Руководство пользователя   
программы ED08   
для исследования методов   
выявления нечетких дубликатов

Оглавление

[Назначение программы ED08 3](#_Toc388712754)

[Настройка программы 3](#_Toc388712755)

[Задание параметров для методов выявления нечетких дубликатов 4](#_Toc388712756)

[Добавление обучающей выборки 5](#_Toc388712757)

[Проведение исследований 6](#_Toc388712758)

# Назначение программы ED08

Программа ED08 предназначена для автоматизация процесса исследования ряда методов выявления нечетких дубликатов (далее — МВНД), а именно: коэффициента ассоциативности Джаккарда, метода шинглов, коэффициента Жаро и Жаро-Винклера, расстояния Левенштейна, Winnowing. Программа ED08 позволяет производить исследование МВНД по таким показателям качества работы, как полнота (recall), точность (precession), F-мера, процент верно принимаемых решений (accuracy), а также производить настройку методов.

# Настройка программы

Для работы программы необходимо произвести её настройку. Требуется указать словарь стоп-слов, а также файл, где хранится набор разделителей слов. Для этого следует нажать кнопку «Настройка», находящуюся на главном меню (рис. 1).

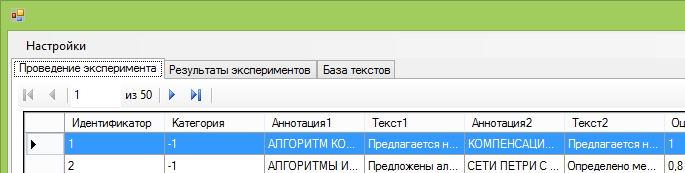


Рисунок 1 — Главное меню программы и кнопка «Настройки»

В появившемся окне «Настройки» (рис. 2) на вкладке «Общие настройки» необходимо нажать кнопку «Изменить» и выбрать необходимые файлы, содержащие словарь стоп-слов и набор разделителей (файлы поставляются вместе с программой и находятся в папке «Словари»).

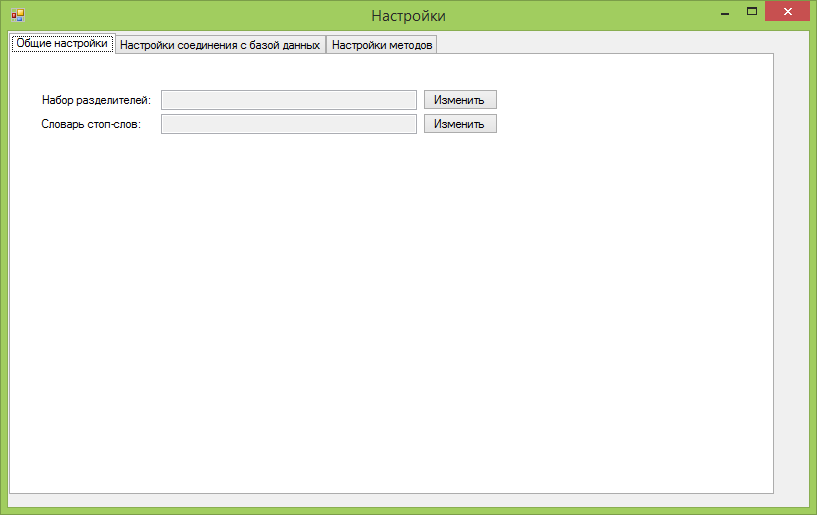


Рисунок 2 — Окно «Настройки» и вкладка «Общие настройки»

Также, в случае необходимости, требуется задать параметры соединения с базой данных (по умолчанию настройка уже произведена, и дополнительных действий со стороны пользователя не нужно). Настройки соединения производятся на вкладке «Настройка соединения с базой данных», находящейся в окне «Настройки» (рис. 3).

Параметры соединения с базой данных можно узнать у лаборантов.

