

Википедия решила облегчить доступ к своим данным для разработчиков искусственного интеллекта (ИИ), предлагая специально подготовленный набор данных. Сотрудничая с платформой Kaggle, Wikimedia Foundation анонсировала публикацию набора структурированных данных на английском и французском языках, предназначенных для тренировки моделей ИИ.

Новый набор данных включает в себя аннотации, краткие описания, ссылки на изображения, информацию из инфобоксов и разделы статей. Однако в нем отсутствуют такие элементы, как ссылки и аудиофайлы. Данные предоставляются в открытой лицензии, что позволяет использовать их для различных целей, включая моделирование, настройку и анализ.

Цель этого шага — предоставить более удобную альтернативу для ИИ-разработчиков, нежели скрейпинг сырых текстов с сайта, что уже приводит к перегрузке серверов Википедии из-за активности автоматических ботов. Wikimedia надеется, что данная инициатива снизит нагрузку на платформу, предоставив разработчикам доступ к более упорядоченной и машиночитаемой информации.