**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**МОСКОВСКИЙ ЭНЕРГЕТИЧЕСКИЙ ИНСТИТУТ**

**Кафедра Управления и Информатики**

Задание и методические указания   
к лабораторной работе по дисциплине   
«Интеллектуальные и информационные системы»   
по теме «Выявление нечетких дубликатов»

Москва, 2014 г.

Оглавление

[Подготовка 3](#_Toc388709087)

[Задание 3](#_Toc388709088)

[Варианты заданий 4](#_Toc388709089)

[Краткое описание методов выявления нечетких дубликатов 4](#_Toc388709090)

[Коэффициент ассоциативности Джаккарда 4](#_Toc388709091)

[Метод шинглов 4](#_Toc388709092)

[Коэффициент Жаро и Жаро-Винклера 5](#_Toc388709093)

[Winnowing 5](#_Toc388709094)

[Расстояние Левинштейна 7](#_Toc388709095)

[Показатели качества работы методов выявления нечетких дубликатов 7](#_Toc388709096)

[Полнота 8](#_Toc388709097)

[Точность 8](#_Toc388709098)

[Аккуратность 9](#_Toc388709099)

[F-мера 9](#_Toc388709100)

# Подготовка

1. Выставить экспертные оценки парам документов, представленным в файле «Выборка.xls», на предмет того, являются ли документы нечеткими дубликатами или нет. Выставляется оценка «0» (документы оригинальные) или «1» (документы — нечеткие дубликаты). Необходимо выставить каждой паре документов одну оценку. Выставленные оценки используются для заполнения столбца «Экспертная оценка 1» в документе «Выборка.xls».
2. Сформировать обучающую выборку. Для этого необходимо взять экспертные оценки, выставленные бригадами с номерами и  (где  — номер бригады), и заполнить ими столбцы «Экспертная оценка 2» и «Экспертная оценка 3». Оценки в столбце «Экспертная оценка 1», занесенные вашей бригадой, остаются.

# Задание

1. Рассчитать значения полноты-точности для методов: коэффициент ассоциативности Джаккарда, коэффициент Жаро, коэффициент Жаро-Винклера, расстояние Левенштейна для различных значениях порога на обучающей выборке. Построить графики для полученных результатов. Выбрать пороговые значения для методов, при котором достигается максимальное значение точности, при значении полноты не менее 85 %
2. Рассчитать значения полноты точности для методов шинглов (длина шингла 3 слова), Winnowing (k = 2, t = 5) при различных пороговых значениях на обучающей выборке. Построить графики для полученных результатов.

Для каждого из методов добиться заданного целевого критерия путем настройки данных методов (изменением параметров методов). Построить графики полнота-точность для полученных оптимальных настройках методов.

1. Получить значения полноты-точности для всех методов на экзаменационной выборке при выбранном пороге и настройках. Сравнить с результатами, полученными на обучающей выборке. Объяснить полученные результаты.
2. Повторить пункты 1-3 при использовании стемминга.

# Варианты заданий

|  |  |
| --- | --- |
| Вар. № | Задание |
| 1 | Добиться максимальной точности при полноте не менее 90% |
| 2 | Добиться максимальной полноты при точности не менее 75% |
| 3 | Добиться максимального значения F-меры (alfa = 35%) |
| 4 | Добиться максимального значения F-меры (alfa = 50%) |
| 5 | Добиться максимального значения F-меры (alfa = 70%) |
| 6 | Добиться максимальной точности при полноте не менее 80% |
| 7 | Добиться максимальной полноты при точности не менее 85% |
| 8 | Добиться максимального значения F-меры (alfa = 25%) |
| 9 | Добиться максимального значения F-меры (alfa = 40%) |
| 10 | Добиться максимального значения F-меры (alfa = 60%) |

# Краткое описание методов выявления нечетких дубликатов

## Коэффициент ассоциативности Джаккарда

Первоначально использовался в математической статистике в качестве меры сходства между объектами, описанными бинарными признаками. Он основан на анализе соотношения совпадающих и признаков у различных объектов. Коэффициент Джаккарда вычисляется по следующей формуле , где A — число терминов, присутствующих в обоих документах, B — число терминов, присутствующих в первом документе и отсутствующих во втором, C — число терминов, присутствующих во втором документе и отсутствующих в первом.

## Метод шинглов

Данный метод основан на расчёте контрольных сумм (шинглов) для каждой последовательности слов в тексте. Последовательности имеют фиксированную длину, которая может настраиваться в зависимости от длины документа. Таким образом, получается набор чисел (шинглов) характеризующих документ. Для проверки на оригинальность документа сравнивают набор шинглов, характеризующий проверяемый документ, с набором шинглов, характеризующим оригинальный документ.

Примечание: суть метода заключается в том, чтобы производить сравнение не последовательности слов, а чисел, однозначно поставленные в соответствие этим последовательностям. Таком образом, производиться операция хеширования данных. В лабораторной работе используется не нахождение контрольных сумм, а стандартная функция хеширования данных SHA-1.

## Коэффициент Жаро и Жаро-Винклера

Расстояния Ж*аро* () и *Жаро-Винклера* () относятся к специализированным расстояниям и рассчитываются следующим образом:

Здесь:  — количество соответствующих символов (символы считаются соответствующими, если они равны и находятся не далее, чем  друг от друга);  — количество перестановок (вычисляется как число соответствующих символов, расположенных в различном порядке, деленное на 2);  — длины сравниваемых строк;  — число общих начальных символов (не больше 4‑х);  —масштабный коэффициент.

