Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»
(Университет ИТМО)

Факультет **Инфокоммуникационных технологий**

Образовательная программа **Мобильные и сетевые технологии**

Направление подготовки **09.03.03 Прикладная информатика**

О Т Ч Е Т

об учебной, ознакомительной практике

Тема задания: Анализ аудитории тематических сообществ о генеративных моделях

Обучающийся: Хисаметдинова Динара Наилевна, группы К3241

Руководитель практики от университета: Белка Алёна Александровна, ассистент факультета инфокоммуникационных технологий

Санкт-Петербург 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном мире социальные сети стали неотъемлемой частью общественной жизни, играя ключевую роль в формировании мнений и распространении информации. Однако глубокое понимание динамики и структуры сообществ в социальных сетях представляет собой сложную задачу из-за постоянно меняющихся интересов аудитории. Это создает необходимость в гибких методах анализа для выявления предпочтений пользователей. Особенно актуален такой анализ для разработки стратегий создания и развития новых сообществ, где важно определить оптимальные темы для обсуждения и время публикации, чтобы максимально повысить вовлеченность аудитории.

Целью курсовой работы является комплексный анализ тематического наполнения и пользователей сообществ в социальных сетях. Исследование направлено не только на понимание тенденций среди интересов подписчиков групп определенной тематики, но и на выявление эффективных стратегий для создания и развития новых сообществ, учитывая текущие предпочтения аудитории. Работа также включает анализ различий между узкоспециализированными сообществами и теми, что привлекают более широкий круг подписчиков, с целью выявления влияния тематической направленности на уровень вовлеченности и активности аудитории.

Задачи проекта:

- поиск подходящих для анализа сообществ и выдвижение гипотез;
- изучение документации Reddit API и соответствующей библиотеки на Python (async PRAW), а также VK API;
- написание кода для скрейпинга данных о пользователях и постах;
- описание, анализ и визуализация данных с использованием библиотек;
- структуризация полученных данных и сравнение двух сообществ;
- проверка гипотез.

1 Описание посещений лекций-бесед

1.1 Первая лекция

8 февраля был проведён доклад Артамоновой Валерии Евгеньевны на тему «Кто такой системный аналитик и почему в каждой компании у него разные задачи». Спикер подробно рассказала про своё образование, о том, как вс и опыт работы.

В ходе выступления Валерия Евгеньевна подробно рассказала про своё образование, о том, как разные его аспекте пригодились в последующих проектах, про опыт работы и работодателях, таких как Ediweb, Samokat.Tech. Такая информация позволила очертить круг обязанностей системного аналитика, примерить на себя эту роль, поразмыслить, хватило ли бы у меня компетенций для данной позиции и понять, что факультет ИКТ позволяет получить достойную базу для работы в разных сферах в рамках информационных технологий.

Спикер также дала полезную информацию о важности системного аналитика как связующего звена между разработкой и клиентом, о том, какой широкий пул задач закрывает такой специалист, о последовательности обязанностей, а также интересные случаи из рабочих будней.

Для меня показалось неожиданным то, насколько широкий кругозор должен иметь системный аналитик, чтобы быть востребованным: в её практике пригождалось знание Javascript, REST API, понимание методологий управления, архитектуры приложения, моделирования бизнес процессов, используя BPMN, умение вести документацию, работать с базами данных и компьютерными сетями, проверять наличие багов и даже набрасывать лендинги в Figma.

Встреча вызвала у меня много положительных эмоций благодаря замечательной презентации, хорошей подаче и тому, что сама рассматриваю варианты развития именно в этой профессии, поэтому мне важно было узнать о положительных и отрицательных сторонах работы. После выступления

Артамоновой В.Е. я перестала недооценивать те профессии в IT, которые не связаны непосредственно с разработкой, технической поддержкой и дизайна продукта, а больше направлены на коммуникацию и управление, ведь работа с людьми полна казусов, сложных моментов. Ещё я в очередной раз убедилась, что не нужно бояться собеседований и обилия математики.

