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 № 2
«Закони Ціпфа та Парето для слів у текстах»

Виконав:
Студент групи
Фес-32с
Бойко Кирило

Львів 2024

Завдання

1. Використовуючи програму +proj6stats&plots, дослідити I і II-й закони Ціпфа та закон Парето для одного із обраних Вами текстів англійською мовою на рівні слів. Побудувати графіки F(r), p(F) і P(F). Використовуючи лінійну апроксимацію даних знайти коефіцієнти статистичних законів α, β, k.
2. Дослідити I, II-й закони Ціпфа та закон Парето для одного з запропонованих текстів українською (або будь-якою іншою) мовою на рівні слів та провести порівняння отриманих результатів із результатами дослідження тексту англійською мовою.

Хід виконання роботи

Для виконання цієї лабораторної роботи я використав два текстові файли, один з яких було вибрано на попередньому занятті.

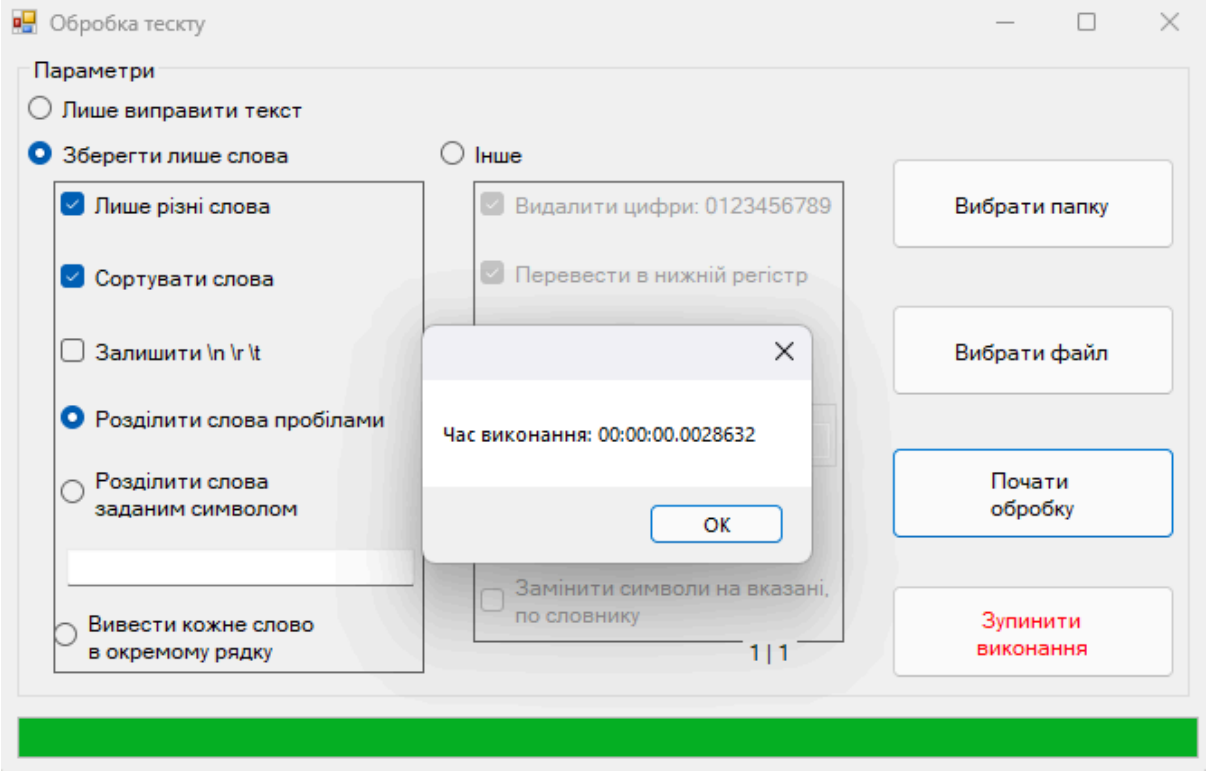
DONALD J. TRUMP January 20, 2017

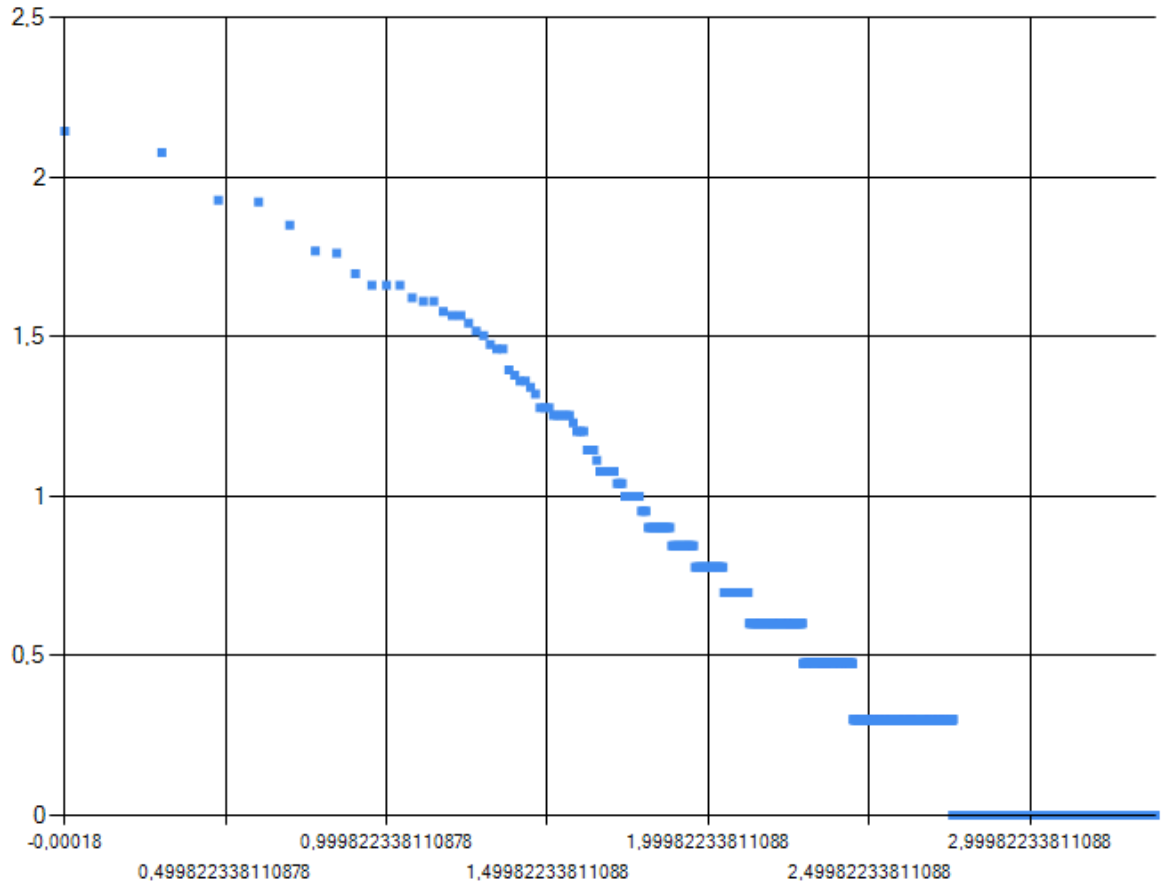
Людина-маятник

Для дослідження I та II законів Ціпфа, а також закону Парето для тексту українською мовою на рівні слів, необхідно виконати кілька етапів:

