

Проект №

1.	Название	Изучение качества жизни по цифровым следам пользователей
2.	Руководитель	Щекотин Евгений evgvik1978@mail.ru
3.	Команда	Дунаева Дарья ddo@data.tsu.ru Басина Полина polya.aleskandrovna@yandex.ru Сайфулин Эдуард sayfulin@data.tsu.ru Бакулин Вячеслав slava0505@data.tsu.ru
4.	Задача	В современном обществе по мере расширения и углубления процесса цифровизации появился еще один источник информации о поведении, чувствах и настроениях людей – это цифровые следы, которые пользователи оставляют в Интернете. К их числу относятся поисковые запросы пользователей, информация о посещении сайтов, сведения из социальных сетей (информация на страницах пользователей, публикации и комментарии, лайки и т.д.), сообщения в мессенджерах, геоданные о перемещении пользователей и другие. Цифровые следы могут служить источником информации для изучения широкого спектра социальных, экономических, политических, психологических, правовых и культурных явлений, связанных с деятельностью человека в онлайн-пространстве. Например, этот источник информации о поведении людей в Интернете очень активно используются в современном маркетинге для повышения эффективности рекламы. Для изучения цифровых следов пользователей применяются нереактивные методы исследования, в рамках которых не происходит взаимодействие исследователя с исследуемой социальной группой. Особенность нереактивных методов применительно к Интернету заключается в том, что пользователи выражают свое мнение по собственному желанию, без запроса со стороны исследователя, реагируя на события в оффлайн- (и онлайн-) пространстве. Такое «непрошенное» мнение (unsolicited opinion), то есть суждение, которое не было спровоцировано исследователем, позволяет лучше понять, что действительно волнует людей, что занимает их мысли в настоящий момент времени.
5.	Кейс	Одной из научных проблем, которую можно исследовать, используя цифровые следы пользователей, является проблема изучения качества жизни населения. В фокусе проекта «Изучение качества жизни по цифровым следам пользователей» – субъективные оценки пользователями условий жизни в своем регионе. Проект реализуется на протяжении 3 лет (2019–2021 гг.). Исследованием охвачено порядка 80 регионов РФ. Источником данных служит социальная сеть ВКонтакте, наиболее популярная на территории Российской Федерации. Методология исследования включает в себя последовательную реализацию следующих этапов: 1) разработка модели системы показателей качества жизни (выделены такие показатели: образование, здравоохранение, безопасность, социальное обеспечение, политика, экология, доступность товаров и услуг, ЖКХ и инфраструктура);

		2) отбор региональных сообществ в социальной сети ВКонтакте, в которых обсуждаются различные стороны жизни в данном регионе (отбираются сообщества с количеством подписчиков от 100 до 1 000 000 чел., из которых более 50% проживают в данном регионе и в которых присутствуют публикации, касающиеся тех или иных показателей качества жизни); 3) классификация сообщений и постов в сообществах по тематике; 4) автоматический анализ контента в отобранных сообществах; 5) построение индекса субъективного качества жизни для различных регионов России.
6.	Данные	1) реестр региональных сообществ с указанием их метаданных и цифровых следов пользователей (7 тыс. сообществ); 2) обучающая выборка для автоматической классификации сообщений по категориям качества жизни – 100 тысяч сообщений, размеченных вручную; 3) база сообщений отобранных региональных сообществ за 2019–2021 гг., автоматически классифицированных по категориям качества жизни с помощью алгоритма машинного обучения (8 млн. сообщений); 4) индексы качества жизни населения по регионам РФ.
7.	Научные статьи в рамках проекта (при наличии)	1) Shchekotin E., Myagkov M., Goiko V., Kashpur V. Digital methods of analysis of subjective quality of life: case of Russian Regions // Administratie si Management Public. 2021. № 36. P. 25–48. https://doi.org/10.24818/amp/2021.36-02 2) Shchekotin E., Goiko V., Myagkov M., Dunaeva D. Assessment of quality of life in regions of Russia based on social media data // Journal of Eurasian Studies. 2021. https://doi.org/10.1177/18793665211034185 3) Щекотин Е.В. Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций // Вестник Томского государственного университета. 2021. № 467. С. 170–181. https://doi.org/10.17223/15617793/467/21 4) Щекотин Е. В., Мягков М. Г., Гойко В. Л., Кашпур В. В., Коварж Г. Ю. Субъективная оценка (не)благополучия населения регионов РФ на основе данных социальных сетей // Мониторинг общественного мнения: Экономические и социальные перемены. 2020. № 1. С. 78-116. https://doi.org/10.14515/monitoring.2020.1.05
8.	СМИ о проекте	1) Статистика благополучия https://www.youtube.com/watch?v=JaIwmh1OrFI 2) Цифровое качество жизни населения https://www.youtube.com/watch?v=HP0J_yRAwRg 3) Индекс качества жизни, вычисленный искусственным интеллектом, удивил ученых // Российская газета, 20 октября 2021 г., № 8590, полоса 12. https://rg.ru/2021/10/20/reg-sibfo/indeks-kachestva-zhizni-vydannyj-iskusstvennym-intellektom-udivil-uchenyh.html

Источники данных

Российская Федерация

2018 год

- ✓ 83 региона
- ✓ 258 крупных населенных пункта
- ✓ 2 580 сообществ в социальной сети ВКонтакте
- ✓ Более 8 млн. активных пользователей



2019-2020 год

- ✓ 85 регионов
- ✓ 1 076 населенных пунктов
- ✓ 463 292 сообществ в социальной сети ВКонтакте
- ✓ Более 21 млн. активных пользователей



Составление реестра региональных сообществ



Анализ подписчиков всех сообществ социальной сети ВКонтакте с количеством подписанных пользователей от 100 до 1 млн. человек



Расчет доли подписчиков, указавших в профиле населенный пункт из определенного субъекта Российской Федерации



Оценка соотношения доли подписчиков из конкретного региона с общим количеством подписчиков сообщества



Выгрузка публикаций за необходимый временной промежуток



Очистка сообществ с помощью алгоритма автоматической классификации текстового контента

Разработка и обучение алгоритма автоматической классификации

1

Ручная классификация текстового контента

2019

Сумма размеченных сообщений: 69 032

Размечено мусорных сообщений: 31 862

Размечено по категориям: 33 640

2020

Сумма размеченных сообщений: 70 751

Размечено мусорных сообщений: 63 950

Размечено по категориям: 6 800

2

Предварительная обработка данных и обучение модели

Очистка текста от знаков пунктуации, различных тегов, эмодзи и нерелевантных слов

Перевод текста в нижний регистр

Удаление стоп-слов, к которым относятся предлоги, причастия, междометия, цифры, частицы

Реализация стемминга из библиотеки Natural Language Toolkit

Данные приводились к векторному виду с помощью метода TF-IDF

Обучение модели