1.2 Вторая лекция

12 февраля состоялась встреча с Анастасией Колгановой, на которой обсуждалось, как образование, полученное в ИТГС, способствует успеху в профессиональной деятельности, особенно в сфере дизайна интерфейсов. Анастасия поделилась опытом применения этих знаний в своей карьере и представила свой выпускной проект, посвященный анализу трудовой мотивации в России с использованием данных из открытых источников.

В рамках своего выступления, спикер детализировала процесс своей работы, начиная с классификации видов бонусов и заканчивая разработкой информационных панелей (дашбордов), которые иллюстрируют полученные результаты. Она также упомянула о вкладе центра учебной аналитики в изучение вопросов мотивации студентов и их сознательного выбора образовательного пути.

Несмотря на то, что дизайн интерфейсов не является тем, чем я хочу заниматься в качестве основной работы, лекция дала представления о том, чем будут заниматься мои коллеги, а также я услышала полезные советы по выбору научного руководителя, написанию и защите дипломной работы, что предстоит мне, как студенту.

Для меня это мероприятие оказалось весьма информативным и расширило мои знания. Немаловажным открытием для меня стал ITMO TRACK, позволяющий подстраивать, казалось бы, строгое академическое образование под свои карьерные предпочтения. Встреча с Анастасией стала отличной возможностью узнать больше о применении аналитических навыков

в реальных проектах и о том, как данные из открытых источников могут быть использованы для глубокого понимания социальных процессов, в частности, в контексте трудовой мотивации.

1.3 Третья лекция

13 февраля состоялась презентация Виктора Гляненко, посвященная актуальным направлениям в мире видеоигр. Виктор поделился своим опытом и знаниями о последних тенденциях и инновациях в индустрии видеоигр, обсудив новшества в игровых механиках и предпочтениях пользователей, а также роль новых технологий в адаптации игр к запросам сегодняшних геймеров.

Во время своего выступления Гляненко эффективно демонстрировал эволюцию игровой сферы, акцентируя внимание на ключевых и впечатляющих тенденциях. Он детально анализировал, как технологический прогресс и изменяющиеся вкусы игроков влияют на формирование игровых процессов, подчеркивая необходимость их адаптации к ожиданиям аудитории. Отмечу то, что он высоко оценил важность роли аналитика в индустрии, об особенностях тестирования.

Презентация была насыщенной, логично построенной и захватывающей, предоставив слушателям важные сведения о современных течениях в игровой индустрии. Самое интересное – вопросы и последующие ответы на них специалистом. Несмотря на мой низкий уровень заинтересованности в развитии в сфере видеоигр, доклад оказался крайне полезным благодаря советам по трудоустройству и рассказу о том, как кардинально при желании можно сменить специализацию в сфере информационных технологий на примере тестировщиков.

2 Извлечение данных о пользователях и постах

2.1 Выбор подходящих тематических сообществ и выдвижение гипотез

Критериями для выбора тематических групп для последующего анализа были такие параметры, как принадлежность тематике, предложенной куратором, большое число подписчиков, так как чем больше выборка, тем точнее результаты, наличие возможности выгрузки данных пользователей, а также актуальность, то есть важно, чтобы пользователи группы были достаточно активны, а создатели регулярно выкладывали посты.

Таким образом, было решено выбрать группу в социальной сети Вконтакте ‘Midjourney — арты от нейросети’, посвященной сгенерированным одноименной нейросетью изображениям, а также сообщество в социальной сети Reddit, посвящённой LLM (Large Language Model).

Первая группа для широкого круга пользователей, в то время как вторая – узкоспециализирована, заточена под разработчиков, аналитиков. Исходя из этого выдвинута гипотеза о том, что интересы подписчиков первого сообщества более разнообразны, как и возраст, профессия. Также активность во второй группе предположительно должна быть меньше подвержена изменениям в зависимости от времени выхода записей, количества вложений

2.2 Работа с Reddit API

Мной были рассмотрены две Python библиотеки для работы с Reddit API – PRAW (Python Reddit API Wrapper) и Async PRAW. Так как предстояла выгрузка довольно большого количества данных, и асинхронный код сильно ускоряет процесс передачи множества запросов, была выбрана вторая.

