Міністерство освіти і науки України
Львівський національний університет імені Івана Франка

Звіт про виконання лабораторної роботи №8

Закони статистичної лінгвістики на лінгвістичних рівнях букв (символів) і буквених (символьних) n-грам для окремих текстів

Виконав:
студент групи ФеІ-43
Горак Т. А.
Перевірив:
доц. Катеринчук І. М.

Львів-2024

Завдання

1. Використовуючи програму +proj6stats&plots, дослідити закони статистичної лінгвістики на лінгвістичних рівнях букв (символів).
2. Розглянути статистичні закони для буквених і символьних n-грам для окремих випадків n = 1–4 для деякого тексту.
3. Побудувати спільну статистику для цих n-грам.

Теоретичні матеріали

Рангові залежності, частотні розподіли і закон зростання словника для буквених і символьних n-грам є важливими концепціями в аналізі тексту та обробці природної мови. Основна ідея полягає в тому, що частота вживання символів чи n-грам у тексті може бути описана певним математичним законом, який вказує на те, як часто певні символи або n-грами зустрічаються у тексті.

Для букв: Рангова залежність для букв наближено описується логарифмічною функцією, що означає, що частота вживання букви залежить від її позиції в рейтингу по частоті. Тобто, чим менше ранг букви, тим вона вживається частіше.

Для n-грам: З ростом n, рангова залежність для n-грам поступово змінюється від логарифмічної до степеневої функції. Це означає, що збільшенням довжини n-грами її частота вживання може рости не пропорційно до логарифму рангу, а, можливо, за іншою, більш складною функцією.

Частотні розподіли:

Закон Парето: Для букв та n-грам може застосовуватися закон Парето, який є експоненціальною функцією. Цей закон вказує на те, що є деякі символи або n-грами, які вживаються значно частіше за інші, що є характерною рисою багатьох мов.

Закон зростання словника: Для букв та n-грам закон зростання словника може бути різним. З ростом n словник n-грам росте експоненційно, але з певної точки може зупинитися через обмеженість унікальних n-грам у текстах.

Ці концепції допомагають у розумінні та моделюванні вживання мови, а також у побудові ефективних методів для стиснення, кодування та генерації тексту.

