# Question-Answer Dataset

В данном датасете имеется все, что нужно для того, чтобы опробовать свои силы в задаче

вопросно-ответного поиска. Датасет содержит 3 файла с вопросами, а также почти 700 000 слов

очищенного текста с Википедии, с помощью которого были созданы вопросы. Также имеется

файл вопрос-ответ, который содержит пары вопрос-ответ, название статьи в Википедии, оценка

сложности вопроса, рейтинг сложности и имя файла, содержащего статью. Также из датасета

исключены «плохие» вопросы.

# Stanford Question Answering Dataset

Датасет, созданный Стэнфордским университетом. Этот набор создан для понимания

прочитанного текста. Датасет также содержит вопросы, составленные на основе более 500

статей из Википедии, а ответом на каждый вопрос является сегмент текста. База содержит

около 100 000 пар вопрос-ответ.

# SberQuad

Датасет, созданный российской компанией Сбер. Этот набор данных создан для понимания текста, состоит из вопросов, заданных краудворкерами на основе статей с Википедии, где ответом на вопрос является фрагмент текста. Это российский аналог Stanford Question Answering Dataset, но он не был должным образом представлен научному сообществу. На основе данного датасета существует очень мало исследований.

# DaNetQA

Датасет DaNetQA – это набор данных, созданный для задачи бинарной классификации, т.е. ответов на вопросы, с ответом «да» или «нет». Вопросы возникают естественным образом, что значит, что генерируются без ограничений. Набор состоит из трех файлов – test, train, val. В каждом файле находится набор данных в форме (вопрос, созданный человеком/данный, отвечающие на него/ответ «да» или «нет»). Данные были собраны из открытой энциклопедии – Википедия.

Данный датасет является частью проекта Russian SuperGLUE, который является рейтингом NLP-систем. Проект создан с целью научить нейросети работать с русским языком, а также помогает решить любые задачи обработки естественного языка, выбирать наиболее оптимальные метрики для решения конкретных задач.

# RWSD

Также, как и предыдущий датасет, RWSD является частью российского проекта Russian SuperGLUE. Данный датасет направлен на обучение модели различать двусмысленность предложений. Ответом на вопрос также является простой ответ «да» или «нет», но отличается от датасета DaNetQA тем, что данными для ответа являются два предложения, которые незначительно отличаются, но имеют разный смысл. Задача модели – сделать логический вывод, и ответить на поставленный вопрос верно. Для данной задачи не подходит классическое ориентирование на порядок слов и структуру предложения, так как неоднозначность предложения не может быть устранена данными способами.