Для начала необходимо было зайти на страницу для разработчиков Реддит, чтобы зарегистрировать приложение и получить client id и client secret, что позволяет работать с API. Ниже представлен листинг извлечения данных за один год в async_posts_details.csv.

```python
import asyncpraw
import asyncio
import csv
from datetime import datetime, timezone
from config import client_id, client_secret

async def main():
    reddit = asyncpraw.Reddit(
        client_id=client_id,
        client_secret=client_secret,
        user_agent='reddit_data_mining'
    )

    subreddit_name = 'LLM'
    subreddit = await reddit.subreddit(subreddit_name)

    start_time = datetime(2023, 1, 1, tzinfo=timezone.utc).timestamp()
    end_time = datetime(2024, 1, 1, tzinfo=timezone.utc).timestamp()

    with open('async_posts_details.csv', mode='w', newline='', encoding='utf-8') as file:
        writer = csv.writer(file)
        writer.writerow(['post_title', 'upvotes', 'num_comments', 'date_created'])
        async for submission in subreddit.new(limit=1000):
            if start_time <= submission.created_utc <= end_time:
                created_date = datetime.fromtimestamp(submission.created_utc, tz=timezone.utc).strftime('%Y-%m-%d %H:%M:%S')
                writer.writerow([submission.title, submission.score, submission.num_comments, created_date])

    await reddit.close()

if __name__ == '__main__':
    asyncio.run(main())
```

Рисунок 1 – Парсинг информации о постах

Таким образом, получился файл с названием поста, числом лайков, комментариев и датой создания.

```
post_title,upvotes,num_comments,date_created
Decoding the preprocessing methods in the pipeline of building LLMs,9,4,2023-07-17 09:31:09
Running LLMs Locally,41,34,2023-07-17 08:41:17
"There's an Azure OpenAI sub for which a primary key and a secondary key exists which is managed via the Azure portal. Current rate
"Does USA have free online legal databases in the same format as BAILII, CanLII, CommonLII, etc?",5,1,2023-07-15 23:14:23
Jobs with a LLM,5,6,2023-07-14 18:38:43
How do you Monitor Your Production LLM based Application?,0,6,2023-07-14 08:03:17
"Hey folks! Ever wished you could get a mind-blowing brainstorm report generated by AI agents in just 20 minutes? Well, guess what
Best way to map user questions to code functions,0,10,2023-07-12 01:08:54
Falcon 40B - impressive local LLM,1,4,2023-07-10 14:03:06
Fine-Tuning Insights: Lessons from Experimenting with RedPajama Large Language Model on Flyte Slack Data,0,2,2023-07-10 13:30:07
"Are ""Language Models"" simply Decoder-Only Transformers?",0,2,2023-07-10 06:48:34
AI player companions?,1,1,2023-07-10 05:48:27
"Introduction to Language Models (LLM's, Prompt Engineering, Encoder/Deco...",0,1,2023-07-09 12:16:26
LLM + SQE,1,1,2023-07-08 19:08:22
GPT Alternatives,1,7,2023-07-08 07:36:57
Is there decent open-source LLMs faster than Falcon-7b-instruct?,2,4,2023-07-08 07:28:54
Is someone here who is applying or applied already to Washington State to sit for the bar as an LLM?,4,0,2023-07-07 15:33:26
"General LLM VS LLM with concentration,which one is better for future career ?",3,0,2023-07-07 10:18:47
NSQL: First Ever Fully OpenSource SQL Foundation Model,1,1,2023-07-07 05:52:04
Book recommendation please,2,1,2023-07-07 04:49:28
Who chooses what tools you can use for work?,0,4,2023-07-06 17:30:34
What GPA did you get in your LLM program?,3,3,2023-07-06 13:52:06
I need a personal model,0,3,2023-07-06 09:32:55
"Seeking Advice: Building Language Models for Non-English Languages (e.g., Spanish or Japanese)",0,4,2023-07-05 15:30:01
How many credits should I take?,3,1,2023-07-04 15:48:40
Best Law books,1,1,2023-07-03 21:24:41
LLM IN UK,2,0,2023-07-03 20:09:07
customizing llm with subreddit data.,0,3,2023-07-02 23:41:26
A Blueprint for AI Regulation,0,1,2023-06-29 23:05:00
```

