Датасет для обучения бинарного классификатора

Датасет записывать в формате JSONL.

Общее описание датасета: В датасете должно быть 37.500 примеров, из которых 15.000 «отрицательных» примеров (не по теме помощи с bash-терминала), 15.000 «положительных» примеров (по теме помощи с bash-терминалом), а также 7.500 примеров для тестирования после обучения. Примеры не должны быть однотипными, чтобы классификатор не переобучился

1. **«Отрицательные примеры»:** На каждую тему должно быть по 300 примеров. 75 примеров должны быть написаны без ошибок, 75 примеров должны быть написаны с ошибками. На каждый из 75 примеров (без ошибок) приходится 2 переформулированных примера без ошибок.
2. **«Положительные примеры»:** Всего 5 тем (см. «Техническое задание», функциональные требования, пункт 4), на каждую тему должно быть 3.000 примеров. 750 примеров должны быть написаны без ошибок, 750 примеров с ошибками. На каждый из 750 примеров (без ошибок) приходится 2 переформулированных примера без ошибок.
3. **Тестовые примеры:** Не должны включать в себя те примеры, которые были в обучающей выборке. В тестовой выборке должно быть 3.750 «положительных» примеров и 3.750 «отрицательных» примеров. То есть, по 75 примеров на каждую тему из «отрицательных» примеров обучающей выборки, и по 750 примеров на каждую тему из «положительных» примеров.

Формат примера в датасете: {«text»: «<Текст запроса>», «label»: <0 или 1>} — один пример. Каждый пример хранится в отдельной строке.