От Шубхама:
Моей матери Гаятри,
которая всегда верила в меня

От Сандры:
Посвящаю эту книгу Руи
за его бесконечную поддержку во всем

Отзывы о книге

Эта книга – идеальная отправная точка для практиков и разработчиков, которые хотят освоить языковую модель GPT-3 и научиться создавать приложения на API OpenAI.

– Питер Велиндер,
вице-президент по продукту и партнерским отношениям, OpenAI

Главная особенность этой книги в том, что ее могут прочитать люди с самым разным техническим образованием и создать решения мирового уровня с использованием ИИ.

– Ноа Гифт,
исполнительный директор Университета Дьюка,
основатель Pragmatic AI Labs

Если вы хотите использовать GPT-3 или любую другую большую языковую модель для создания своего приложения либо службы, в этой книге найдется все, что вам нужно. В книге подробно рассматривается GPT-3, и примеры использования помогут вам применить эти знания к вашему продукту.

– Дэниел Эриксон,
основатель и генеральный директор Viable

Авторы проделали замечательную работу по изучению технических и социальных аспектов использования GPT-3. Прочитав эту книгу, вы будете уверенно рассуждать о современном состоянии искусственного интеллекта.

– Брэм Адамс,
основатель Steganography

Отличная книга для начинающих! В ней даже есть мемы и очень нужная глава об ИИ и этике, но ее главное достоинство – пошаговые процедуры работы с GPT-3.

– Рикардо Хосе Лима,
профессор лингвистики Университета Эстадо-ду, Рио-де-Жанейро

Это всестороннее глубокое погружение в работу с одной из ключевых генеративных моделей обработки естественного языка с практическим акцентом на том, как использовать API OpenAI и интегрировать его в ваши собственные приложения. Помимо очевидной технической ценности, я считаю особенно важными изложенные в последних главах мысли в отношении предубеждений и конфиденциальности моделей и их роли в демократизации ИИ.

– Рауль Рамос-Поллан,
профессор искусственного интеллекта
Университета Антиокии в Медельине, Колумбия

Содержание

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com, указав название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Благодарности

От Сандры

Я хочу выразить признательность моему соавтору Шубхаму, который пригласил меня сотрудничать с ним в работе над этой книгой и постоянно оказывал мне огромную поддержку.

Я также хочу выразить огромную благодарность нашим техническим редакторам Даниэлю Ибаньес и Маттеусу Танха, которые помогли нам окончательно оформить идею, а также Владимиру Алексееву и Натали Пистунович, которые дали нам отличные предложения по техническим правкам.

Большое спасибо следующим организациям и отдельным лицам в сообществе GPT-3, которые согласились поделиться с нами своим опытом, помогая написать главы 4 и 5 и разобраться в продуктовой экосистеме GPT-3: Питеру Велиндеру из OpenAI, Доминику Дивакаруни и Крису Ходер из Microsoft Azure, Дастину Коутсу и Клэр Хельме-Гизон из Algolia, Клэр Берд из Wing VC, Дэниелу Эриксону из Viable, Фрэнку Кэри и Эдварду Саатчи из Fable Studio, Брэму Адамсу из Stenography, Петру Грудзеню из Quickchat, Анне Ванг и Шегуну Отулане из Copysmith, Мустафе Эргиси из AI2SQL, Джошуа Хаасу из Bubble, Дженни Чу и Оге де Муру из GitHub, Бакзу Авану и Яннику Килчеру.

Я также благодарю мою мать Терезу, мою сестру Паулину, моего дедушку Тадеуша, мою кузину Мартину и моего супруга Руи, а еще моих друзей и коллег, которые были рядом со мной, когда я была занята писательством.

От Шубхама

Я благодарю моего соавтора Сандру, которая, как идеальный партнер, заполнила пробелы и дополнила мои навыки. Несмотря на трудности, с которыми мы столкнулись при написании этой книги, мы испытали огромное удовольствие от работы благодаря способности Сандры превращать даже самые стрессовые ситуации в приятные.

