Естественный отбор: проранжируй комментарии с помощью ML

Помоги команде по работе с данными разработать механизм ранжирования комментариев к постам на основе методов машинного обучения для соцсети ВКонтакте. Это супер-важный проект: пользователи хотят видеть, в первую очередь, топовые комментарии, повышающие ценность контента. Успехов!

Оглавление

Введение 3

Применение Natural Language Processing для анализа текста 7

Карьерные возможности в VK 12

Примеры применения нейросетей и машинного обучения ВКонтакте 6

О компании 10

Введение

Во время обеденного перерыва в общей зоне отдыха компании VK сидели четыре человека — дата сайентист Юлия[1], аналитик данных Андрей и два стажера, Римма и Юрий, сотрудники команды по исследованиям искусственного интеллекта. Они играли в настольные игры: сейчас вся четвёрка с большим энтузиазмом составляла слова из одного длинного слова – «распределение».

— И кто выбирал контрольное слово? Одни «е» и «р», — потирая лоб от напряжения, прервал молчание Андрей.

— Конечно же я, — с ноткой иронии сказала Юля. — У меня была идея дать слово «ранжирование», но кажется из него можно составить ещё меньше слов, — и она вздохнула, вертя в руках карандаш. — Просто мне не дает покоя одна рабочая задача — вот оттуда и слова.

— А что за задача? — спросила Римма.

— Это новый и интересный проект, связанный с ранжированием комментариев к постам, — ответила Юля. — Ведь комментарии низкого качества ухудшают общее восприятие поста и сайта в целом. Именно поэтому важно ранжировать комментарии пользователей, а впоследствии и модерировать их, чтобы вначале показывать только те, которые действительно дополняют ценность контента.

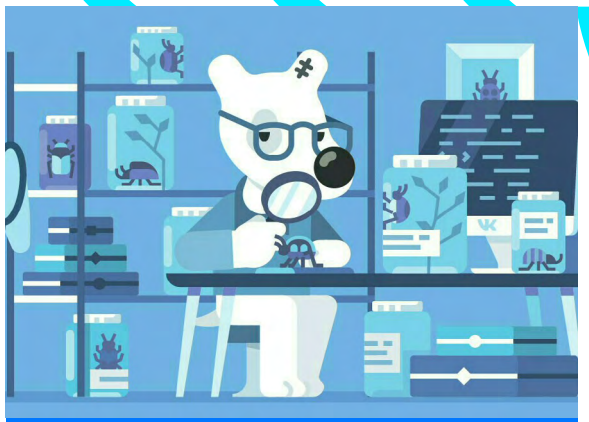
— Но как мы будем это делать? — спросил Юрий.

— Я думаю, что вначале нам нужно создать прототип сервиса по ранжированию комментариев на основе открытых данных, — сказала Юлия. — Это позволит нам иметь уже какие-то наработки к тому времени, когда получим от коллег данные внутренних исследований. Вдобавок, значительно ускорит процесс работы над проектом. Мы использовали такой подход при работе над автоматическим переводом постов.

— Да, ещё мы можем использовать различные инструменты для предобработки текста: например, заменить ссылки на специальный токен, — добавила Римма.

— Я слышал, что для ранжирования комментариев можно применять и методы машинного обучения, — сказал Юрий. — А какую модель будем использовать мы?

— Так как наша большая задача состоит из двух более мелких — представления текста и затем ранжирования, то для первой можно обратиться к готовым моделям, например с Hugging Face, — предложила Юлия.



VK делится опытом и лучшими практиками с рынком и влияет на развитие всей индустрии. Так, программа тестирования VK Testers, объединяющая свыше 30 тысяч начинающих, опытных и продвинутых тестировщиков с января стала доступна всем компаниям на платформе VK Cloud. Сервис позволяет найти ошибки, которые не выявили на предыдущих этапах тестирования, и ускоряет процесс вывода продукта на рынок[2].

[1] Все имена и названия вымышленные, данные кейса могут быть изменены в целях конфиденциальности.
[2] Источник: Lenta.ru, 2023

Введение

— А что насчёт нейронных сетей? — спросил Андрей. — Сейчас почти все дата сайенс проекты используют их, даже популярный чат-бот.