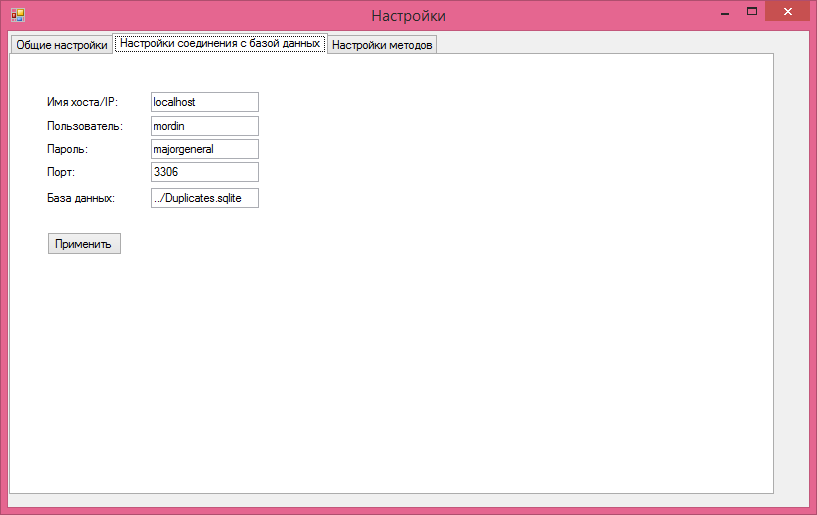


Рисунок 3 — Окно «Настройки»   
и вкладка «Настройка соединения с базой данных»

# Задание параметров для методов выявления нечетких дубликатов

Для задания параметров МВНД необходимо выбрать вкладку «Настройка методов» в окне «Настройки» (рис. 4), задать необходимые значения и нажать кнопку «Применить».

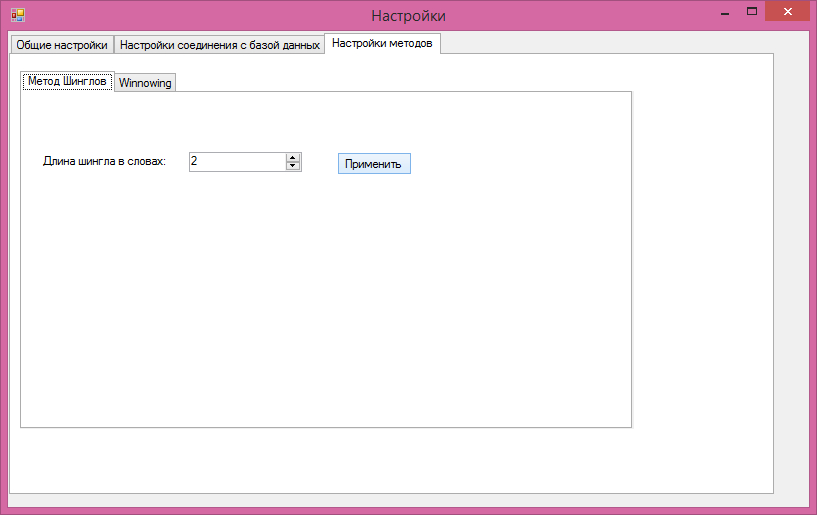


Рисунок 4 — Вкладка «Настройки методов» в окне «Настройки»

# Добавление обучающей выборки

Для добавления обучающей выборки в программу необходимо перейти на вкладку «База текстов» на главной форме (рис. 5) и нажать кнопку .

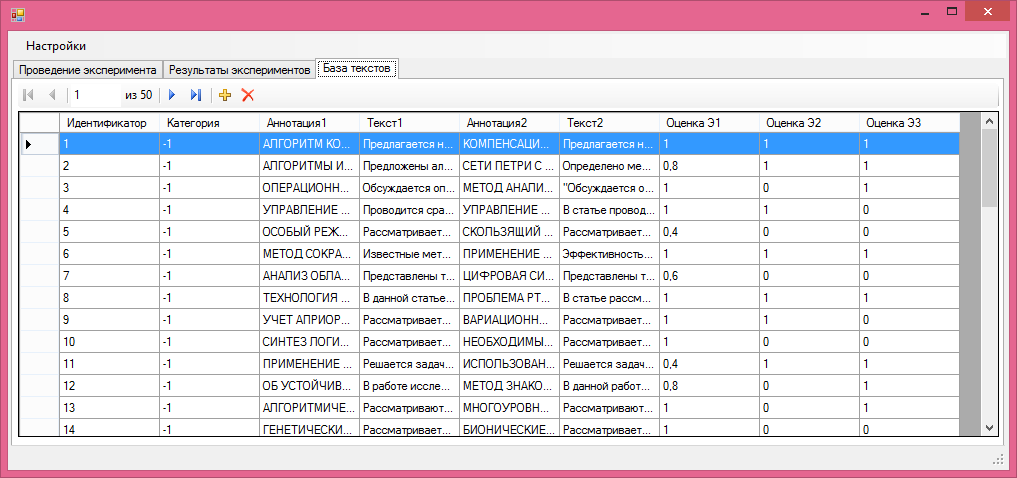


Рисунок 5 — Вкладка «База текстов»