Наши технические редакторы Даниэль Ибаньес и Маттеус Танха сыграли решающую роль в том, чтобы дать нам отличную обрат-

ную связь о том, где стоит поднажать и где вовремя остановиться. Огромное спасибо команде OpenAI, особенно Питеру Велиндеру и Фрейзеру Келтону, за их постоянную поддержку и советы на протяжении всего пути. Я также хотел бы поблагодарить всех основателей и лидеров отрасли, с которыми мы беседовали, за их драгоценное время и ценные идеи.

Спасибо моей маме Гаятри, моему отцу Сурешу, моему брату Сараншу и всем моим друзьям и коллегам, которые поддерживали меня на протяжении всего процесса работы над книгой. Отдельная признательность профессорско-преподавательскому составу и основателям Университета Плакша, которые дали мне возможность выйти за рамки повседневной работы. Мое образование и опыт участия в программе Plaksha Tech Leaders Program позволили мне написать эту книгу.

Об авторах

Сандра Кублик – предприниматель в области ИИ, популяризатор и общественный деятель, которая продвигает бизнес-инновации, связанные с ИИ. Наставник и тренер нескольких компаний, занимающихся ИИ, соучредитель программы ИИ-акселераторов для стартапов и сообщества хакатонов ИИ Deep Learning Labs. Она является активным представителем сообщества NLP и генеративного ИИ. Ведет канал на YouTube, где берет интервью у различных действующих лиц экосистемы стартапов и обсуждает новаторские тенденции в области искусственного интеллекта с помощью забавного и образовательного контента.

Шубхам Сабу занимался разными видами деятельности, от специалиста по данным до консультанта по ИИ в известных фирмах по всему миру, где участвовал в разработке общеорганизационных стратегий работы с данными и технологической инфраструктуры для создания и масштабирования практики обработки данных и машинного обучения с нуля. Его работа в качестве популяризатора ИИ привела к появлению собственной широкой аудитории, где он продвигает идеи применения ИИ. Движимый страстью к изучению нового и обмену знаниями с сообществом, он ведет технические блоги о достижениях в области ИИ и экономических последствиях этого. В свободное время путешествует по стране, что позволяет ему погрузиться в разные культуры и развить свое мировоззрение на основе опыта.

Предисловие

Знаменитая GPT-3, или Generative Pretrained Transformer 3, представляет собой большую языковую модель на основе архитектуры Transformer, разработанную OpenAI. Она состоит из ошеломляющих 175 млрд параметров. Любой желающий может получить доступ к этой огромной языковой модели через API OpenAI – простой в использовании пользовательский интерфейс «текст на входе – текст на выходе» без каких-либо серьезных технических требований. Это первый случай в истории, когда модель искусственного интеллекта такого масштаба была размещена на удаленной платформе и доступна для широкой публики с помощью простого вызова API. Этот новый режим доступа называется «модель как услуга» (model-as-a-service, MaaS). Из-за этого невиданного ранее режима доступа многие люди, включая авторов этой книги, рассматривают GPT-3 как первый шаг к демократизации искусственного интеллекта (ИИ).

С появлением GPT-3 стало проще, чем когда-либо, создавать приложения ИИ. Эта книга в деталях покажет вам, как легко начать работу с API OpenAI. Кроме того, мы познакомим вас с инновационными способами использования этого инструмента в разных областях. Мы рассмотрим успешные стартапы, созданные на основе GPT-3, и корпорации, использующие его в своей продуктовой линейке, а также обсудим проблемы и перспективы развития.

Эта книга предназначена для людей с любым образованием и любого рода занятий, а не только для технических специалистов. Она будет особенно полезна, если вы:

- специалист по обработке данных, желающий приобрести навыки в области ИИ;
- предприниматель, который хочет построить следующий проект в области ИИ;
- руководитель компании, который хочет расширить свои знания об искусственном интеллекте и использовать их для принятия ключевых решений;
- писатель, подкастер, менеджер социальных сетей или другой создатель языковых продуктов, желающий использовать лингвистические возможности GPT-3 в творческих целях;

- любой, у кого есть идея, основанная на искусственном интеллекте, которая когда-то казалась технически невозможной или слишком дорогой для реализации.

