Лабораторна робота №09

«Вивчення основних рандомних моделей. Генерування рандомних текстів»

Теоретичні відомості:

1. конспект лекцій
2. стаття Zipf's and Heaps' laws for the natural and some related random texts2018.pdf
3. стаття Zipf's, Heaps' and Taylor's Laws are Determined by the Expansion into the Adjacent Possible2018.pdf
4. стаття Large-Scale Studies of the Repetition Characteristic for Different Models of Symbolic Sequences2021.pdf
5. довідковий файл recipes_lin&log-monkey_corr.pdf для роботи з допоміжною програмою +monkey_text_generator(letters).

Порядок виконання роботи

1. Вивчити теорію, що стосується основних рандомних лінгвістичних моделей.

2. Теоретична підготовка до цієї роботи і конкретні дослідницькі завдання до неї ті самі, що в лабораторній роботі №11.

3. Згенерувати *по одному тексту* з орієнтовними розмірами L ~ 100 000 слів або L ~ 500 000 букв таких типів:
   а. глобально рандомізований текст на лінгвістичному рівні слів (програма +shuffler(words only) з лабораторної роботи №11; кількість циклів рандомізації N ~ 50 000–500 000, вихідний природний текст – це один із обраних Вами в попередніх лабораторних роботах текстів;
   б. текст мавпи Міллера (програма +monkey_texts з лабораторної роботи №11; алфавіт M від 1 до 33; однакові ймовірності всіх букв; ймовірність пробілу 0.2; barrier = 1);
   в. текст Маркова (програма +Markov texts з лабораторної роботи №27; тип ланки – або «букви», або «слова»; довжина ланцюжка N – від 3 до 8);
   г. текст Саймона (програми +Simon_3in1 або +Simon_with memory із лабораторної роботи №11; імовірність нових слів з мішка 0.1–3.0%; решта параметрів за бажанням);
   д. текст Поля (програма +Polya texts з лабораторної роботи №11; початкові розміри урни зі словами від 10 000 до 100 000 слів; цілочисельні параметри ρ і ν обрати так: $\nu = 1, 2, ..., 5, \rho \leq \nu$).

4. Додатково слід згенерувати тексти, які є рандомізованими версіями обраного природного тексту, із використанням рандомізації інших типів. Це рандомізація на лінгвістичних рівнях букв і речень, а також локальна рандомізація на рівні слів. Для цього можна використати програму +shuffler2023 (виконавчий файл у папках зі шляхом Build -> exe.win32-3.6) або програму +shuffler_universal із лабораторної роботи №11.

   Разом за цим пунктом Ви одержите 3 рандомізовані тексти: глобально рандомізований на рівні букв (символів), глобально рандомізований на рівні речень і локально рандомізований на рівні слів.

5. Використовуючи програму +monkey_text_generator(letters) із даної лабораторної роботи, слід додатково згенерувати *по одному* тексту мавпи Міллера із:
   1) однаковими частотами букв

2) лінійною ранговою залежністю частоти букв
3) логарифмічною ранговою залежністю частоти букв.

Разом за цим пунктом Ви одержите 3 рандомні тексти мавпи Міллера.

У режимах 2) і 3) слід коректно обрати чисельні коефіцієнти на інтерфейсі (див. файл recipes_lin&log-monkey_corr.doc, а також кольорову підказку на інтерфейсі)!

6. Перевірити перший та другий закони Ціпфа, а також закони Парето і Гіпса на лінгвістичному рівневі слів для кожного із генерованих текстів типів (б)–(д) так, як це робилося в лабораторних роботах №2 і №4.

Зауваження: перевірка дотримання статистичних законів для рандомізованих текстів (тобто текстів типу (а)) *не потрібна*, бо рандомізація не впливає на частоти символів або слів. Іншими словами, закони Ціпфа, Парето та Гіпса для рандомізованого тексту тотожні до тих же законів для початкового природного тексту, який рандомізували.

7. Перевірити, чи виконуються основні статистичні закони для слів для текстів мавпи Міллера, Саймона, Маркова та Поля.

8. У висновку підсумуйте отримані Вами результати. Зокрема, з'ясуйте, для яких типів текстів виконуються або не виконуються закони Ціпфа, Парето та Гіпса? Для яких типів текстів взаємні зв'язки між показниками різних статистичних законів α, β, k і θ збігаються з теоретично передбаченими, а для яких ні? Чому?