

Міністерство освіти й науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
з предмета: Комп'ютерна лінгвістика

Звіт
про виконання лабораторної роботи № 8а
**«Закони статистичної лінгвістики на лінгвістичних рівнях букв (символів) і
буквених (символьних) n-грам для окремих текстів»**

Виконав:
Студент групи
Фес-32с
Бойко Кирило

Львів 2024

Завдання

Використовуючи програму +proj6stats&plots, дослідити закони статистичної лінгвістики (див. лабораторні роботи №2 і №4) на лінгвістичних рівнях букв (символів). Розглянути статистичні закони для буквених і символьних n-грам для окремих випадків $n = 1-4$ для деякого тексту. Побудувати спільну статистику для цих n-грам

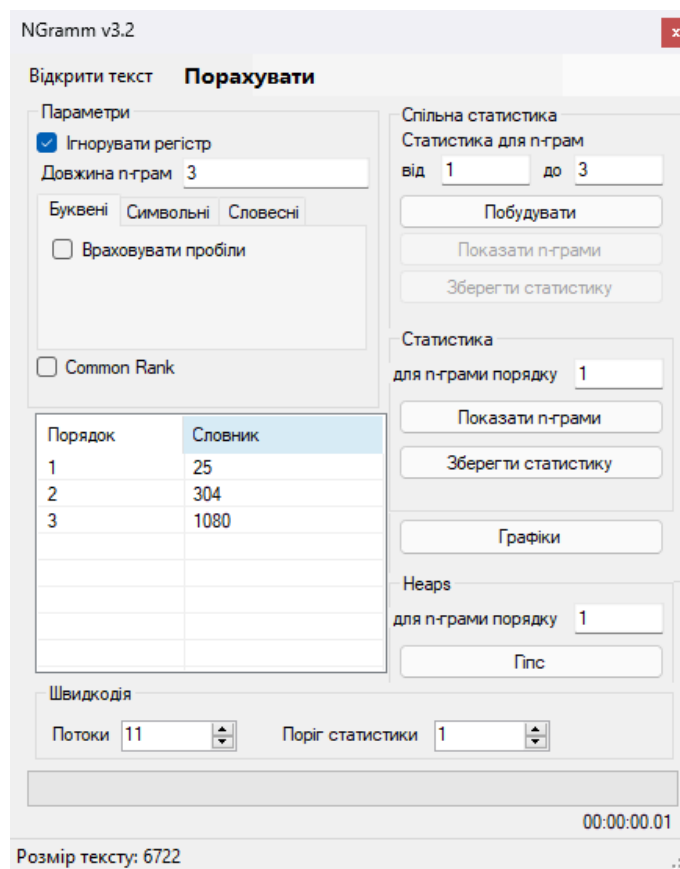
Хід виконання роботи

Для виконання цієї лабораторної роботи я вибрав текст, який був обраний ще в першій лабораторній роботі

DONALD J. TRUMP January 20, 2017

Я запустив програму **+proj6stats&plots**, завантажив текстовий файл та обрав довжину n-грам, встановивши її на значення 3. Це дозволить аналізувати послідовності з трьох слів, що допомагає детальніше вивчити частотний розподіл та структуру тексту на рівні трьохграм.

Результати



NGramm v3.2

Відкрити текст **Порахувати**

Параметри

- ☒ Ігнорувати регістр
- Довжина n-грам: 3
- Буквені Символьні Словесні
- ☐ Враховувати пробіли
- ☐ Common Rank

| Порядок | Словник |
|---------|---------|
| 1 | 25 |
| 2 | 304 |
| 3 | 1080 |
| | |
| | |
| | |
| | |

