# Задание 3. Предобработка данных

## Задание

Совместно в группах выполнить предобработку данных для того корпуса текстов, с которым вы работали в Задании 2.

1. Определить необходимые этапы предобработки для ваших данных:
   1. разделение на предложения / токены;
   2. удаление стоп-слов;
   3. удаление знаков препинания, чисел, символов другого алфавита;
   4. удаление коротких / длинных слов;
   5. приведение к одному регистру;
   6. дополнительные методы обработки для ВАШИХ данных.
2. Выполнить поиск похожего текста с данными после предобработки.
3. Сравнить результаты ДО и ПОСЛЕ предобработки.

## Ожидаемый результат

1. Файл jupyter-notebook, который содержит решение все перечисленных выше задач с поэтапным выводом результатов (как было в примере в лекции).
2. Текстовый документ с интерпретацией результатов - словесное описание результатов (дало ли результат изменение параметров векторизатора, нормирование данных).