Рисунок 2 – Сниппет файла с данными

Далее, необходимо было извлечь информацию об интересах пользователей, однако, Reddit ограничивает доступ к ним напрямую. Тогда, мной было принято решение поступить так: пройтись по списку комментаторов, а потом уже произвести выгрузку названий групп, в которых

эти пользователи оставляли комментарии. Дело в том, что названия сообществ этой социальной сети всегда отражают их тематику, таким образом, были найдены данные о том, чем ещё интересуются подписчики группы ‘LLM’. Детали можно увидеть в листинге кода ниже.

```python
import praw
import csv
from collections import defaultdict
from config import client_id, client_secret

reddit = praw.Reddit(
    client_id=client_id,
    client_secret=client_secret,
    user_agent='reddit_data_mining',
    username = 'data_mining by NaughtyChinchilla'
)
subreddit_name = 'LLM'
subreddit = reddit.subreddit(subreddit_name)
subreddits_commented = defaultdict(int)

for submission in subreddit.new(limit=8):
    submission.comments.replace_more(limit=0)
    for comment in submission.comments.list():
        user = comment.author
        if user:
            for user_comment in reddit.redditor(str(user.name)).comments.new(limit=80):
                subreddits_commented[user_comment.subreddit.display_name] += 1

with open('midjourney_subreddits_commented.csv', mode='w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerow(['subreddit_name', 'comment_amount'])
    for subreddit, count in subreddits_commented.items():
        writer.writerow([subreddit, count])
```

Рисунок 3 – Извлечение интересов аудитории

2.3 Работа с VK API

Была детально разобрана документация. Затем во вкладке, предназначенной для разработчиков, было создано standalone-приложение, с помощью которого был получен токен для парсинга. Я написала функции, содержащие информацию о запросах для получения таких данных о пользователях, как пол, возраст, названия групп, на которые они подписаны, тематики этих групп, профессии, город, дату рождения. Среди выгруженной информации о постах были текст публикации, дата и время, количество комментариев, лайков, просмотров, приложенных файлов. Для соблюдения ограничений API VK информация выгружалась в несколько файлов по 1000 строк по очереди, используя параметр offset. Ниже представлена малая часть листинга для скрейпинга данных о подписчиках.

```
group_id = '215205015'

def get_user_groups(user_id, access_token):
    """Получение списка групп пользователя с их названиями."""
    groups = []
    try:
        response = requests.get('https://api.vk.com/method/groups.get', params={
            'user_id': user_id,
            'extended': 1,
            'access_token': access_token,
            'v': '5.131',
            'count': 40
        }).json()
        if 'response' in response:
            groups = [group['name'] for group in response['response']['items']]
    except Exception as e:
        print(f"Ошибка при получении групп пользователя {user_id}: {e}")
    return groups

def get_user_careers(user_id, access_token):
    """Получение информации о должностях в карьере пользователя."""
    careers_positions = ''
    try:
        response = requests.get('https://api.vk.com/method/users.get', params={
            'user_ids': user_id,
            'fields': 'career',
            'access_token': access_token,
            'v': '5.131'
        }).json()
        if 'response' in response:
            user_data = response['response'][0]
            if 'career' in user_data and user_data['career']:
                # только должности из каждой записи карьеры
                positions = [career.get('position') for career in user_data['career'] if career.get('position')]
                careers_positions = '; '.join(positions)
    except Exception as e:
        print(f"Ошибка при получении информации о карьере пользователя {user_id}: {e}")
    return careers_positions

def get_user_interests(user_id, access_token):

    interests = ''
    try:
        response = requests.get('https://api.vk.com/method/users.get', params={
            'user_ids': user_id,
            'fields': 'personal',
```