Швидкодія

Потоки: 11

Поріг статистики: 1

00:00:00.01

Розмір тексту: 6722

Спільна статистика

Статистика для n-грам

від 1 до 3

Побудувати

Показати n-грами

Зберегти статистику

Статистика

для n-грам порядку 1

Показати n-грами

Зберегти статистику

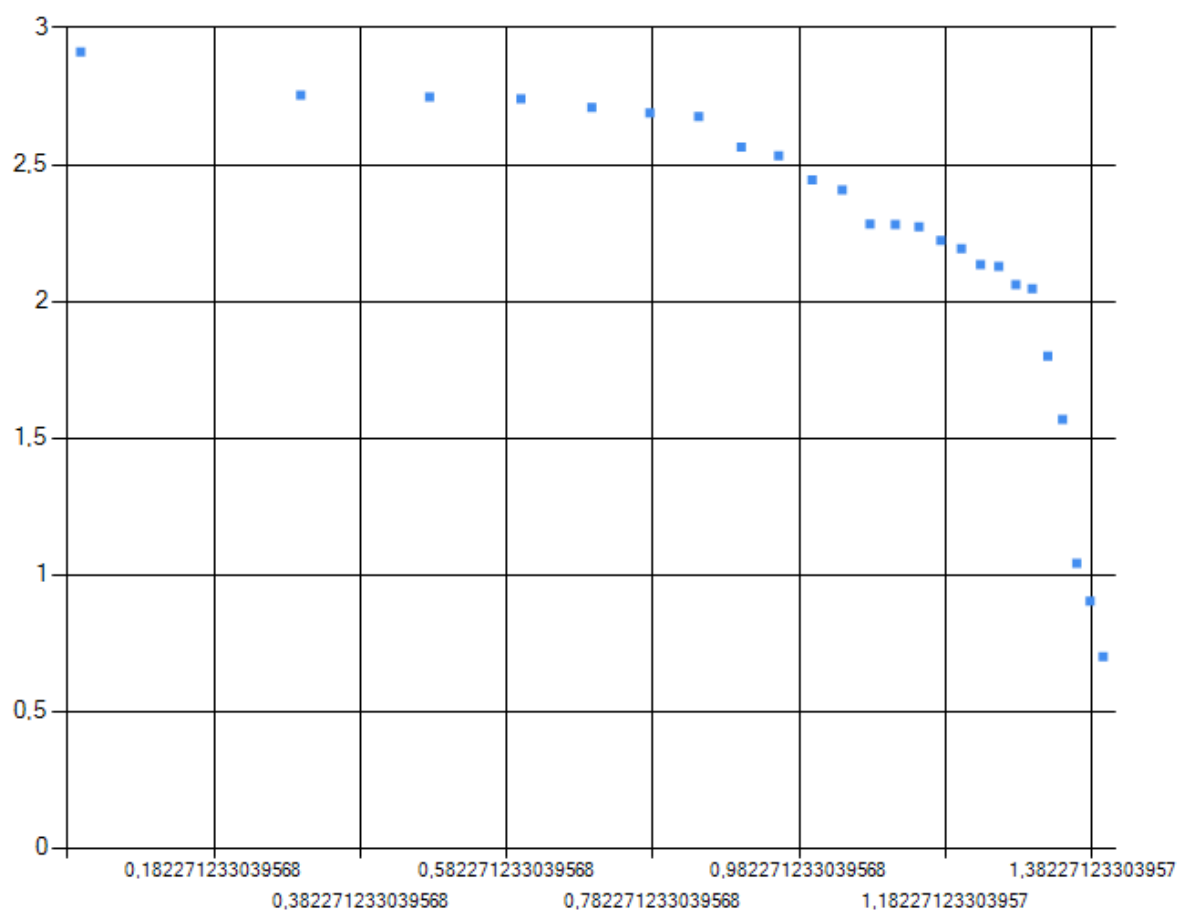
Графіки

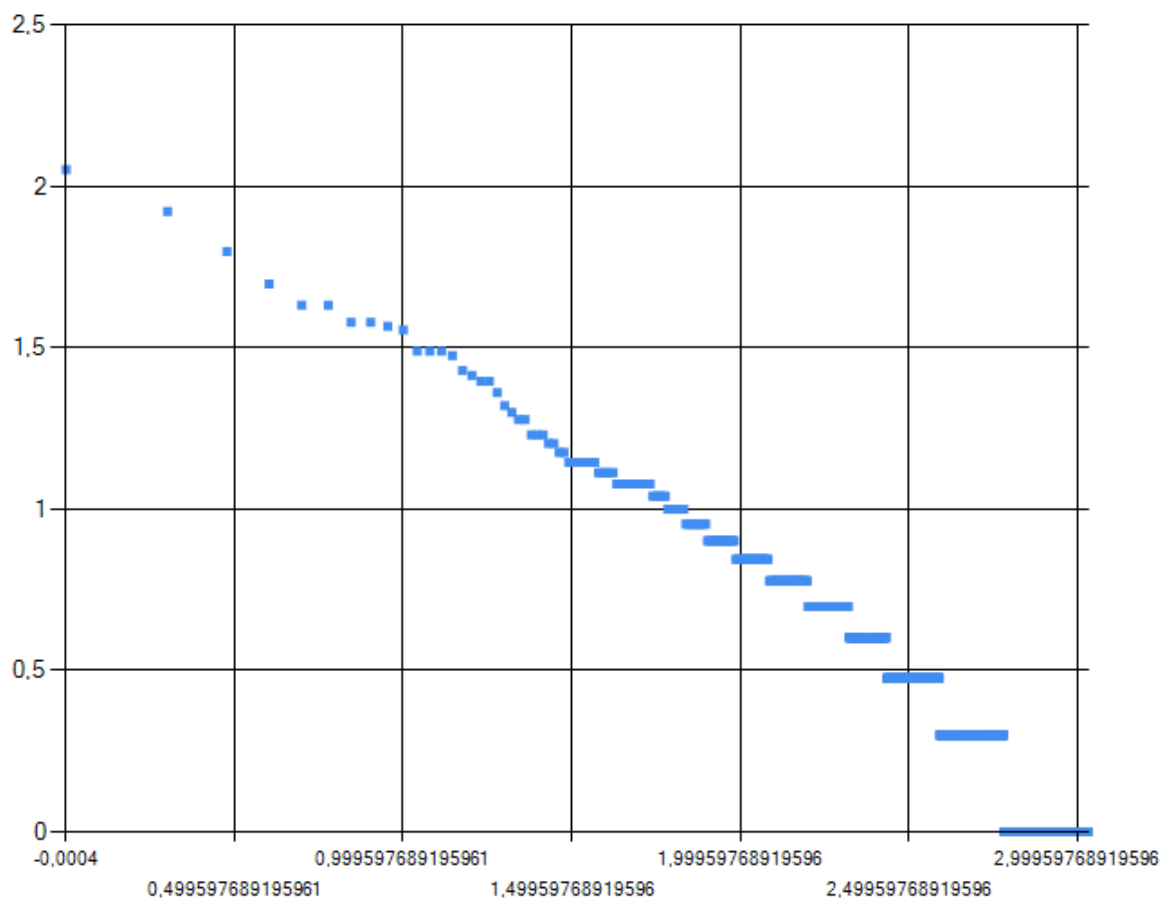
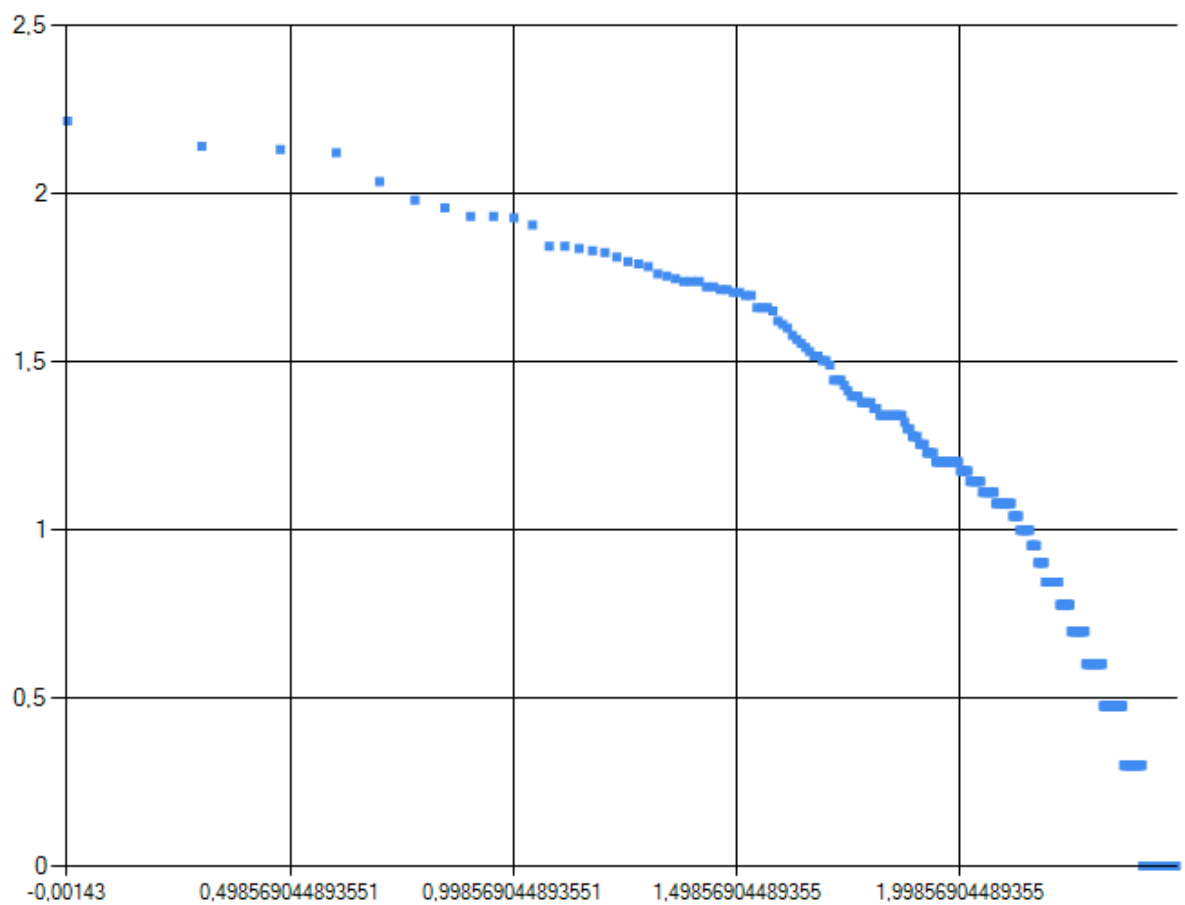
Неарс