## Winnowing

Суть данного метода состоит в том, чтобы сравнивать не тексты документов, а набор значений хеш-функций, характеризующих эти документов. Данный метод напоминает метод шинглов, но главное отличие Winnowing заключается в том, что для описания исходного документа используется не весь набор значений хеш-функций (как в методе шинглов), из которых состоит текст, а только его малая часть.

Логически можно разделить рассматриваемый метод на два этапа: предварительный и основной.

На предварительном этапе из документа удаляются все неинформативные признаки: местоимения, предлоги, союзы и т. д. Далее рассматриваемый документ преобразуется в одну строку путем удаления пробелов:

*A do run run run, a do run run*

*Adorunrunrunadorunrun*

На основном этапе преобразованный документ разбивается на k‑граммы, где параметр *k* задаетсяпользователем. Для примера разобьем преобразованный выше текст на 5-грамм. В результате получим:

*adoru dorun orunr runru unrun nrunr runru*

*unrun nruna runad unado nador adoru dorun*

*orunr runru unrun*

Затем для каждой *i*‑й *k*-граммы вычисляется хэш-фунция. Ниже приведен пример хэшированных k-грамм:

*77 74 42 17 98 50 17 98 8 88 67 39 77 74 42*

*17 98*

Далее набор хэшей *k*-грамм разбивается на т. н. окна размером . t —шумовой порог, т.е. минимальная длина подстроки, при которой общие подстроки не игнорируются. Ниже приведен пример разбиения на «окна» длиной 4:

(77, 74, 42, 17) (74, 42, 17, 98)

(42, 17, 98, 50) (17, 98, 50, 17)

(98, 50, 17, 98) (50, 17, 98, 8)

(17, 98, 8, 88) (98, 8, 88, 67)

( 8, 88, 67, 39) (88, 67, 39, 77)

(67, 39, 77, 74) (39, 77, 74, 42)

(77, 74, 42, 17) (74, 42, 17, 98)

Из каждого j‑го окна выбирается минимальное значение хэш-функции. Если в соседних окнах минимальные значения хэшей одинаковы, в результирующий набор попадает только одно значение. Получаем:

17 8 39 17

На заключительном этапе осуществляется сравнение полученных наборов значений хэш-функции каждого документа. В рамах данной работы сравнение осуществлялось при помощи коэффициента ассоциативности Джаккарда.

Данный алгоритм обладает высокой скоростью работы и гарантирует, что если у двух сравниваемых текстовых документов есть общая подстрока длиной не менее *t*, она будет найдена.

**Примечание.** В примере, где показано разбиение на k-граммы, в качестве синтаксической единицы используется буква, но в проводимых исследованиях в качестве синтаксической единицы используется слово.

## Расстояние Левенштейна

Расстояние Левенштейна относится к т. н. редакционным расстояниям. Считается, что в текст могут вноситься элементарные правки — удаление, вставка и замена слов, — и подсчитывается количество таких правок, необходимое для преобразования одного текста в другой. Если тексты одинаковы, требуется 0 правок. Чтобы привести расстояние Левенштейна к значению от 0 до 1, где 1 означает одинаковые тексты, из 1 вычитается отношение количества правок к суммарному числу слов в текстах. Это соответствует отнесению количества слов, которые не потребовалось править, к их суммарному количеству. Расстояние Левенштейна может быть вычислено очень быстро и часто применяется на практике.

# Показатели качества работы методов выявления нечетких дубликатов

Для оценки качества работы методов применяются различные оценки, основанные на анализе результатов работы системы. При этом «идеальным» алгоритмом считается тот, для которого выводы, сделанные системой, согласуются с мнением экспертов-оценщиков. Существуют следующие метрики оценки качества:

* полнота (recall);
* точность (precision);
* аккуратность (accuracy);
* F-мера.

Эти метрики основаны на матрице классификации, представленной в таблице 1:

Таблица 1 — Матрица классификации

|  |  |  |
| --- | --- | --- |
|  | Является дубликатом | Не является дубликатом |
| Найдено методом | a | b |
| Не найдено методом | c | d |

Здесь а — количество документов, найденных системой и являющихся дубликатами с точки зрения экспертов; b — количество документов, найденных системой, но не являющиеся дубликатами с точки зрения экспертов; c — количество дубликатов, не найденных системой; d — количество документов, которые являются оригинальными и метод расценил их как оригинальные.

## Полнота

Полнота (recall) вычисляется как отношение найденных дубликатов к их общему количеству:



Полнота характеризует способность метода находить дубликаты, но не учитывает количество документов, не являющихся дубликатами, выдаваемых пользователю. Например, если полнота равна 50 %, это значит, что половина дубликатов не найдена.

## Точность

Точность (precision) вычисляется как отношение найденных дубликатов к общему количеству найденных документов:



Точность характеризует способность системы выдавать в списке результатов только дубликаты. Например, если точность равна 50 %, это значит, что среди документов половина — дубликаты, а половина — нет.

## Аккуратность

Аккуратность (accuracy) вычисляется как отношение правильно принятых системой решений к общему числу решений. Формально:



## F-мера

F-мера используется для того, чтобы объединить точность и полноту в одной усреднённой величине. F-мера вычисляется по формуле:

 F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}, \qquad \alpha \in [0,1]. 

При  F-мера придает одинаковый вес точности и полноте и называется сбалансированной, или -мерой (в нижнем индексе принято указывать величину). Выражение для неё упрощается:



F-мера при исследованиях используется как единая метрика, по которой возможно делать вывод о качестве работы метода.