

Начало

Извлечение текста с
изображения

Приведение всех букв к
нижнему регистру

Удаление знаков
препинания и пустых строк

Очистка текста от
функциональных слов
(союзов, предлогов)

Индексация документа -
модель мешок слов

Определение весов
признаков документа
TF-IDF

Конец