Рисунок 4 – Извлечение данных о подписчиках

3 Анализ сообществ

3.1 Анализ аудитории группы LLM в Reddit

Важнейшим показателем динамики активности пользователей стала тепловая карта, изображающая, в какие дни недели и время дня аудитория активно реагирует на посты. Как и ожидалось, наиболее благоприятным временем для постинга является промежуток с 14:00 по 21:00, в основном, будние дни, ибо группа больше не развлекательная, а образовательная.

Для визуализации интересов сообщества было составлено облако слов, которое представлено ниже. Результат соответствует гипотезе – в нём много специализированных терминов. Интересно, что в нём присутствуют слова, относящиеся к высшему образованию – University, NYU, scholarship, degree, GPA, Berkeley. Это показывает, что этой отраслью интересуются в основном люди, имеющие серьёзную академическую подготовку. Для анализа текста использовалась библиотека nltk, которая активно применялась далее в работе.

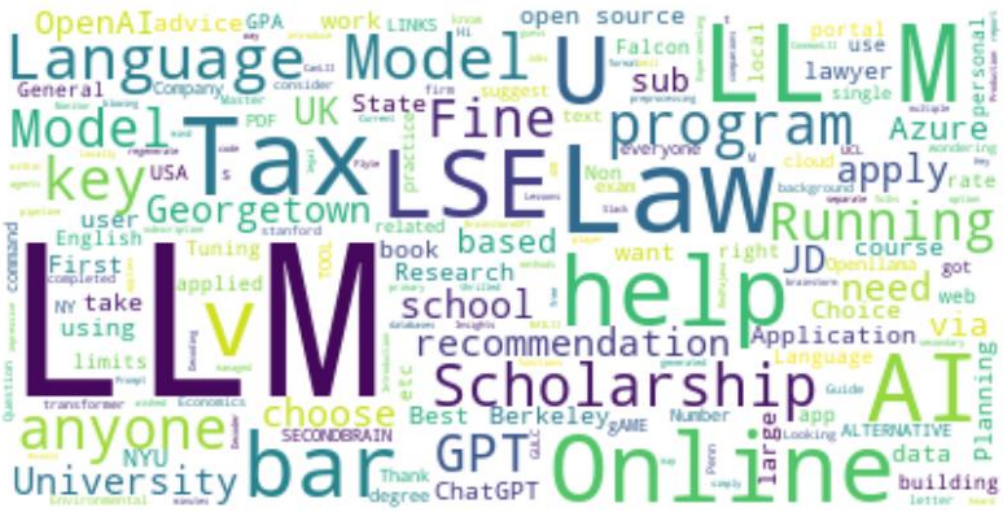

Рисунок 5 – Облако ключевых слов для сообщества ‘LLM’

Итогом анализа интересов самих пользователей стала визуализация того, посты каких других групп комментируют подписчики. Как выяснилось, это не только связанные с программированием, машинным обучением и языковыми моделями сообщества, но и новостные политические, а также посвящённые поиску работы.

3.2 Анализ аудитории группы Midjourney

В тексте постов в группе, посвящённой изображениям, сгенерированным нейросетью, наиболее популярными оказались, как и ожидалось, оказались слова, связанные с визуальными характеристиками (рис. 6).

Также были визуализированы взаимосвязи между тематиками групп (рис. 7) посредством тепловой карты. Для каждого пользователя анализируются тематики групп, с которыми он взаимодействовал. Если он подписан на две группы с разными тематиками, то счётчик взаимосвязи между ними в матрице увеличивается на единицу. Это позволяет выявить, как часто пользователи, интересующиеся одной тематикой, также интересуются и другой.