Первая часть книги посвящена основам API OpenAI. Во второй части книги мы исследуем пеструю экосистему, органично и стремительно возникшую вокруг GPT-3.

В *главе 1* изложен контекст и основные определения, необходимые для комфортного изучения дальнейших тем. В *главе 2* мы глубоко погружаемся в API, разбивая его на наиболее важные элементы, такие как базовые модели и конечные точки, описывая их назначение и способы использования для читателей, которые хотят взаимодействовать с ними на более глубоком уровне. *Глава 3* содержит простой и интересный рецепт для вашего первого приложения на базе GPT-3.

Затем, переместив акцент на увлекательную экосистему ИИ, в *главе 4* мы берем интервью у создателей некоторых из самых успешных продуктов и приложений на основе GPT-3 и спрашиваем их о проблемах и опыте взаимодействия с моделью в коммерческом масштабе. В *главе 5* будет рассказано, как предприятия относятся к GPT-3 и каков потенциал внедрения этой модели. В *главе 6* мы обсуждаем потенциально проблематичные последствия более широкого внедрения GPT-3, такие как неправомерное использование и предвзятость, а также прогресс в решении этих проблем. Наконец, в *главе 7* мы заглядываем в будущее, знакомя вас с наиболее интересными тенденциями и возникающими возможностями, по мере того как GPT-3 все шире внедряется в коммерческую экосистему.

1

Революция большой языковой модели

«искусство – это обломки от столкновения души и мира»
«технологии стали мифом современного мира»
«революции начинаются с вопроса, но не заканчиваются ответом»
«природа украшает мир разнообразием»

Твиты, сгенерированные нейросетью GPT-3

Представьте, что вы проснулись прекрасным солнечным утром. Сегодня понедельник, и вы знаете, что неделя будет беспокойной. Ваша компания собирается запустить новое приложение для отслеживания личной продуктивности под названием Taskr и начинает кампанию в социальных сетях, чтобы рассказать миру о вашем гениальном продукте.

На этой неделе ваша главная задача – написать и опубликовать серию интересных постов в блоге. Вы начинаете с составления списка дел:

- написать информативную и забавную статью о лайфхаках для повышения производительности с упоминанием о Taskr. Не более 500 слов;
- создать список из пяти броских заголовков статей;
- выбрать визуальное оформление.

Вы нажимаете клавишу ввода, делаете глоток кофе и наблюдаете, как на вашем экране возникает статья, предложение за предло-

жением, абзац за абзацем. Через 30 секунд у вас готов содержательный высококачественный пост в блоге, идеальный старт для вашей серии публикаций в социальных сетях. Современное и красочное визуальное оформление привлекает внимание читателей. Готово! Вы выбираете лучшее название из пяти предложенных вариантов и приступаете к публикации.

Это не фантазия из далекого будущего, а зарисовка новой реальности, ставшей возможной благодаря достижениям в области искусственного интеллекта. Пока вы читаете эту книгу, одно за другим появляются новые приложения для креативной генерации текста и изображений, доступные всем желающим.

GPT-3 – это передовая языковая модель, созданная компанией OpenAI, которая находится на переднем крае исследований и разработок в области искусственного интеллекта. Первый официальный релиз OpenAI, в котором объявляется о создании GPT-3, был выпущен в мае 2020 года, а уже в июне 2020 года был открыт доступ к GPT-3 через API OpenAI. С момента запуска GPT-3 во всем мире были придуманы сотни, если не тысячи интересных применений этой модели в самых разных областях, включая технологии, искусство, литературу, маркетинг… и этот список постоянно растет.

GPT-3 может с невероятной легкостью решать общие языковые задачи, такие как создание и классификация текста, свободно перемещаясь между различными стилями текста и целями. Круг задач, которые она может решить, огромен.

