в настоящем со стороны третьих лиц.

Список литературы

1. Что такое датасеты [Электронный ресурс] // Annotate. – Режим доступа: https://annotate.ru/shto_takoye_datasety (дата обращения: 18.03.2024).
2. Как разделять набор данных [Электронный ресурс] // VC.RU. – Режим доступа: https://vc.ru/newtechaudit/485313-kak-razdelyat-nabor-dannyh (дата обращения: 18.03.2024).
3. Очистка данных: кто их загрязняет и что аналитику с этим делать [Электронный ресурс] // Practicum Yandex. – Режим доступа: https://practicum.yandex.ru/blog/chto-takoe-ochistka-dannyh/ (дата обращения: 18.03.2024).
4. Как подготовить данные для машинного обучения [Электронный ресурс] // Practicum Yandex. – Режим доступа: https://practicum.yandex.ru/blog/podgotovka-dannyh-k-analizu/ (дата обращения: 19.03.2024).
5. Подготовка датасета для машинного обучения: 10 базовых способов совершенствования данных [Электронный ресурс] // Хабр. – Режим доступа: https://habr.com/ru/articles/684580/ (дата обращения: 19.03.2024).

Молодёжный вестник НФ БГТУ. 2024. Том 04. № 03 (15)
https://rio-nb-bstu.science

Formation of a dataset for the artificial intelligence system «Your way»

M.V. Kolomina, R.N. Akhatov, A.V. Maryshev, E.V. Narimanyan, D.D. Samsonov

Astrakhan State University named after V.N.Tatishchev , 20a Tatishcheva str., Astrakhan, 414056, Russia

email: pmiiagu@gmail.com

Abstract

The article describes the process of collecting and processing a dataset for training a neural network. The dataset is used in the development of the artificial intelligence system "Your way". The system is a recommendation service and helps applicants decide on the direction of bachelor's degree when applying to a university. Special attention is paid to the methods and algorithms used in data collection and processing. The article will be useful for those who want to learn more about the process of collecting and processing data for training neural networks.

Keywords: dataset, VK, communities and subscriptions, entrant.

References

1. What are datasets [Electronic resource] // Annotate. – Access mode: https://annotate.ru/shto_takoye_datasety (date of application: 03/18/2024).
2. How to divide a data set [Electronic resource] // VC.RU . – Access mode: https://vc.ru/newtechaudit/485313-kak-razdelyat-nabor-dannyh (date of application: 03/18/2024).
3. Data cleaning: who pollutes them and what an analyst should do with it [Electronic resource] // Practicum Yandex. – Access mode: https://practicum .yandex.ru/blog/chto-takoe-ochistka-dannyh / (date of access: 03/18/2024).
4. How to prepare data for machine learning [Electronic resource] // Practicum Yandex. – Access mode: https://practicum .yandex.ru/blog/podgotovka-dannyh-k-analizu / (date of access: 03/19/2024).
5. Preparing a dataset for machine learning: 10 basic ways to improve data [Electronic resource] // Habr. – Access mode: https://habr.com/ru/articles/684580 / (date of access: 03/19/2024).