

Лабораторна робота №04

«Закон Гіпса для слів в окремих текстах»

Завдання

використовуючи програму +proj6stats&plots, на лінгвістичному рівневі слів дослідити виконання закону Гіпса для обраних Вами текстів у лабораторній роботі №1. Провести дослідження для однограм ($n = 1$) за різних опцій (умов).

Теоретичні відомості

1. конспект лекцій;
2. стаття Zipf's and Heaps' laws for the natural and some related random texts.pdf;
3. стаття Statistical regularities of the linguistics of computer programs.pdf,

Порядок виконання роботи та вказівки до оформлення звіту

1. Звіт має містити титульну сторінку, текст завдання, коротку теоретичну частину, опис виконання завдання, висвітлення та аналіз одержаних результатів і висновок.
2. У теоретичній частині сформулювати закон Гіпса і співвідношення між показниками степеневих законів лінгвістичної статистики α , β , k , θ .
3. Ознайомитися із вкладкою «Гіпс» програми +proj6stats&plots.
4. Провести дослідження для уніграм ($n = 1$) за різних умов:
 - фіксована позиція біжучого вікна;
 - фіксований розмір біжучого вікна (режим вивчення параметра TTR – type-to-token ratio);
 - змінні розмір вікна і позиція біжучого вікна.
1. Змінюючи початкові дані програми (розміри вікна, крок його переміщення та ін.), отримати за кожної з умов по кілька варіантів даних для аналізу. Із отриманих даних обрати три найбільш інформативні, на Вашу думку, варіанти наборів параметрів біжучого вікна.
2. Для режимів «фіксована позиція біжучого вікна» і «змінні розмір вікна і позиція біжучого вікна» побудувати графіки залежності розмірів словника V від розмірів тексту L в звичайному і подвійному логарифмічному масштабах.
3. Використовуючи лінійну апроксимацію даних в подвійному логарифмічному масштабі, знайти числові значення коефіцієнту нахилу θ . Порівняти відповідні коефіцієнти кореляції для обох режимів. У якому режимі коефіцієнт кореляції найвищий, тобто графік $\log V(\log L)$ найближчий до лінійного?
4. У звіті вказати значення основних кількісних параметрів графіків, а також загальну статистичну інформацію про апроксимацію.
5. Перевірити виконання співвідношень між показниками степеня α , β , k з лабораторної роботи №1 і одержаним зараз показником степеня θ для даного тексту.
6. Перевірити виконання закону Гіпса для кількох інших текстів тою ж мовою і текстів з мов іншим родин. Порівняти значення коефіцієнтів θ для цих текстів і мов. Які з цих мов можна вважати синтетичними, а які аналітичними, виходячи з Ваших даних?
7. Висновок повинен містити аналіз отриманих результатів.