В появившемся диалоговом окне необходимо указать файл с обучающей выборкой (рис. 6). Обучающая выборка должна быть представлена в виде файла формата \*.csv.

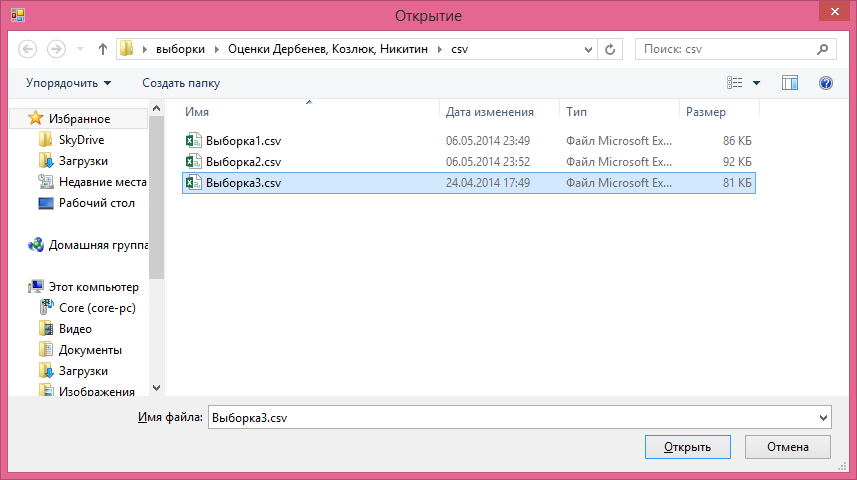


Рисунок 6 — Выбор файла с обучающей выборкой в диалоговом окне

После успешного добавления документов появится окно с указанием количества добавленных в базу документов (рис. 7).

**Внимание!** При добавлении обучающей выборки уже находящаяся базе выборка будет стерта.

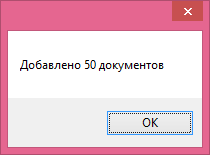


Рисунок 7 — Сообщение о количестве добавленных пар документов в БД

# Проведение исследований

Для проведения исследований необходимо выбрать вкладку «Проведение эксперимента» главного окна приложения.

Необходимо задать параметры эксперимента (рис. 8), указав: исследуемый метод, выборку, пороговое значение, решающее правило для экспертных оценок, значение  для F-меры, а также необходимо ли удалять стоп-слова и производить стемминг. Решающее правило определяет, как будут трактоваться оценки экспертов; рекомендуется использовать «Средний голос > 0,5».

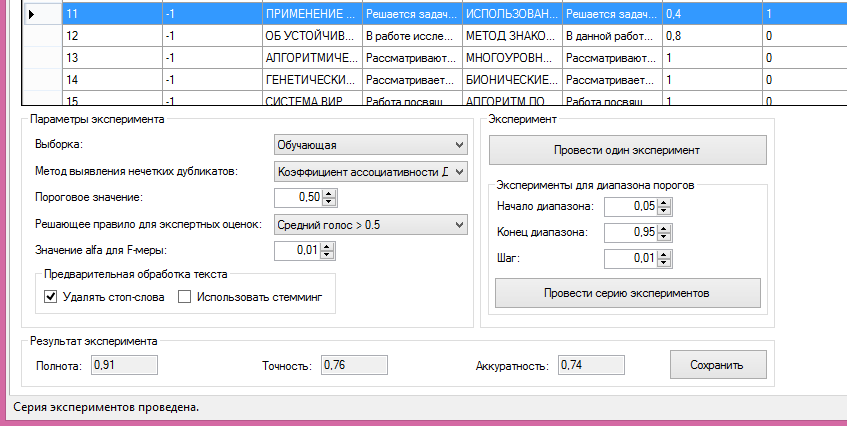


Рисунок 8 — Задание параметров исследования

Чтобы провести эксперимент для заданного порогового значения, необходимо нажать кнопку «Провести один эксперимент». По завершении эксперимента в строке состояния будет выведено сообщение «Эксперимент проведен», а в секции «Результат эксперимента» будут выведены его результаты.