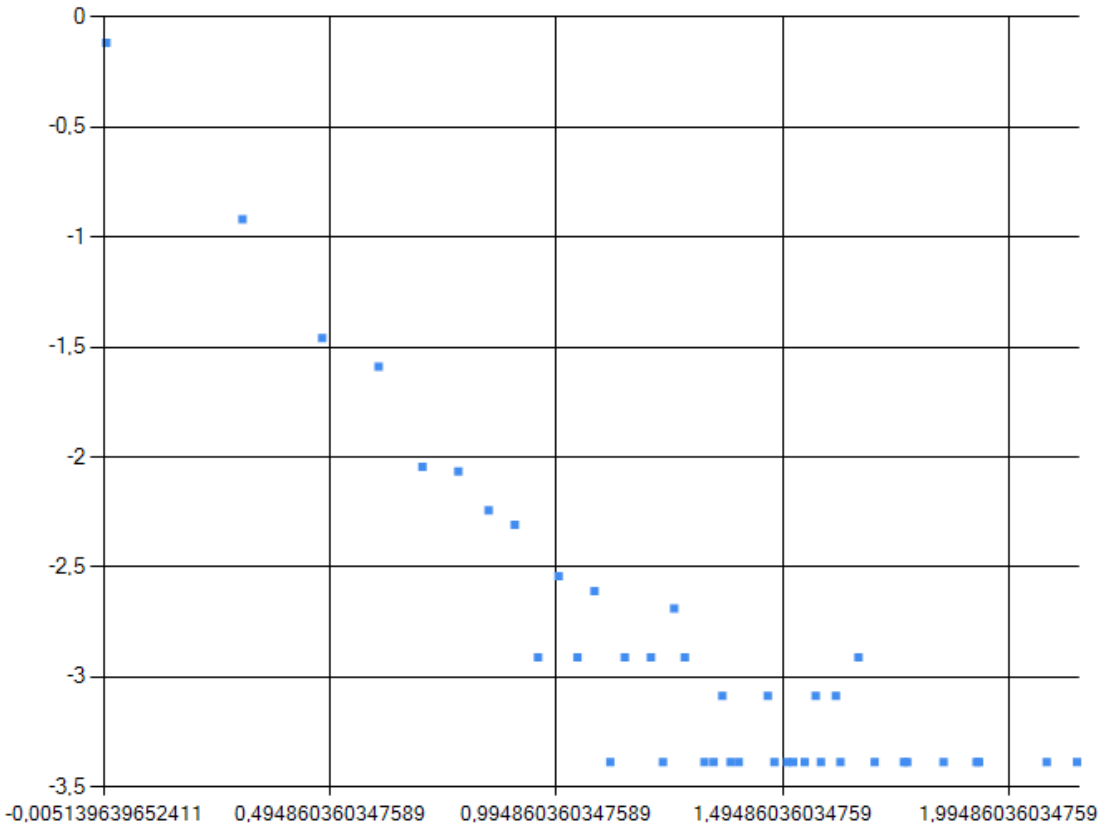
- **Підготовка тексту**: Спочатку я завантажу текстовий файл з українським текстом і проведу його обробку, видаливши спеціальні символи та залишивши лише слова.
- **Побудова частотного розподілу**: Після завантаження та обробки файлу я отримаю результати аналізу частотного розподілу слів.

Результати

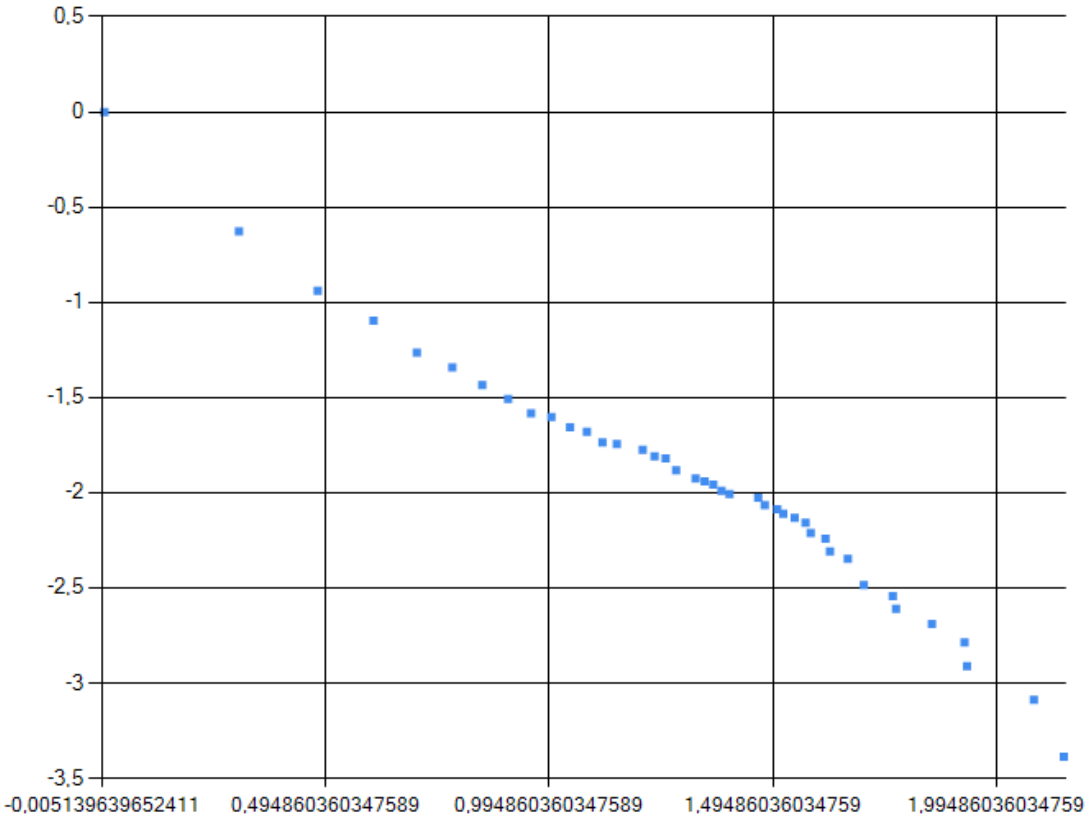




1-й закон Зіпфа



2-й Закон Зіпфа



Парето

Після проведення дослідження я зробив висновок, що:

І закон Ціпфа:

- Перший графік демонструє зниження частоти слів із збільшенням їх рангу. За І законом Ціпфа, частота слова повинна обернено пропорційно зменшуватися залежно від його рангу. Спад на графіку відповідає цьому правилу: найпоширеніші слова мають високий ранг, і їх частота поступово знижується з підвищенням рангу. Це підтверджує виконання І закону Ціпфа.

ІІ закон Ціпфа:

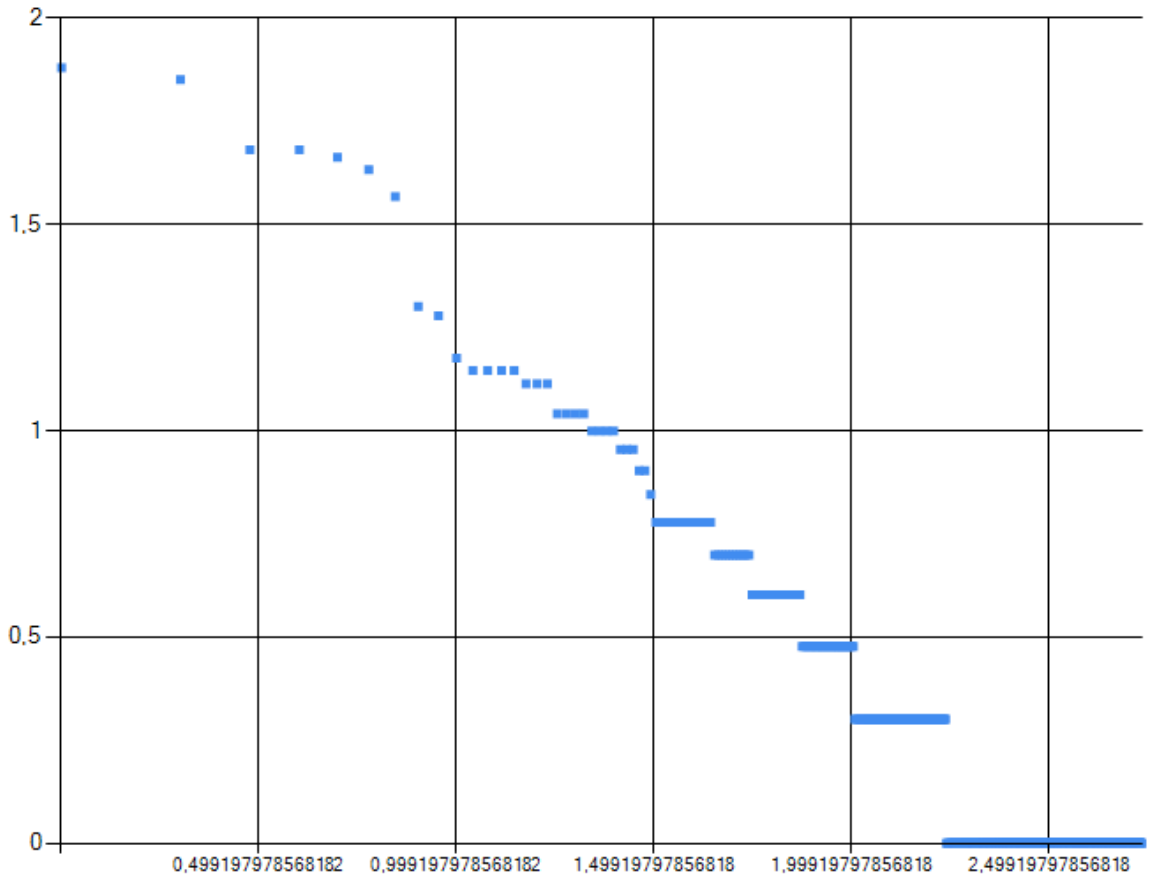
- На другому графіку зображена залежність логарифму частоти від логарифму рангу. Лінійний спад підтверджує виконання ІІ закону Ціпфа, згідно з яким між логарифмами частоти та рангу існує лінійна залежність. Отже, дані відповідають ІІ закону Ціпфа.