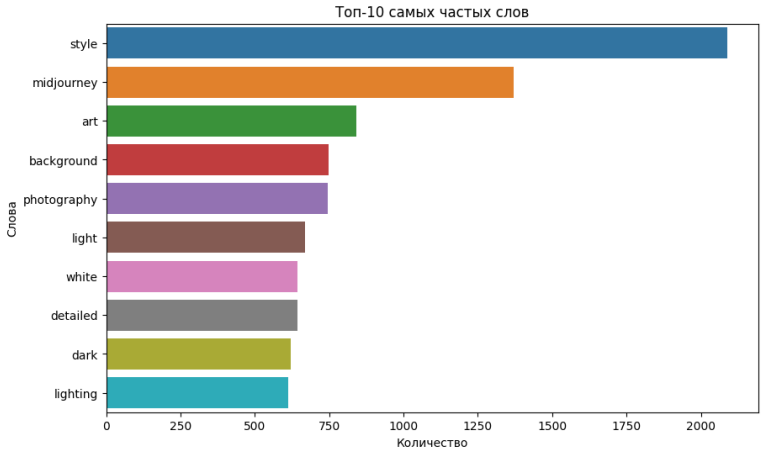

Рисунок 6 – Визуализация самых часто встречающихся слов в публикациях

Далее был проведён полезный для прикладных задач анализ, который показывает, когда выгоднее выкладывать публикации в группе по данной тематике. Выяснилось, что корреляция не слишком выраженная, однако в субботу активность несколько выше. Необычно, что активность поднимается не вечером, а днём. Это связано с тем, что midjourney используют в работе.

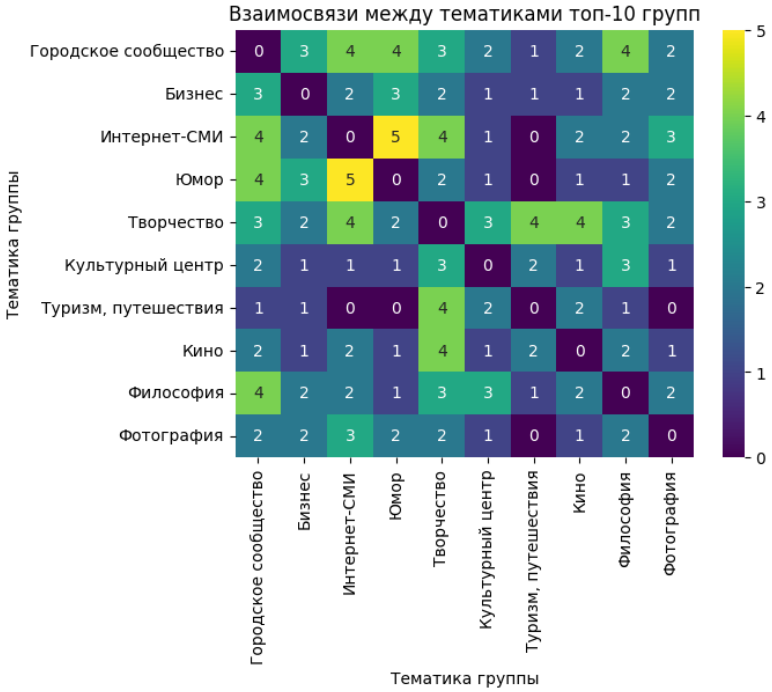

Рисунок 7 – Тепловая карта взаимосвязи между тематиками групп

Медианный возраст участников сообщества - 36 лет, что удивительно, так как было предположение, что наиболее активные пользователи нейросетей – молодёжь до 25 лет.

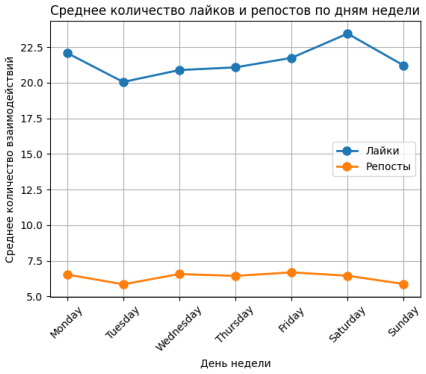

Рисунок 8 – Зависимость активности от дня недели публикации

Наиболее информативным оказался дашборд с самыми популярными тематиками групп, на которые подписаны подписчики исследуемого сообщества, представленный ниже.

