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.

1. **«Отрицательные примеры»:** На каждую тему должно быть по 40 примеров. Примеры не должны быть однотипными, чтобы классификатор не переобучился. Из этих 40 примеров 5 должны быть синетическими без ошибок в тексте (парафразинг), 5 должны быть написаны с ошибками в тексте без синтетики, 10 должны быть и синетическими (парафразинг), и с ошибкми.
2. **«Положительные примеры»:** Всего 5 тем (см. «Техническое задание», функциональные требования, пункт 4), на каждую тему должно быть 3.000 примеров, которые также должны быть разнообразными, а не однотипными, что не добиться переообучения. Из этих 3.000 примеров 375 должны быть синетическими без ошибок в тексте (парафразинг), 375 должны быть написаны с ошибками в тексте без синтетики, 750 должны быть и синетическими, и с ошибкми (парафразинг).
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 10 примеров на каждую тему из «отрицательных» примеров обучающей выборки, и по 750 примеров на каждую тему из «положительных» примеров. В каждом из 10 «отрицательных» примеров на тему должно быть 3 с ошибками в тексте, и 3 сгенерированных синтетически. В каждом из 750 примеров на тему должно быть 190, сгенерированных синтетически, и 190 с ошибками в тексте.

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