Обнаружение каверов
музыкальных треков

Для задачи связывания (группирования) каверов и исходного трека мы привлекли алгоритм кластеризации DBSCAN. Этот алгоритм обладает возможностью угадывать количество кластеров, а в нашем случае как раз неизвестно сколько именно у нас есть в наборе данных групп каверы+оригинал.

Определили задачи:

- Очистить текст
- Обработать TfidfVectorizer
- Уменьшить размерность до 2 признаков с помощью PCA
- Обучить модель с помощью алгоритма DBSCAN(eps=1, min_samples=2,
- algorithm='kd_tree')
- Оценить метрикой silhouette_score

Модель предсказала кластеры (количество менялось в зависимости от выбора eps
и min_samples). Но где-то что-то было не так и мы не смогли вывести график и получить удовлетворяющую метрику.
Основные сложности:

- данные в разноязычном тексте
- сложно определить к какому track_id принадлежит песня (нескольким текстам
- присвоен одинаковый track_id)