Университет ИТМО, факультет программной инженерии и компьютерной техники

Двухнедельная отчётная работа по «Информатике»: аннотация к статье

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Дата прошедшей лекции | Номер прошедшей лекции | Название статьи/главы книги/видеолекции | Дата публикации (не старше 2022 года) | Размер статьи (от 400 слов) | Дата сдачи |
| 10.09.2025 | 1 | Компьютеры на основе тернарной логики и перспективы их развития. | 12.10.2022 | ~12500 | 24.09.2025 |
| 24.09.2025 | 2 | Алгоритм сжатия информации с использованием нумерации перестановок. | 2025 | ~1750 | 08.10.2025 |
| 08.10.2025 | 3 | Enhanced Regular Expression as a DGL for Generation of Synthetic Big Data | 2023 | ~6800 | 22.10.2025 |
|  | 4 |  |  |  |  |
|  | 5 |  |  |  |  |
|  | 6 |  |  |  |  |
|  | 7 |  |  |  |  |

Выполнил(а) Шалабодов Я.Д. , № группы *Р3110* , оценка

Фамилия И.О. студента не заполнять

|  |
| --- |
| **Прямая полная ссылка на источник или сокращённая ссылка:**  [**https://jips-k.org/digital-library/manuscript/file/38429/01-(1\_16)%2021E01-007-ME-ed(0222)-r1(0227)%20end.pdf**](https://jips-k.org/digital-library/manuscript/file/38429/01-(1_16)%2021E01-007-ME-ed(0222)-r1(0227)%20end.pdf) |
| **Теги, ключевые слова или словосочетания:**  Big Data Analytics, Data Generation Language (DGL), Performance Analysis, Regular Expression, Synthetic Data Generation, Type/format Inference |
| **Перечень фактов, упомянутых в статье:**  1. Основа для языка генерации данных DGL - регулярные выражения.  2. Для DGL предложены важные расширения: data domains, type/format inference, resource reference.  3. Два основных типа доменов для синтеза осмысленных данных: *set* и *range.*  4. Для выборки данных из доменов используются два типа примитивных генераторов: seq() и rng().  5. Поддерживаются различные вероятностные распределения: равномерное, нормальное, Пуассона, Ципфа, для создания скошенных данных, характерных для реальных наборов.  6. В синтаксис регулярных выражений введена новая конструкция — ссылка на ресурс %{name}.  7. На примере бенчмарка TPC-H показано, что предложенный DGL может легко определять значения для отдельных колонок, а также генерировать внешние ключи.  8. Для повышения эффективности генерации предложено и экспериментально проверено RegExp Caching и DB Caching. |
| **Позитивные следствия и/или достоинства описанной в статье технологии.**  1. Расширенные регулярные выражения позволяют точно описывать сложные шаблоны данных, включая ссылки на внешние словари, базы данных и последовательности.  2. Технология позволяет генерировать данные со сложными вероятностными распределениями, внешними ключами и data skew.  3.  Возможность интеграции с файлами упрощает процесс подготовки данных для тестирования и анализа. |
| **Негативные следствия и/или недостатки описанной в статье технологии.**  1. Язык может стать сложным для чтения и сопровождения при описании очень сложных иерархических данных или данных с многочисленными взаимосвязями.  2. Генерация данных на основе сложных регулярных выражений с множественными ссылками на внешние ресурсы и вероятностными распределениями может быть затратной при создании очень больших наборов данных.  3. Для обеспечения сложной логики и семантической целостности потребуется дополнительный уровень валидации или генерации, выходящий за рамки языка, т.к. DGL лучше подходит для определения формата и статистических распределений данных. |