Хід роботи
A tale of two cities
N = 1

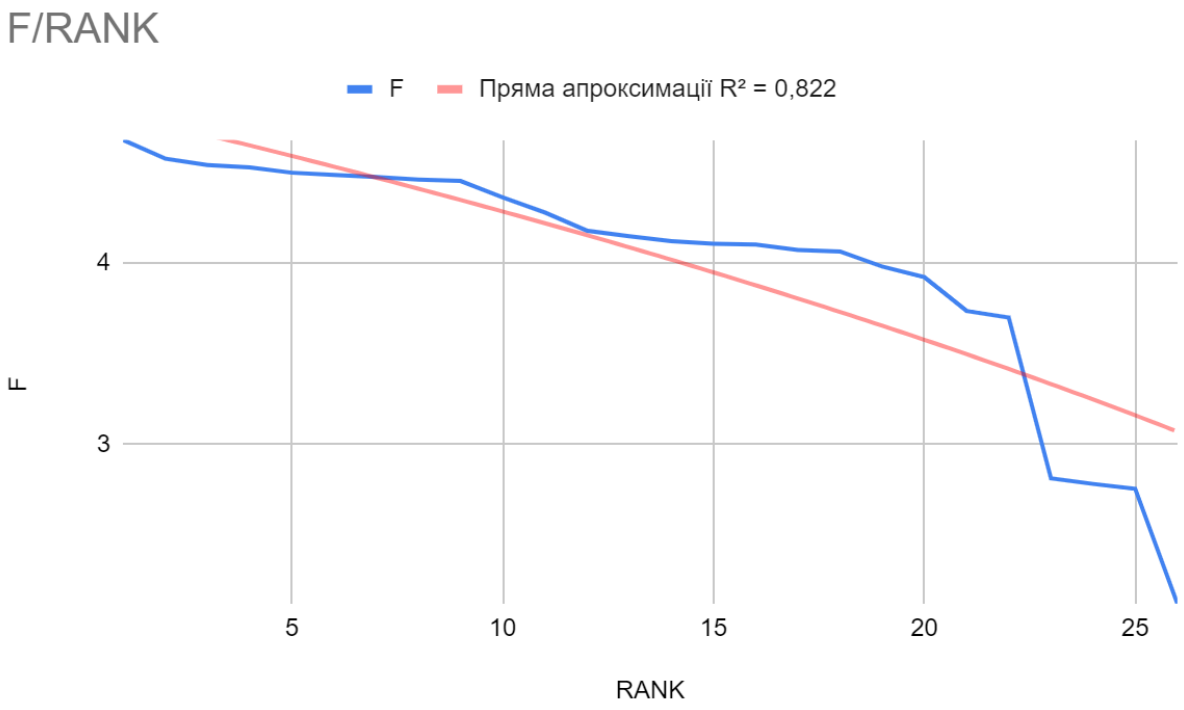

Рис. 1. Напівлогарифмічний масштаб

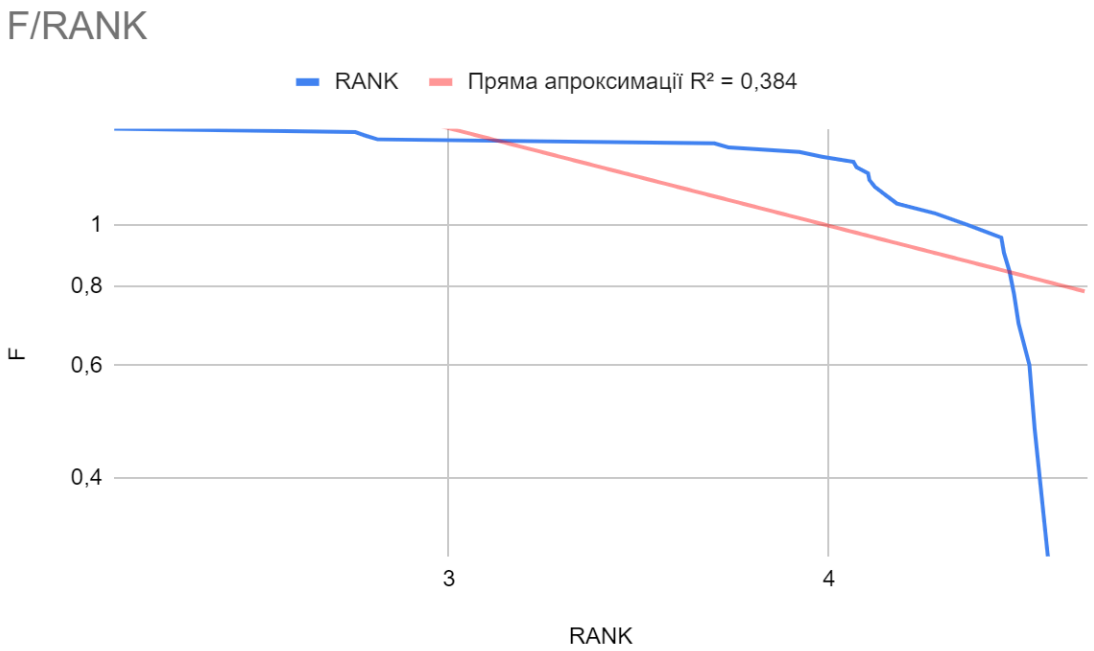

Рис. 2. Подвійний логарифмічний масштаб

N=2

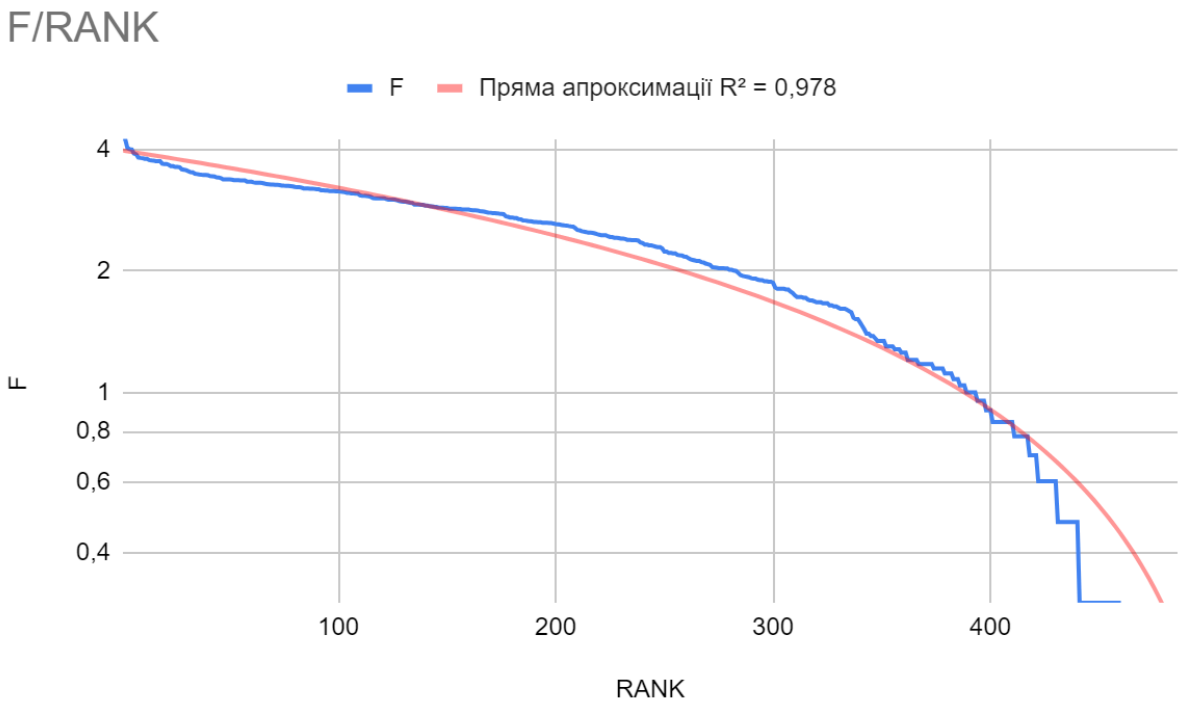

Рис. 3. Напівлогарифмічний масштаб

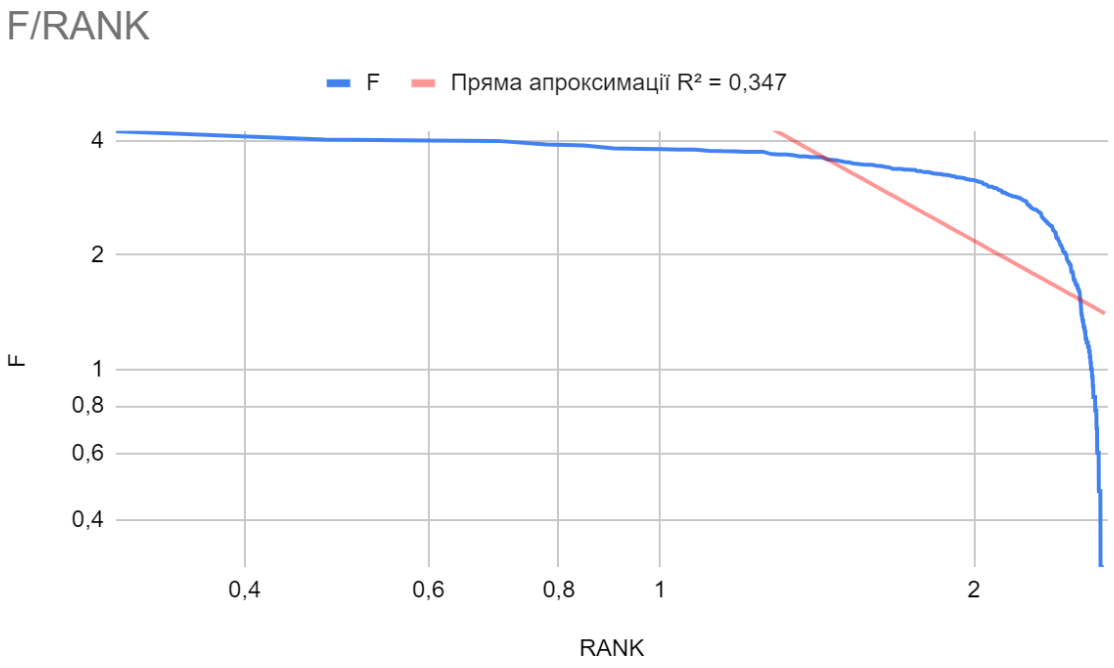