— Нейронные сети могут быть хорошим выбором, но они требуют большого количества данных и вычислительных ресурсов, — ответила Юлия. — Но мы обязательно попробуем применить и их тоже. А после того, как мы обучим модель, нам нужно будет проанализировать полученные результаты и сформулировать полезные инсайты о том, что обычно содержит популярный комментарий.

— А по взаимодействию с комментаторами? — спросил Андрей. — Нам бы не помешал какой-нибудь механизм на момент написания комментария, когда мы можем заранее понять и попросить пользователей переписать неудачный комментарий.

— Это хорошая идея, — ответила Юлия. — Начинающим комментаторам надо помочь, чтобы они не писали неинтересные комментарии и ответы (реплаи). Можно подумать как мотивировать человека, чтобы он мог переписать свой комментарий, если наш сервис покажет ему плохое ранжирование.

Таким образом, команда продолжала обсуждать гипотезы и идеи по решению поставленной перед ними задачи. Их цель была ясна — разработать механизм ранжирования комментариев на основе методов машинного обучения и в будущем создать более интересное и вовлекающее сообщество комментаторов.

С первого дня в компании все сотрудники получают доступ к обширной системе тренингов и других инструментов для личностного и профессионального развития. На корпоративной платформе обучения Стади команде доступно более 380 материалов, а при необходимости можно пройти внешнее обучение или поучаствовать в профессиональной конференции по своему профилю.

Введение

Постановка задачи

Предложите механизм сортировки комментариев к постам по их популярности на основе методов машинного обучения, чтобы модель могла как можно лучше ранжировать пользовательские комментарии.

Для этого:

1. Проведите проверку и разведочный анализ данных (EDA) и подумайте, какие готовые решения можно использовать для представления текста.

2. Используя тренировочную и тестовую выборки датасета, обучите модель ранжировать текстовые комментарии в порядке их популярности (от популярных к менее популярным) и проведите валидацию своей модели. Выбор модели необходимо аргументировать, основываясь на результатах обучения.

3. Проанализируйте полученные результаты и сформулируйте полезные инсайты о том, что обычно содержит популярный комментарий, чтобы команда VK могла использовать эту информацию для улучшения комментариев своих пользователей.

4. Предложите методы взаимодействия с комментаторами, а также механизмы поддержки для разных групп пользователей, включая тех, чьи комментарии непопулярны.

Примеры применения нейросетей и машинного обучения ВКонтакте

Для того, чтобы вдохновиться перед началом нового проекта, у Юлии и Андрея была традиция вспоминать успешные кейсы из недавнего прошлого.

ВКонтакте использует нейронные сети с 2015 года — на тот момент самым большим проектом в этой области была алгоритмическая лента, ведь компания заинтересована в том, чтобы что-то наиболее релевантное для пользователя появлялось на ней повыше, а неинтересное — наоборот.

В 2020 году ВКонтакте стала единственной из крупных соцсетей в России, кто запустил расшифровку голосовых сообщений. Сообщения пользователей прогонялись через нейросети с использованием нескольких алгоритмов, включая акустический, языковой и пунктуационный. Основной особенностью проекта стало применение собственного решения, а не готового, что позволило компании стать не только пионером, но и занять ведущую позицию по распознаванию голосовых сообщений на русском языке.

В том же году социальная сеть запустила нейросеть для борьбы с оскорблениями в комментариях. Она удаляет комментарии с угрозами и рекомендует пользователям перед публикацией отказаться от оскорблений.

В 2021 году пользователям ВКонтакте стал доступен автоматический перевод размещённых в сообществах публикаций с русского на английский язык. Надёжность машинного перевода и определения языка материалов обеспечивает собственная технология, основанная на нейросетевой модели, обученной на открытых данных. Она адаптирована под манеру общения пользователей социальной сети и учитывает особенности лексики в разных сообществах: позволяет переводить на уровне специализированных сервисов как художественные тексты, так и разговорные фразы или официально-деловые публикации.