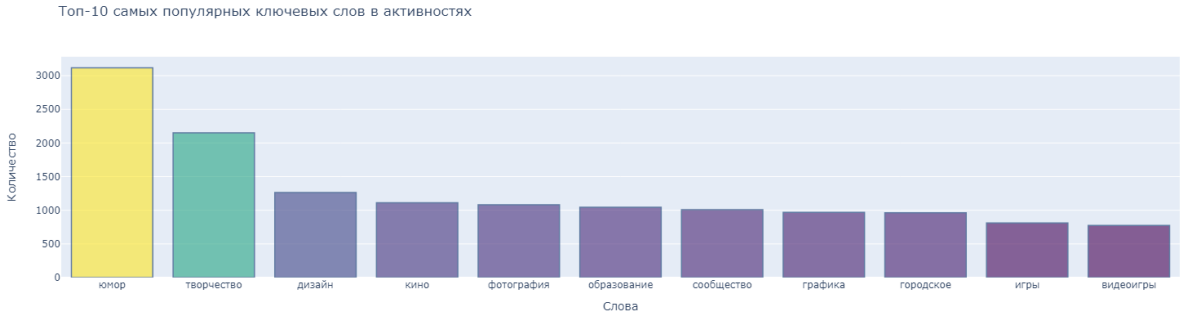

Рисунок 9 – Интересы подписчиков сообщества

Здесь представлена творческая сфера – творчество, дизайн, кино, игры, графика, фотография. Это полностью подтверждает гипотезу о том, что пользователи самой нейросети midjourney – креативные люди, использующие её в работе или хобби.

Сравнивая два сообщества, узкоспециализированного и с широкой целевой аудиторией, становится ясно, что интересы людей из второго больше интересов, не связанных напрямую с темой, посты могут быть более вариативными и зависимость активности от времени публикации меньше.

ЗАКЛЮЧЕНИЕ

В результате выполнения учебной практики был успешно проведён анализ сообществ в тематике нейросетей и LLM. Благодаря тщательному подбору сообществ для анализа и эффективному использованию инструментов Reddit API, VK API, а также библиотек Python, удалось собрать и обработать значительный объем информации о пользователях и постах.

Анализ этих данных не только выявил текущие тенденции и предпочтения аудитории, но и предоставил ценные инсайты для разработки стратегий создания и развития новых сообществ.

В ходе работы были успешно проверены и подтверждены выдвинутые гипотезы, что подчеркивает значимость и актуальность выбранной темы исследования. Описание, анализ и визуализация данных с использованием современных библиотек обеспечили наглядное представление результатов, что делает данное исследование не только научно значимым, но и практически применимым для специалистов в области социальных сетей.

Были выполнены все задачи, поставленные в рамках работы над проектом, а именно:

- поиск подходящих для анализа сообществ и выдвижение гипотез;
- изучение документации Reddit API и соответствующей библиотеки на Python (async PRAW), а также VK API;
- написание кода для скрейпинга данных о пользователях и постах;
- описание, анализ и визуализация данных с использованием библиотек;
- структуризация полученных данных и сравнение двух сообществ;
- проверка гипотез.

В ходе учебной практики, я улучшила навыки визуализации данных, работы в команде, решения нестандартных задач, познакомилась с работой с API. Все это, безусловно, пригодится на пути профессионального развития.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Async PRAW// Joel Payne URL: https://asyncpraw.readthedocs.io/en/stable/getting_started/ratelimits.html (дата обращения: 06.02.2024).

2. Документация API VK// ВКонтакте URL: https://dev.vk.com/ru/reference (дата обращения: 09.02.2024).