Закон Парето:

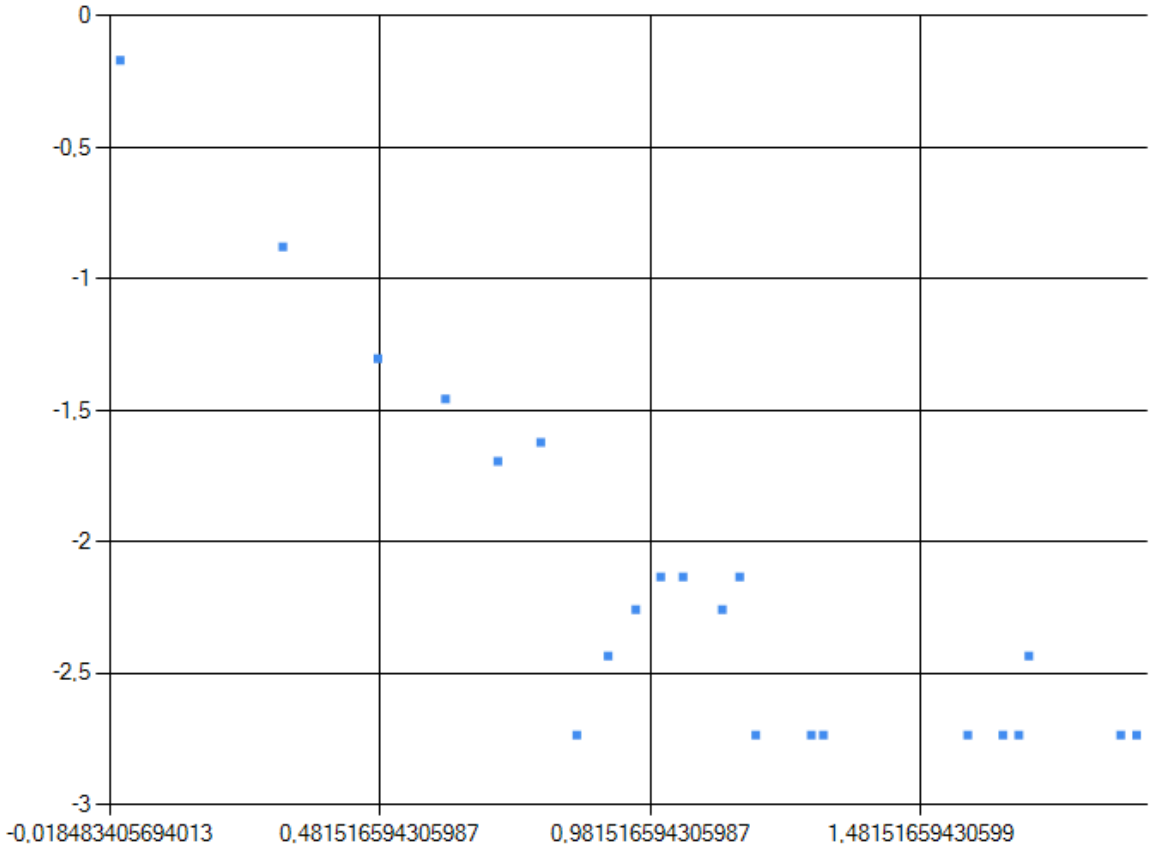
- Хоча графіки безпосередньо не демонструють виконання закону Парето (принципу 80/20), підтвердження І та ІІ законів Ціпфа є непрямим підтвердженням цього принципу. За законом Парето, невелика частина високочастотних слів (приблизно 20%) повинна складати основну частину всього тексту (приблизно 80%). Оскільки найбільш частотні слова домінують у загальній частоті, можна зробити висновок, що закон Парето також ймовірно реалізується у цьому тексті.

Після цього я завантажив текстовий файл англійською мовою для порівняння з текстом українською. Це дозволило порівняти виконання І та ІІ законів Ціпфа, а також закон Парето для текстів на різних мовах і виявити можливі відмінності в розподілі частот слів.