В этой книге мы предлагаем вам подумать о том, какие задачи вы могли бы самостоятельно решить с помощью GPT-3. Мы обещаем рассказать вам, что это за модель и как ее использовать, но сначала хотим получше ввести вас в тему. В оставшейся части данной главы мы обсудим, откуда взялась эта технология, как она устроена, с какими задачами она лучше всего справляется и какие потенциальные риски она несет. Давайте пойдем короткой дорогой и начнем прямо с *обработки естественного языка* (natural language processing, NLP), посмотрим, как с ней связаны *большие языковые модели* (large language model, LLM) и GPT-3.

Что скрывается за кулисами NLP

NLP – это область информационных технологий, посвященная взаимодействию между компьютерами и человеческими языками. Цель исследователей – создать системы, способные эффективно

и качественно обрабатывать естественный язык, с помощью которого люди общаются друг с другом.

NLP сочетает в себе компьютерную лингвистику (моделирование человеческого языка на основе правил) с машинным обучением для создания интеллектуальных машин, способных определять контекст и понимать смысл естественного языка.

Машинное обучение – это ветвь ИИ, в которой исследователи развивают способность машин решать различные задачи с помощью опыта, без явного программирования. *Глубокое обучение* – это область машинного обучения, которая основана на использовании глубоких нейронных сетей, смоделированных по образцу человеческого мозга, для выполнения сложных задач с минимальным вмешательством человека.

Глубокое обучение появилось в 2010-х годах, и спустя некоторое время были созданы большие языковые модели на основе плотных нейронных сетей, состоящих из тысяч или даже миллионов простых рабочих элементов, называемых искусственными нейронами. Нейронные сети стали первым значительным прорывом в области NLP, позволив реализовать сложную обработку естественного языка, что до той поры считалось возможным только в теории. Второй важной вехой стало появление *предварительно обученных моделей* (таких как GPT-3), которые впоследствии можно точно настроить для различных задач, что позволяет сэкономить много часов обучения. (Предварительно обученные модели мы обсудим позже в этой главе.)

NLP лежит в основе многих прикладных применений ИИ, таких как:

Обнаружение спама
: Система фильтрации спама в вашем почтовом ящике использует NLP, чтобы определить, какие электронные письма выглядят подозрительно, и отправить их в корзину.

Машинный перевод
: Google Translate, DeepL и другие программы машинного перевода используют NLP для перевода предложений в почти произвольных языковых парах.

Виртуальные помощники и чат-боты
: В эту категорию попадают чат-боты наподобие Alexa, Siri, Google Assistant и многочисленные службы поддержки клиентов по всему миру. Они используют NLP, чтобы понимать и анализировать смысл обращения, определять приоритетность вопросов и запросов пользователей и быстро и правильно реагировать на них.

Анализ настроений в социальных сетях
: Маркетологи собирают в социальных сетях сообщения о конкретных брендах, темы разговоров и ключевые слова, а затем используют NLP для анализа индивидуального и коллективного отношения людей к бренду. Это помогает брендам в исследовании клиентов, оценке своего имиджа и определении социальной динамики.

Обобщение текста
: Обобщение текста – это уменьшение его размера при сохранении ключевой информации и основного смысла. Наиболее распространенными примерами обобщения текста являются заголовки новостей, анонсы фильмов, информационные бюллетени, финансовые обзоры, анализ юридических контрактов, сводки писем в электронной почте и приложения, доставляющие ленты новостей, отчеты и электронные письма.

Семантический поиск
: Семантический поиск использует глубокие нейронные сети для интеллектуального поиска данных. Вы взаимодействуете с ним каждый раз, когда выполняете поиск в Google. Семантический поиск полезен при поиске чего-либо на основе контекста, а не определенных ключевых слов.

«Мы взаимодействуем с другими людьми посредством языка, – говорит Янник Килчер (https://www.youtube.com/@YannicKilcher), один из самых популярных ютуберов и авторитетов в области NLP. – Язык является частью каждой бизнес-транзакции, любого совместного действия людей, и даже с машинами мы взаимодействуем посредством того или иного языка, будь то программа либо пользовательский интерфейс». Поэтому неудивительно, что компьютерная обработка естественного языка стала источником самых захватывающих открытий и местом самых впечатляющих применений ИИ за последнее десятилетие.