Кроме того, нейросети помогают создавать более эффективную и персонализированную рекламу, автоматизировать службу поддержки клиентов, делать автоматические заголовки для новостей, использовать рекомендательные системы для подбора похожих плейлистов в VK Музыке.

Недавно ВКонтакте представила сервис для генерации обложек на основе пользовательских интересов: подписок на различные сообщества, реакций и пр. Искусственный интеллект сгенерирует сразу несколько обложек в разных стилях, например футуристичном или мозаичном. После этого останется только выбрать один из вариантов в сервисе и нажать «Установить обложку». После чего изображение автоматически появится в профиле пользователя.

В VK есть 18 корпоративных спортивных команд по 12 видам спорта. 21% сотрудников вовлечены в корпоративный спорт. За регулярные тренировки, участие и победу в соревнованиях в составе одной из команд VK Sport Team сотрудники получают в подарок спортивный мерч.

Применение Natural Language Processing для анализа текста

Прежде чем приступить к созданию работающей модели, Юлия решила рассказать стажёрам основы обработки естественного языка.

Natural Language Processing (NLP) использует искусственный интеллект для обработки письменной и устной речи людей. Основная цель NLP — понимание естественного языка в автоматическом режиме. Использование NLP позволяет решать ряд сложных задач. Например, проблемы ранжирования, когда при поиске учитываются не только слова, но и их смысловое значение. Также это могут быть задачи классификации, где с помощью NLP можно определять тональность отзывов (положительную или отрицательную) и задачи извлечения сущностей, при которых из корпуса текста можно выделять имена, фамилии и даты. Наконец, задачи Seq-2-Seq для машинного перевода тоже могут быть решены с помощью NLP.

— Задачи NLP можно решать через стандартное машинное обучение на основе логистической регрессии или любого другого алгоритма, а также за счёт использования глубокого обучения — сетей прямого распространения или рекуррентных нейронных сетей, — сказала Юлия. — Но преимуществом глубокого обучения, по сравнению со стандартным машинным, является отсутствие необходимости отбора признаков, так как нейронная сеть сама определяет их значимость и оптимальный набор параметров.

Перед применением любой модели текст необходимо представить в виде набора токенов. В задачах NLP под токенами понимают текстовые единицы, на которые можно разбить текст, например символы, слова, предложения, параграфы или абзацы. Процесс разбиения текста на токены называется токенизацией. Выбор вида токена влияет на количество токенов в наборе текстов (его ещё называют корпусом) и на размер словаря, представляющего собой совокупность всех уникальных слов в тексте.

Технологии обработки текстов и определения сходства

Jaccard Similarity (Коэффициент сходства Жаккара) — статистическая величина для оценки сходства и разнообразия, определяемая как размер пересечения, разделенный на размер объединения двух множеств.

Перед подсчетом коэффициента сходства Жаккара необходимо использовать лемматизацию, чтобы привести слова к их словарной форме. При лемматизации существительные приводятся к именительному падежу единственного числа («лица» — «лицо»), а глаголы, причастия и деепричастия — к глаголу в инфинитиве («вижу» — «видеть»). Лемматизацию следует отличать от стемминга, когда слова приводятся к основе слова без окончания («нужно» — «нужн»)

VK ведёт работу по созданию собственного игрового движка – бесплатного и основанного на открытом коде. Компания планирует сделать движок доступным и за счёт понятной документации обеспечить привлечение новых специалистов в отрасль. Бета-версию движка планируют представить в 2024 г[3].

[3] Источник: Cnews, 2023

Применение Natural Language Processing для анализа текста

Разработка VK – платформа in-memory вычислений Tarantool – это open source решение, которое позволяет эффективно работать с высоконагруженными системами. Tarantool в компании используется для динамического контента: сеансов пользователей, мгновенных сообщений и прочего, а клиентам помогает снижать нагрузку на бэкэнд-системы до 85%[4] и ускорять их работу до 5010 раз.

Cosine Similarity (Косинусное подобие) — рассчитывает сходство путем измерения косинуса угла между двумя векторами.

Word2vec — алгоритм, который использует контекст, чтобы сформулировать численные представления слов. Так, для вектора слова «кошка» близким будет «собака», так как они чаще используются в одном контексте.

