Лабораторна робота №16

«Характеристика повторюваності лінгвістичних елементів у тексті»

Завдання:
використовуючи видану локально розміщену програму ++repetition3.3jar і онлайн-програму (див. файл +Colab program_explanation.pdf), вивчити характеристику повторюваності для кількох природних і рандомних текстів, генерованих в попередніх лабораторних роботах.

Теоретичні відомості:
1. конспект лекцій
2. стаття 1Repetition Characteristic for Single Texts2021.pdf
3. стаття 2Large-Scale Studies of the Repetition Characteristic for Different Models of Symbolic Sequences2021.pdf
4. довідковий файл +Colab program_explanation.pdf.

Порядок виконання роботи та вказівки до оформлення звіту

1. Звіт повинен містити титульну сторінку, текст завдання, коротку теоретичну частину, опис виконання роботи, опис і аналіз результатів і висновки.

2. У теоретичній частині коротко опишіть, що таке характеристика повторюваності, як її розраховують і які її основні практичні застосування (див. додану літературу, зокрема конспект лекцій, а також розділ 1 і підрозділ 2.2 статті 1Repetition Characteristic for Single Texts2021.pdf).

3. Оберіть два природні тексти різними мовами і згенеруйте два рандомні тексти різних типів (найліпше використати тексти, генеровані в попередніх лабораторних роботах №9 і №11).

Зауваження 1. Як і у всіх решті лабораторних роботах, слід чітко зазначити в звіті, які саме тексти було обрано!
Зауваження 2. Якщо Ви раніше виконували лабораторну роботу №11, то вимога обрати для генерування ті типи рандомних текстів, які є відмінними від типів, досліджених у цій лабораторній роботі!

4. Ознайомтеся з програмою, описаною в файлі +Colab program_explanation.pdf. Протестуйте цю програму розрахунку повторюваності на основі ресурсу на colab. Опрацюйте цією програмою обрані Вами тексти і побудувати графіки $v(t)$ залежності параметра повторюваності v від довжини тексту t.

5. Зауважмо, що обсяги даних $v(t)$, збережених за результатами роботи програми, можуть бути гігантськими, оскільки кожному символові у тексті відповідає його номер і число з плаваючою комою. Тому тут цілком виправданий неповний, скорочений запис даних $v(t)$, при якому деякі з точок періодично пропускають. Для такого запису можна використати програму +csv-minifier_for Colab program, яка мініфікує експериментальні дані. Отже, для обмеження кількості точок та наступної побудови графіків $v(t)$ засобами зовнішніх програм слід використати програму +csv-minifier_for Colab program. Пропонуємо використати цю програму та записувати лише мініфіковані файли типу csv або xls.

6. Порівняйте залежності $v(t)$ для різних текстів і визначте, для яких із них ці залежності збігаються до деякого конкретного значення v_0 при великих значеннях t.

7. Ознайомтеся з програмою +repetition3.3jar. Користуючись цією програмою, найперше модифікуйте виконавчий файл Executable_2023_bigXXXGB.bat, де позначення «XXX» відповідає оптимальному обсягу оперативної пам'яті, яку Ви виділяєте на програму. Це дасть змогу оптимізувати роботу програми. В головному меню програми оберіть досліджуваний текстовий файл, в пункті меню «Analyze» оберіть режим роботи «A set of characters». Далі оберіть режим «plus 1 and first repetition», не прописуючи параметрів біжучого вікна справа, натисніть «Analyze» – і одержите графік v(t).

   Для докладнішого аналізу цих даних знову перейдіть до пунктів меню File -> Save result -> v(t) result. Верхня частина спливаючого меню визначає параметри скороченого (мініфікованого) збереження даних: тут поле «Save first *n rows» визначає, скільки перших точок Ви збережете до файлу, а поле «save each *m rows» – через яку кількість точок Ви будете зберігати точки (наприклад, одну точку через 10 точок). Скажімо, якщо обрано n = 5, m = 20, то Ви збережете такі точки: 1, 2, 3, 4, 5, 25, 45, ...

   Нижня частина спливаючого меню визначає особливості розрахунків середнього значення v_0 і с.к.в. Δv параметра повторюваності, які набувають особливого значення у разі, якщо функція v(t) прямує до визначеної величини зі зростанням позиції в тексті t. Величини v(t) усереднюють, починаючи з деякого значення «Start from value t», а розрахунки йдуть із деяким кроком «Step between values t». Після введення цих величин і натискання кнопки «Calculate» одержуємо параметри «Average» v_0 і «Standard deviation» Δv. Параметри v_0 і Δv підсумовують поведінку параметра повторюваності.

8. Дослідіть за допомогою програми +repetition3.3jar ті самі тексти, що й у пункті 4. Оберіть основний режим програми «A set of characters» і режим розрахунку повторень «plus 1 and first repetition».

   Зробіть висновки про поведінку параметра повторюваності зі зростанням розмірів текстів для різних типів текстів.

9. Додатково побудуйте залежності $v(t)$ для кількох інших режимів роботи програми (режим «A set of words» для аналізу повторень слів у довільній комбінації з режимами «врахування лише перших повторень або всіх повторень» і «1 бал або N балів за повторення»).

10. Порівняйте характеристики повторюваності $v(t)$ для різних типів текстів в обраних Вами режимах.

    Зробіть додаткові висновки щодо поведінки параметра повторюваності зі зростанням розмірів текстів для різних типів текстів і обраних Вами режимів роботи програми.

11. Висновки повинні містити короткий аналіз особливостей Ваших даних і графіків і схожості або відмінності поведінки характеристики повторюваності для природних і рандомних текстів.