Языковые модели становятся больше и лучше

Моделирование языка – это задача присвоения вероятности последовательности слов в тексте на определенном языке. Основываясь на статистическом анализе существующих текстовых последова-

тельностей, простые языковые модели могут рассматривать слово и предсказывать следующее слово (или слова), которое, скорее всего, последует за ним. Чтобы создать языковую модель, которая успешно предсказывает последовательности слов, вы должны обучить ее на больших наборах данных.

Языковые модели – это жизненно важный компонент приложений для обработки естественного языка. Их можно рассматривать как инструмент статистического прогнозирования, получающий текст на входе и выдающий прогноз на выходе. Наверняка вы хорошо знакомы с этим инструментом в виде функции автозавершения в телефоне. Например, если вы напечатаете слово «добрый», автозавершение предложит варианты «человек», «день» и «путь».

До GPT-3 не существовало общей языковой модели, которая могла бы хорошо выполнять ряд задач NLP. Языковые модели были разработаны для выполнения *одной* конкретной задачи NLP, такой как генерация текста, обобщение или классификация. В этой книге мы обсудим экстраординарные возможности GPT-3 как общей языковой модели. Мы начнем эту главу с того, что познакомим вас с каждой буквой в аббревиатуре «GPT», чтобы показать, что они обозначают и из каких элементов построена знаменитая модель. Мы дадим краткий обзор истории и покажем, как и почему модели преобразования последовательностей, которые сегодня блистают в различных приложениях, достигли такого успеха. После этого мы расскажем вам о важности доступа к API и о том, как он развивался в зависимости от требований пользователей. Мы рекомендуем зарегистрировать учетную запись на сайте OpenAI, прежде чем переходить к остальным главам.

Что скрывается за названием GPT-3?

Название GPT-3 расшифровывается как «Generative Pre-trained Transformer 3» (генеративный предварительно обученный трансформер). Давайте рассмотрим все эти термины по порядку – это поможет нам понять принцип работы GPT-3.

Генеративные модели

GPT-3 – это *генеративная модель*, поскольку она генерирует текст. Генеративное моделирование – это раздел статистического моделирования. Это метод математической аппроксимации мира.

Нас окружает невероятное количество доступной информации – как в физическом, так и в цифровом мире. Сложность заключается в разработке интеллектуальных моделей и алгоритмов, способных анализировать и понимать эту сокровищницу данных. Генеративные модели являются одним из наиболее многообещающих подходов к достижению этой цели[1].

Чтобы обучить модель, вы должны подготовить и предварительно обработать *обучающий набор данных* – набор примеров, которые помогают модели научиться выполнять определенную работу. Обычно обучающий набор представляет собой большой объем данных в какой-то конкретной области: например, миллионы изображений автомобилей, чтобы научить модель распознавать автомобиль на незнакомых картинках. Обучающие данные могут принимать разнообразную форму. Это могут быть, например, текстовые предложения на естественном языке или фрагменты звуковых файлов (сэмплы). После того как вы показали модели множество примеров, она должна научиться генерировать аналогичные данные – в этом предназначение генеративной модели.

Предварительно обученные модели

Вы слышали о теории 10 000 часов? В своей книге «Выбросы: история успеха» Малкольм Гладуэлл утверждает, что отработки любого навыка в течение 10 000 часов достаточно, чтобы стать экспертом[2]. Это «экспертное» знание закрепляется в связях, которые ваш человеческий мозг развивает между своими нейронами. Модель ИИ делает нечто подобное.

Чтобы создать хорошо работающую модель, ее необходимо обучить с использованием определенного набора переменных, называемых *параметрами*. Процесс определения идеальных параметров для вашей модели называется *обучением*. Модель постепенно усваивает значения параметров, проходя через последовательные итерации обучения.