Программное обеспечение word2vec было разработано в 2013 году компанией Google. Его работа основана на приёме корпуса в качестве входных данных, генерации словаря и последующем вычислении векторного представления слов. Затем метод сопоставляет каждому слову вектор и выдаёт координаты слов на выходе.

TF-IDF — техника, получившая сложное название «частота терминов, обратная частоте документа» (tf-idf), которая придаёт большее значение встречаемости редко встречающихся слов по сравнению с общеупотребительными.

В самом простом варианте алгоритм выглядит следующим образом:

1) Подсчитать частоту для каждого слова или количество присутствия этого слова, делённое на длину документа или предложения. Это «частота термина», или tf. Примечание: при этом знаки препинания и заглавные буквы игнорируются.

2) Подсчитать, в скольких документах встречается каждое слово, и разделить полученное значение на общее количество документов/предложений. Это «частота документа», или df. Если хотят придать меньший вес общеупотребительным словам, то используют обратную величину, 1÷df.

3) Использовать параметр, который бы объединял эти два числа. Существуют разные способы объединения этих двух чисел для присвоения веса каждому слову. Наиболее распространенным является произведение частоты термина и логарифма, обратного частоте документа: tf-idf=tf×log(1÷df).

[4] Источник: Tarantool, 2023

Применение Natural Language Processing для анализа текста

Логика tf-idf здравая: если слово очень часто встречается во всех документах, оно может быть не очень информативно при описании документа, даже если встречается в документе много раз, ведь это могут быть предлоги и союзы. С другой стороны, если слово встречается в корпусе очень редко («эпигенетика», «градиент»), даже одно появление в документе может быть информативным.

В 2022 году VK совместно с НИУ ВШЭ запустили Инженерно-математическую школу, в рамках которой студенты старших курсов работают над реальными технологическими задачами VK вместе со специалистами компании. Занятия проходят в практическом формате мастерских-лабораторий[5].

Рекуррентные нейронные сети и трансформеры. Для решения задач NLP долгое время использовались рекуррентные нейронные сети (Recurrent Neural Networks, RNN), которые умеют обрабатывать потоковые данные. Однако они плохо показали себя при работе с длинными зависимостями, когда для получения связного текста недостаточно перевода отдельных предложений. Исправить проблему помог «механизм внимания» (англ. attention): с его помощью нейросеть оценивала, какая позиция входящей последовательности важна для конкретной позиции последовательности на выходе. Другим недостатком при работе рекуррентных сетей было то, что они требовали последовательных вычислений, что ограничивало возможность применения графических процессоров в ходе обучения моделей.

При дальнейшем развитии идеи рекуррентных сетей ученые из Google Research и Google Brain придумали более продвинутое семейство архитектур машинного обучения — трансформеры (англ. transformers). Трансформеры сочетают в себе параллельную обработку данных, возможность дообучения моделей и широкое применение механизма внимания. При этом нейросеть-трансформер состоит из энкодеров и декодеров. Энкодер извлекает информацию из входящей текстовой последовательности. Декодер использует извлеченную информацию для генерации элементов последовательности на выходе, например текста на другом языке.

Один из наиболее известных трансформеров — языковая модель BERT, разработанная Google в 2018 году. Она состоит из простого набора блоков-трансформеров, который был предварительно обучен на большом корпусе текстов, состоящем из 800 миллионов слов из англоязычных книг и 2,5 миллиарда слов из текстов статей английской Википедии (без разметки). Базовая BERT содержала 110 миллионов параметров, а расширенная — 340 миллионов.

