



Вы работаете с набором данных обзоров продуктов. Сообщалось, что в наборе данных могут быть повторяющиеся отзывы.

**Для методов, которые используются при обнаружении повторяющихся записей в наборе данных, сопоставьте название с описанием.**

Сопоставьте значения из двух списков

Хеширование:

=

Создать уникальный хэш для каждого отзыва на основе его содержания. Если два отзыва имеют одинаковый хэш, скорее всего, они дубликаты.

Точное совпадение

=

Если существуют уникальные идентификаторы, найти повторяющиеся значения в этих полях.

Нечеткое сопоставление

=

Сравнение всех полей каждого отзыва и определение записей, которые имеют идентичное содержимое во всех полях.

Ключевые поля

=

Использование показателей сходства (например, сходство Жаккара, расстояние Левенштейна), чтобы найти похожие, но не точно совпадающие отзывы.

Отправить

Далее >