Рис. 4. Подвійний логарифмічний масштаб

N=3

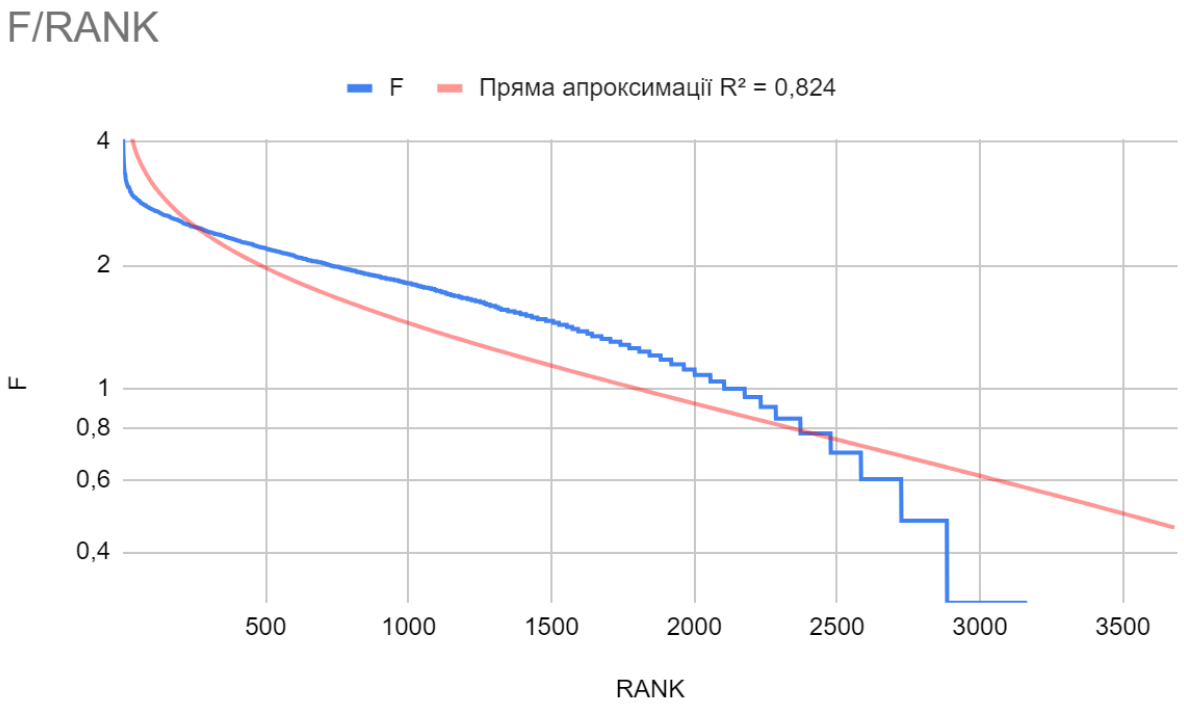

Рис. 5. Напівлогарифмічний масштаб

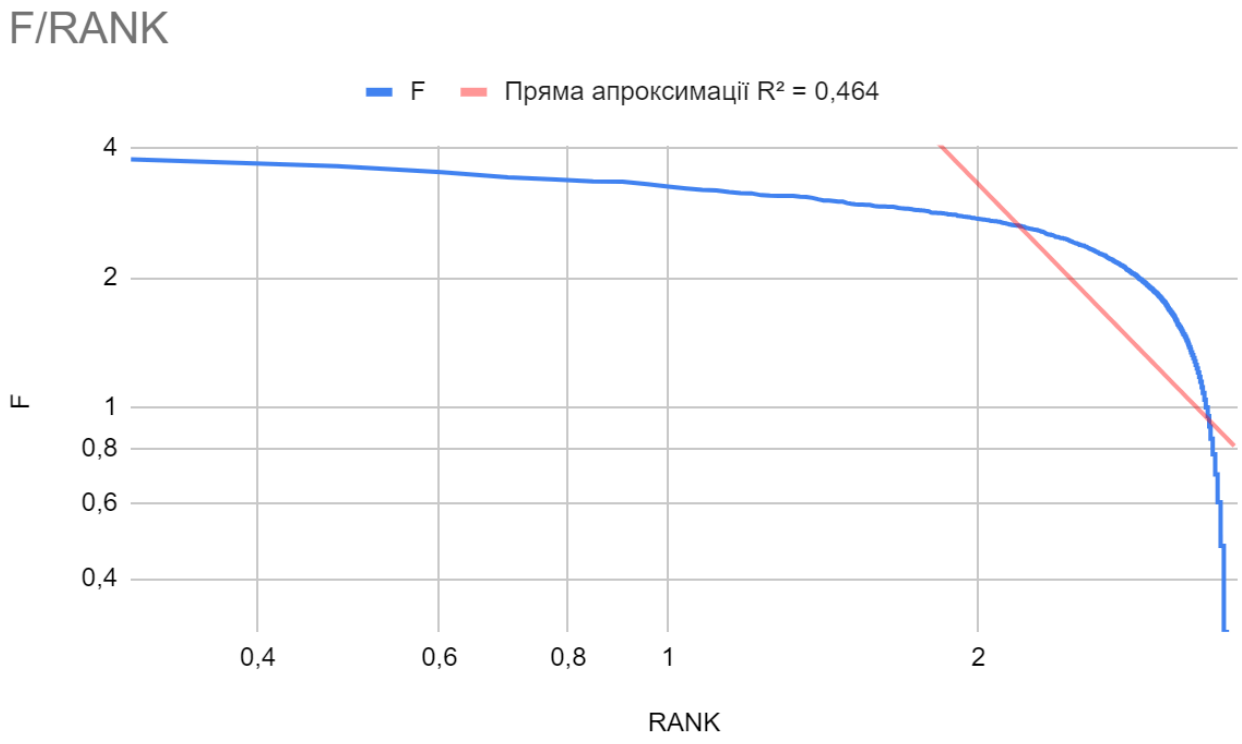

Рис. 6. Подвійний логарифмічний масштаб

N=4

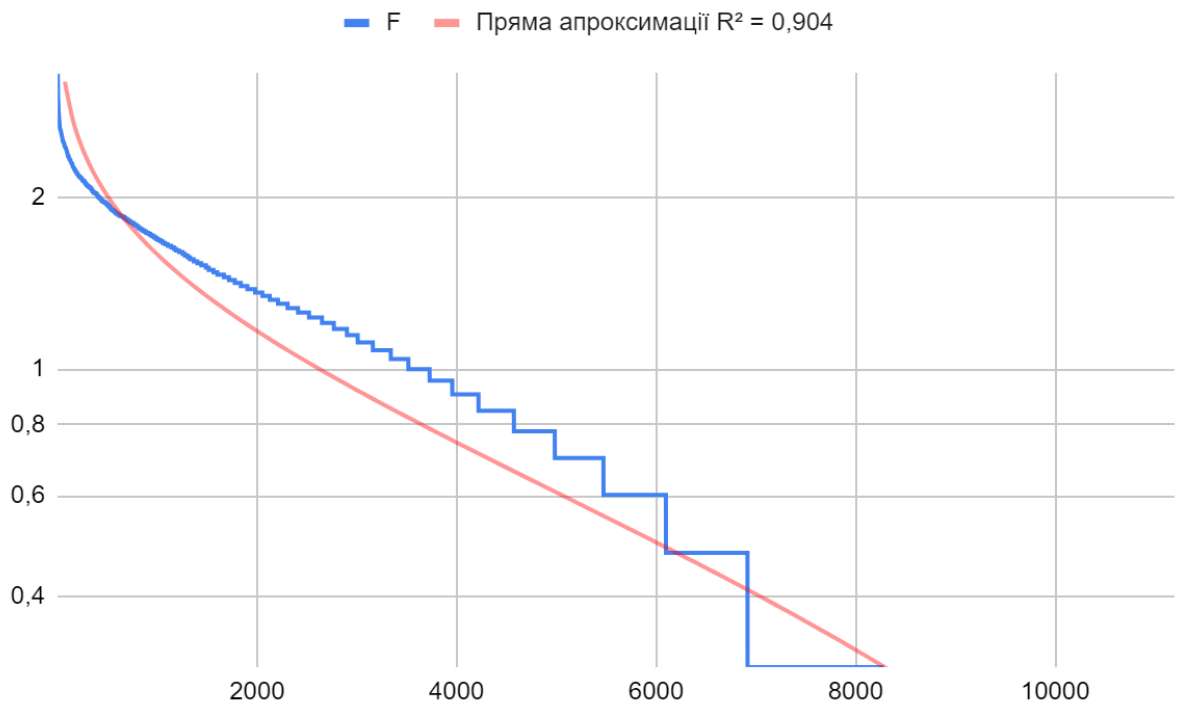

Рис. 7. Напівлогарифмічний масштаб

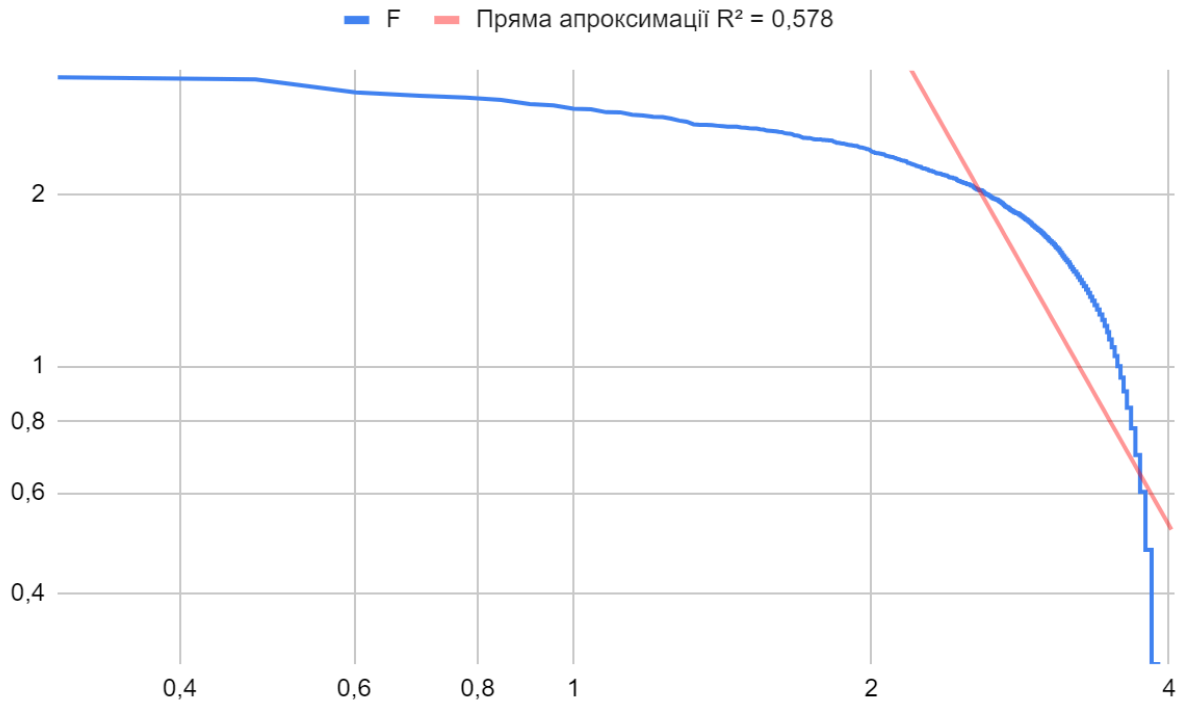

Рис. 8. Подвійний логарифмічний масштаб

Висновки: У цій роботі було проведено аналіз статистичних закономірностей лінгвістичних рівнів букв (символів), використовуючи програму +proj6stats&plots. Завданням дослідження було вивчення характеристик і розподілів n-грам для n від 1 до 4. Було вибрано певний текст, на основі якого було проведено аналіз частоти виникнення одиничних букв, біграм, триграм та тетраграм. Встановлено, що розподіл частот n-грам має тенденцію до зменшення із збільшенням n. Тобто, чим більшою є кількість букв у n-грамі, тим рідше така комбінація зустрічається у тексті. Це спостереження узгоджується з принципом меншої ймовірності появи довших послідовностей букв. Також було виявлено, що деякі n-грами з'являються у тексті значно частіше, ніж інші, що відображає особливості використання мови та її структуру. Зокрема, для одиничних букв було помічено, що деякі літери використовуються набагато частіше, ніж інші, що відповідає закону Ципфа.