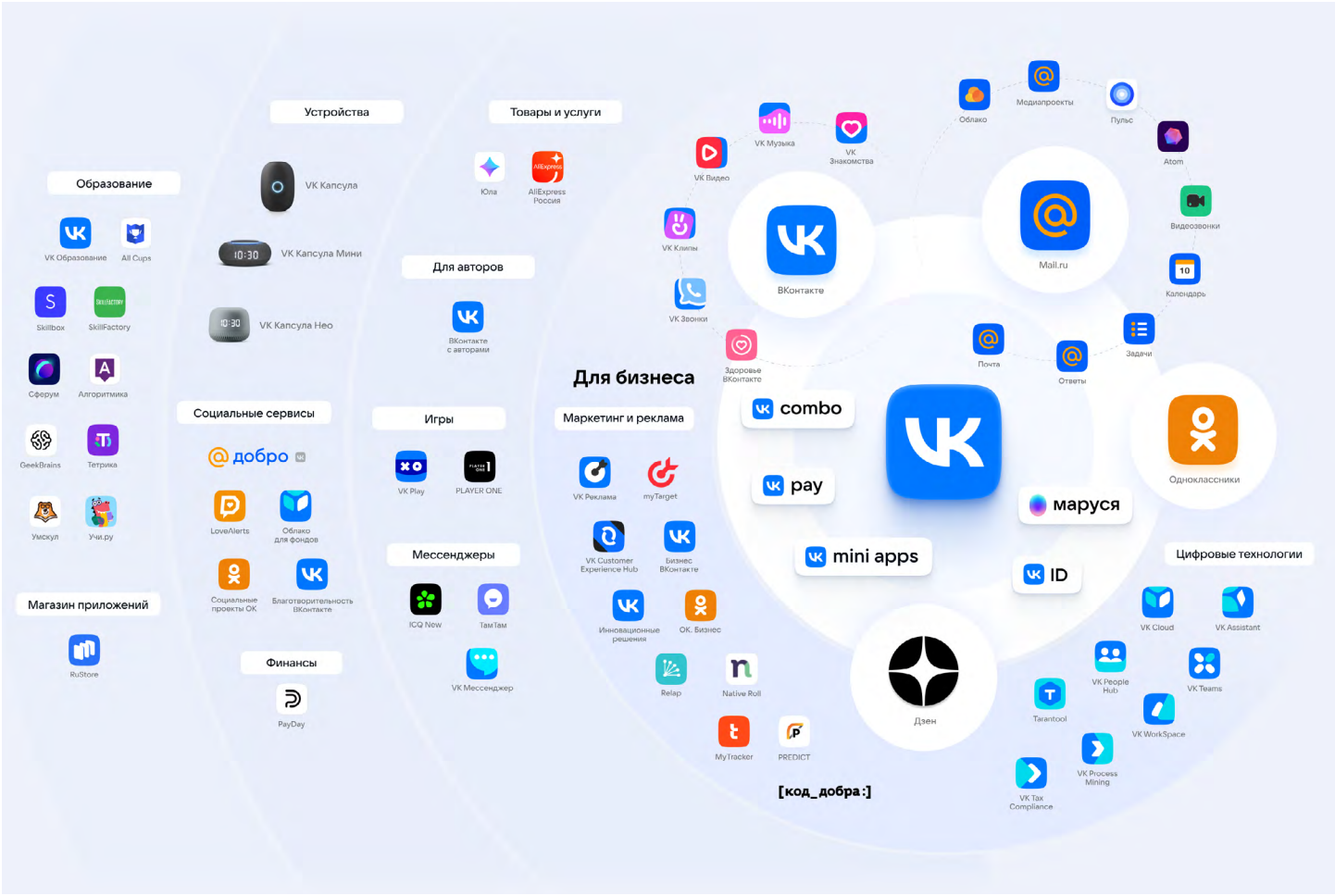
В 2019 году VK запустила образовательную программу в области работы с большими данными – Академию больших данных MADE. Обучение по одному из трех направлений: Data Scientist, Machine Learning Engineer и Machine Learning Operations Engineer, длится год и включает в себя более 25 дисциплин и 250 занятий[6].

[5] Источник: VK, 2022
[6] Источник: Habr, 2022

О компании

VK — крупнейшая российская технологическая компания. Продуктами и сервисами VK пользуются больше 90% аудитории рунета. Они помогают миллионам людей решать повседневные вопросы онлайн: проекты VK позволяют общаться, играть, учиться и осваивать новые профессии, слушать музыку, смотреть и снимать видео, вести бизнес, продвигать своё творчество. В команде VK работают больше 10 000 сотрудников.