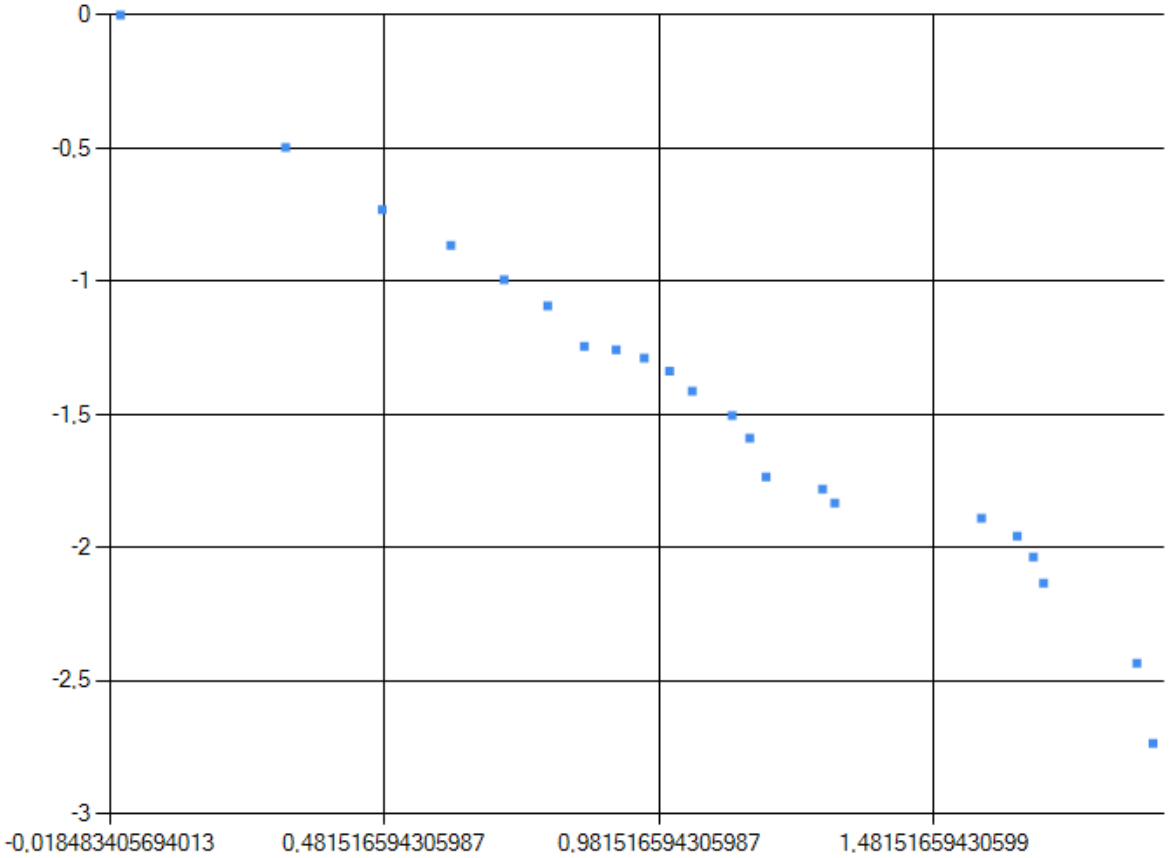
Результати



1-й закон Зіпфа



2-й закон Зіпфа



Парето

Порівняльний аналіз

Результати дослідження підтвердили виконання I та II законів Ціпфа для текстів як українською, так і англійською мовами.

У випадку **I закону Ціпфа** частота слів дійсно зменшується зі збільшенням їх рангу для обох мов.

За **II законом Ціпфа** графіки логарифмів частот і рангів продемонстрували лінійний спад, що також підтверджує виконання цього закону.

Згідно з **законом Парето**, наявність невеликої кількості високочастотних слів, що складають більшу частину тексту, також спостерігалася у обох випадках, що підтверджує виконання цього закону для текстів на двох мовах.

Висновок: У цій лабораторній роботі було досліджено застосування I та II законів Ціпфа, а також закону Парето для текстів українською та англійською мовами. Для кожного тексту було побудовано графіки залежності частоти слів від їх рангу (I закон Ціпфа) та логарифму частоти від логарифму рангу (II закон Ціпфа). Результати показали лінійний спад на графіках, що підтверджує обернено пропорційну залежність частоти слів від їх рангу, відповідно до I та II законів Ціпфа. Крім того, аналіз підтвердив виконання закону Парето для обох текстів, що свідчить про універсальні закономірності в розподілі частоти слів у текстах різних мов.