Глубокой модели, состоящей из множества нейронных слоев с миллионами нейронов, требуется много времени, чтобы найти эти идеальные параметры. Обучение – это длительный процесс, который в зависимости от задачи может длиться от нескольких

---

[1] Андрей Карпати (Andrej Karpathy) и др., публикация в блоге о генеративных моделях, источник: https://openИИ.com/blog/generative-models/.

[2] Malcolm Gladwell, *Outliers: The Story of Success* (Little, Brown, 2008).

часов до нескольких месяцев и требует огромных вычислительных мощностей. Очевидно, что нам очень пригодилась бы возможность повторно использовать результаты этого длительного процесса обучения для других задач. И здесь на помощь приходят предварительно обученные модели.

Если продолжить аналогию с теорией 10 000 часов Гладуэлла, то предварительно обученная модель – это базовый навык, который вы развиваете, чтобы легче было перейти к другому навыку. Например, овладение навыком решения математических задач поможет вам быстрее научиться решать инженерные задачи. Сначала модель обучают (вы или кто-то другой) для более общей задачи, а затем ее можно настроить для решения различных частных задач. Вместо того чтобы создавать совершенно новую модель для решения своей задачи, вы можете использовать предварительно обученную модель, которая уже в общих чертах владеет необходимыми «навыками». Предварительно обученную модель можно настроить в соответствии с вашими конкретными потребностями, предоставив дополнительное обучение с помощью специального набора данных. Этот подход намного быстрее и эффективнее и позволяет повысить производительность по сравнению с построением модели с нуля.

Размер набора данных, на которых обучают модель, во многом зависит от задачи, которую вы решаете, и от ваших возможностей собрать или приобрести необходимые данные. Модель GPT-3 обучена на текстовом корпусе из пяти наборов данных: Common Crawl, WebText2, Books1, Books2 и Wikipedia.

Common Crawl

Корпус Common Crawl содержит петабайты данных, включая необработанные данные веб-страниц, метаданные и текстовые данные, собранные за восемь лет сканирования веб-страниц. Исследователи OpenAI используют проверенную и отфильтрованную версию этого набора данных.

WebText2

WebText2 – это расширенная версия набора данных WebText, внутреннего корпуса OpenAI, созданного путем очистки веб-страниц особенно высокого качества. Чтобы гарантировать качество, авторы извлекли данные по всем исходящим ссылкам с Reddit, которые получили как минимум три кармы (индикатор того, что другие пользователи сочли ссылку интересной, познавательной или просто забавной). WebText содержит 40 ГБ текста, извлеченного из этих 45 млн ссылок и более 8 млн документов.

Books1 и Books2

Books1 и Books2 представляют собой два текстовых корпуса, которые содержат тексты десятков тысяч книг по различным предметам.

Wikipedia

Коллекция, включающая все англоязычные статьи из свободной онлайн-энциклопедии Wikipedia (https://en.wikipedia.org/wiki/МИИп_Page) на момент завершения сбора данных GPT-3 в 2019 году. Этот набор данных насчитывает примерно 5,8 млн статей на английском языке (https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia).

В общей сложности обучающий корпус содержит около триллиона слов.

GPT-3 может распознавать и генерировать тексты не только на английском языке. В табл. 1.1 показана первая десятка языков, наиболее широко представленных в обучающем наборе данных GPT-3 (https://github.com/OpenAI/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv).

Таблица 1.1. Десять наиболее широко представленных языков в наборе данных GPT-3

	Язык	Количество документов	Доля от общего кол-ва документов, %
1.	Английский	235 987 420	93,68882
2.	Немецкий	3 014 597	1,19682
3.	Французский	2 568 341	1,01965
4.	Португальский	1 608 428	0,63856
5.	Итальянский	1 456 350	0,57818
6.	Испанский	1 284 045	0,50978
7.	Голландский	934 788	0,37112
8.	Польский	632 959	0,25129
9.	Японский	619 582	0,24598
10.	Датский	396 477	0,15740

Разрыв между английским и остальными языками огромен. Английский язык занимает первое место с 93 % набора данных; немецкий язык, занимающий второе место, составляет всего 1 %, но даже этого достаточно для создания качественного текста на немецком языке с определенным стилем и решения других за-