В январе 2023 г. VK подключила свою bug bounty-программу к BugBounty.ru и стала первой компанией, которая разместилась на всех трех российских платформах по поиску уязвимостей (bug bounty). В 2022 году VK приняла более 750 отчетов, общий объем выплат превысил 13 миллионов рублей[7].



Экосистема VK

[7] Источник: VK, 2023

О компании

Платформенные решения компании – VK ID, VK Pay, VK ID, VK Combo, VK Mini Apps и голосовой помощник Маруся – связывают многочисленные сервисы VK и позволяют авторизоваться в разных сервисах с единой учетной записью, оплачивать покупки и пользоваться скидками от компании и партнёров.

VK также развивает продукты и услуги для бизнеса, помогая компаниям с цифровизацией бизнес-процессов. Многолетняя экспертиза VK в развитии интернет-сервисов позволяет эффективно решать задачи партнёров, начиная с интернет-продвижения и до внедрения облачных сервисов и ускорения работы IT-систем.

Основные направления технологических решений в VK:

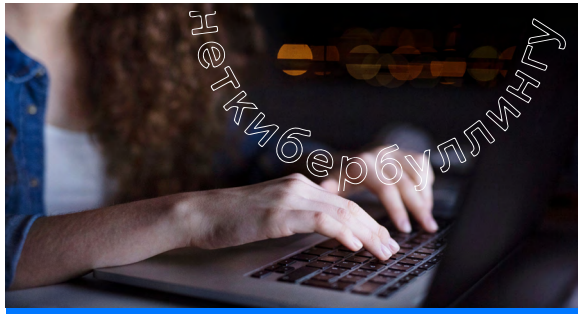
- облачные сервисы и инфраструктура для бизнеса и разработчиков;
- технология In memory, ускоряющая информационные системы за счет хранения и обработки данных в оперативной памяти;
- искусственный интеллект и машинное обучение для роботизации бизнес-процессов.

Ещё одно важное для VK направление — образование. Компания делает продукты для пользователей всех возрастов, охватывая образование на всех ступенях – от школьного до профессионального. В периметр входят образовательные сервисы Skillbox, GeekBrains, SkillFactory, Учи.ру, Умскул, Тетрика, Сферум и другие.

В образовательных проектах для молодёжи VK делает акцент на IT-специальности. Образовательные центры VK есть в 15 ведущих технических вузах России. Студентам ежегодно доступно 70 бесплатных курсов и программ, 70% каждой программы — практика. Это даёт возможность осваивать современные профессии, приобретать новые и совершенствовать навыки в IT и digital.

Узнавать о компании и технологиях студентам помогают амбассадоры VK. Это молодые люди от 16 до 35 лет, которые организуют мероприятия, прокачивают знания в IT-сфере, запускают масштабные проекты. Амбассадоры проходят образовательную программу и общаются с наставниками и трекерами, развиваясь в выбранной сфере.

В 2022 году компания VK вошла в число лучших работодателей страны согласно рейтингу Forbes в 2022 году. VK выделяет забота о пользователях и команде, а также ответственное отношение к окружающей среде и эффективный аппарат управления[8].



VK реализует социальные проекты и с 2019 года ежегодно проводит День борьбы с кибербуллингом 11 ноября. Инициативу поддержали более 30 компаний, а сайт kiberbulling.net набрал 130 млн просмотров[9].

VK активно развивает проекты в сфере образования на всех ступенях – от школьного до профессионального. 12 млн получили новые знания на EdTech-платформах VK. В 2022 году образовательный холдинг Skillbox Holding Limited стал лидером российского EdTech-рынка (данные исследовательского агентства Smart Ranking).

[8] Источник: Forbes, 2022
[9] Источник: VK, 2023

Карьерные возможности в VK

Стажировка VK — это возможность поработать над продуктами, которыми пользуются миллионы людей в команде VK под руководством наставников и с поддержкой HR-экспертов.

Стажировка VK проходит в офисах и онлайн и длится от 2-5 месяцев. Она занимает от 20 часов в неделю в учебное время до 40 часов летом. С первых дней стажеры получают реальный продуктовый и технологический опыт, знакомятся с компанией, развивают свои навыки и набираются опыта для старта карьеры в IT.