Также существует возможность произвести серию экспериментов для диапазона пороговых значений. Для этого необходимо задать диапазон пороговых значений, а также шаг изменения в секции «Эксперименты для диапазона порогов», а затем нажать кнопку «Провести серию экспериментов». Результаты проведенной серии экспериментов будут представлены на вкладке «Результаты экспериментов» главного окна (рис. 9). Для дальнейшей работы с результатами необходимо их выгрузить файл. Это осуществляется нажатие на кнопку «Выгрузить в CSV».

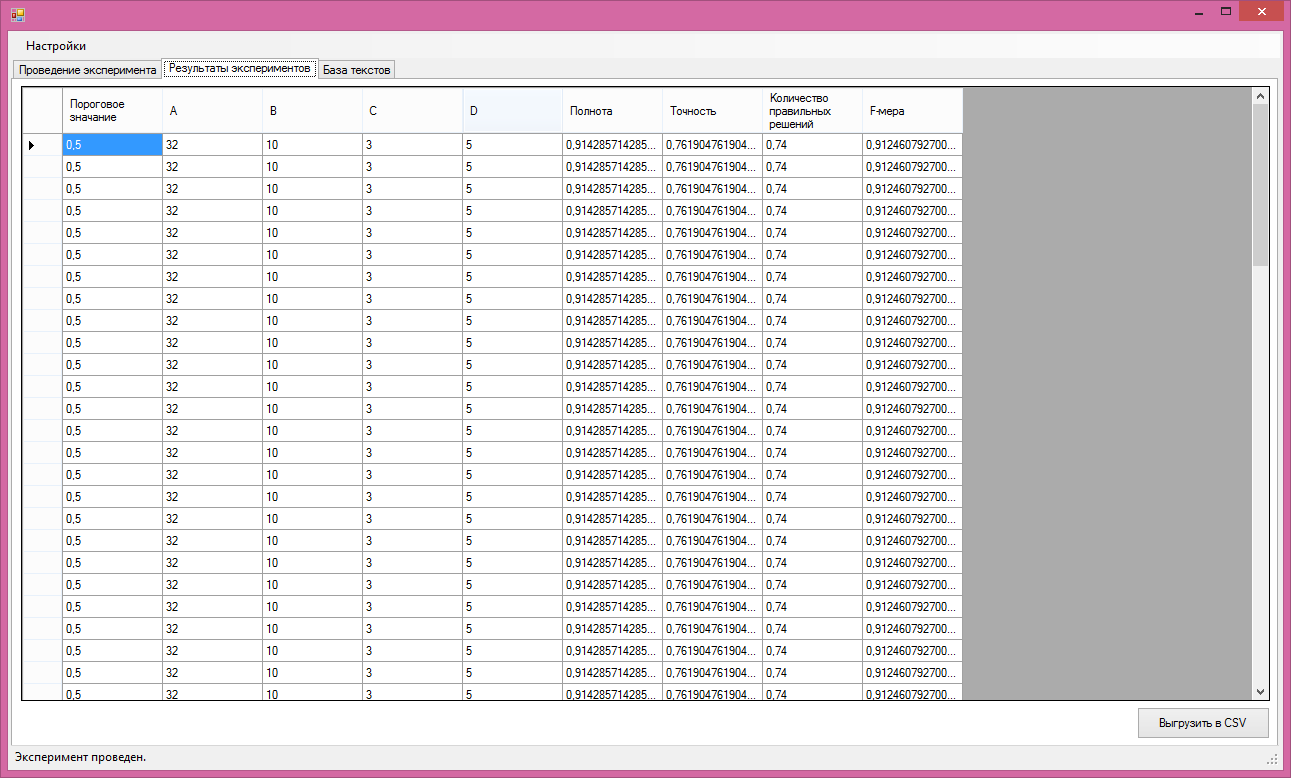


Рисунок 9 — Результаты серии экспериментов