для n-грам порядку 1

Гіпс

| Ранг | N-грама | Кількість | Ранг | N-грама | Кількі... | Ранг | N-г... | Кількість |
|------|---------|-----------|------|---------|-----------|------|--------|-----------|
| 1 | e | 817 | 1 | th | 165 | 1 | the | 113 |
| 2 | o | 567 | 2 | an | 139 | 2 | and | 84 |
| 3 | t | 560 | 3 | er | 136 | 3 | our | 63 |
| 4 | a | 550 | 4 | he | 133 | 4 | ill | 50 |
| 5 | r | 511 | 5 | ou | 109 | 5 | wil | 43 |
| 6 | i | 488 | 6 | nd | 96 | 6 | ame | 43 |
| 7 | n | 474 | 7 | re | 91 | 7 | eri | 38 |
| 8 | s | 367 | 8 | ri | 86 | 8 | ric | 38 |
| 9 | l | 341 | 9 | in | 86 | 9 | ica | 37 |
| 10 | h | 278 | 10 | ll | 85 | 10 | mer | 36 |
| 11 | d | 256 | 11 | on | 81 | 11 | ver | 31 |
| 12 | u | 192 | 12 | to | 70 | 12 | ing | 31 |
| 13 | w | 191 | 13 | il | 70 | 13 | ion | 31 |
| 14 | c | 187 | 14 | or | 69 | 14 | for | 30 |
| 15 | m | 167 | 15 | me | 68 | 15 | all | 27 |
| 16 | f | 156 | 16 | ur | 67 | 16 | you | 26 |
| 17 | g | 136 | 17 | en | 65 | 17 | eve | 25 |
| 18 | y | 134 | 18 | we | 63 | 18 | her | 25 |
| 19 | b | 115 | 19 | es | 62 | 19 | tio | 23 |
| 20 | p | 111 | 20 | at | 61 | 20 | ent | 21 |
| 21 | v | 63 | 21 | ic | 58 | 21 | rea | 20 |
| 22 | k | 37 | 22 | am | 57 | 22 | ies | 19 |
| 23 | j | 11 | 23 | wi | 56 | 23 | one | 19 |
| 24 | z | 8 | 24 | st | 55 | 24 | ere | 17 |
| 25 | x | 5 | 25 | ca | 55 | 25 | ati | 17 |
| | | | 26 | al | 55 | 26 | can | 17 |
| | | | 27 | le | 53 | | | |





Висновок: У процесі виконання лабораторної роботи я дослідив статистичні закономірності для n -грам з довжинами $n=1, 2$ та 3 . Було проведено аналіз рангових залежностей для кожного значення n , створено графіки в напівлогарифмічному та подвійному логарифмічному масштабах, а також виконано лінійну апроксимацію цих залежностей.