Каждому участнику предстоит принять вызов и реализовать решение в одном из проектов компании при поддержке опытного профессионала из команды. В конце стажировки проходит защита проектов, выпускники получают сертификат и фирменный мерч VK, а лучшие стажеры остаются в команде VK.

Направления стажировки:

- Разработка
- Аналитика
- Маркетинг
- Дизайн
- Управление проектами
- Игровое направление
- Информационная безопасность

Следи за обновлениями на сайте, чтобы подать заявку. Выбирай направление и проект в соответствии со своими интересами и возможностями и начинай карьеру в IT вместе с VK.

CHANGELLENGE >>

Уважаемые участники!

Добро пожаловать в первый тур кейс-чемпионата Changellenge >> Cup IT 2023!

В командах по 3-5 человек вам предстоит решить задание кейса.
Перед тем как приступить к подготовке решения, мы рекомендуем:

1

Изучить страницу

первого тура чемпионата

Там вы найдете ответы на все вопросы по решению кейса, а также ключевые ссылки.

Также следите за обновлениями в Телеграм-канале чемпионата по ссылке.

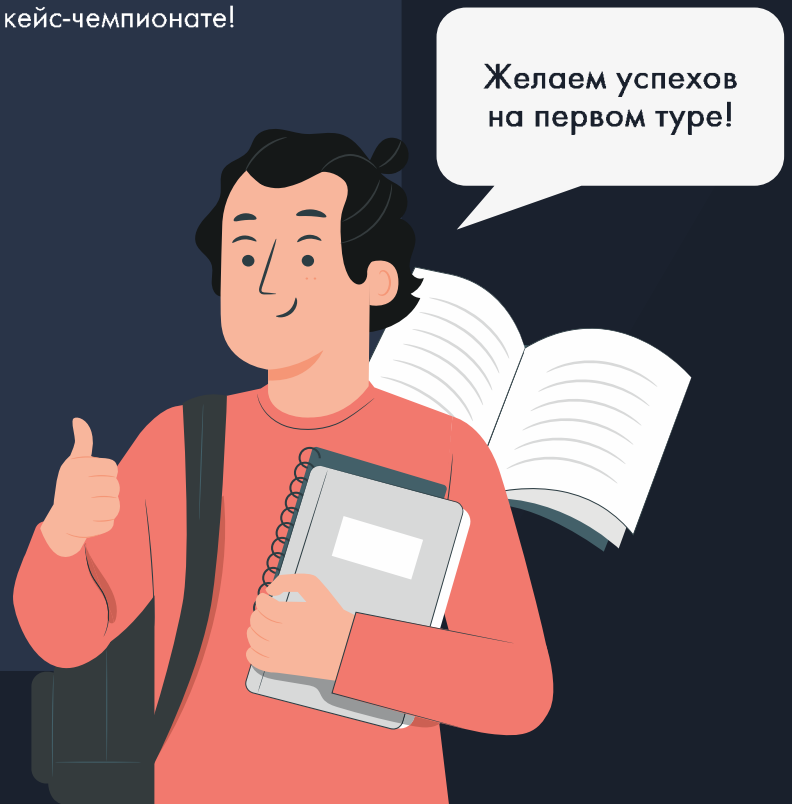
2

Воспользоваться инструментами поддержки кейсера:

- Учебником по решению кейсов: доступен на странице первого тура
- Полезные материалы для кейса: опубликуем в Телеграм-канале чемпионата
- Консультация – Q&A-вебинар с экспертами по кейсу: состоится в Zoom
- Ответы на вопросы по кейсу: в специальном Телеграм-чате
- Если у вас есть организационные вопросы, задайте их тут.

3

Выстроить командную работу. Помните важное правило: вы — команда и успешная организация на старте очень поможет вам в ходе участия в кейс-чемпионате!



CHANGELLENGE >>

Кейс написан и опубликован Changellenge >> — ведущей организацией по кейсам в России.

www.changellenge.com
info@changellenge.com
vk.com/changellengeglobal

Кейс создан по заказу компании VK

www.vk.company