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 № 5
«Зростання словника для корпусу текстів. Практичні рецепти бінування»

Виконав:
Студент групи
Фес-32с
Бойко Кирило

Львів 2024

Завдання

1. Дослідити закон Гіпса для великого корпусу текстів. Із баз текстів, що додаються, для роботи потрібно вибрати щонайменше 150 (а ще ліпше ~ 1000 і більше) текстів, які мають помітно різні розміри. У дослідженні Ви повинні розглянути три методи бінування: (1) з однаковими довжинами бінів; (2) з різними довжинами бінів, кожен з яких містить однакову кількість емпіричних точок (довжин текстів) Li; (3) експоненційне (або логарифмічне) бінування.
2. Побудувати біновані графіки залежності розмірів словників V від розмірів текстів L. Використовуючи лінійну апроксимацію даних, знайти числове значення коефіцієнту θ, яке притаманне тій чи іншій обраній мові.
3. Порівняти дані для параметра Гіпса θ, отримані за допомогою різних способів бінування.

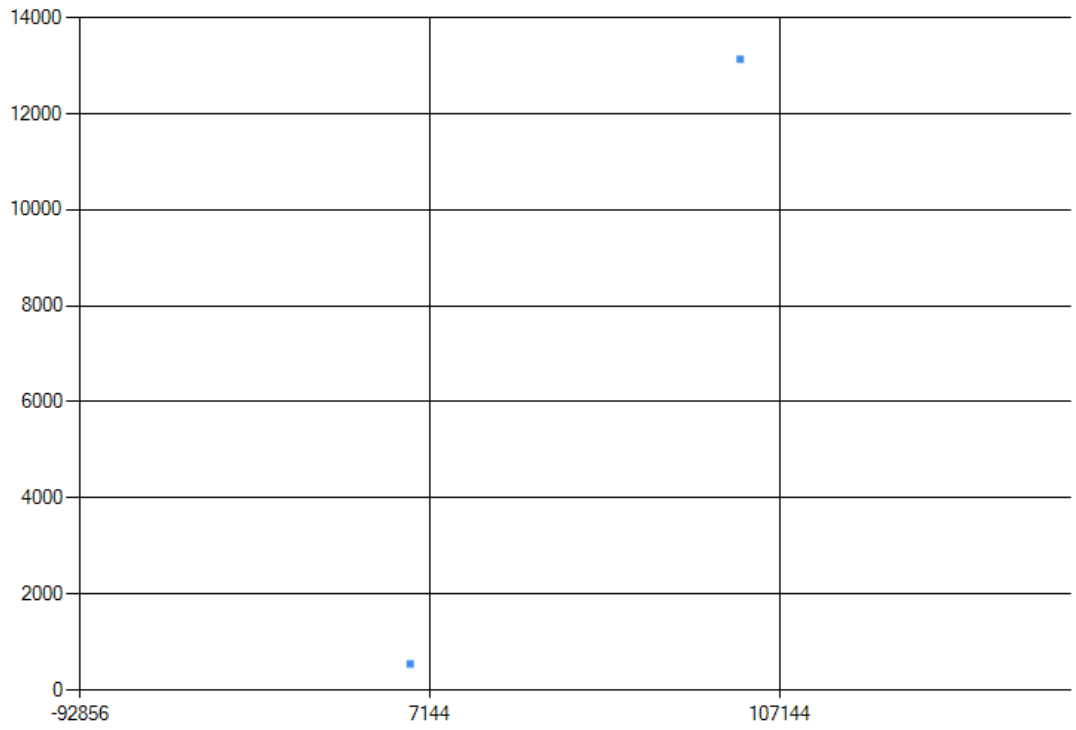
Хід виконання роботи

Для виконання цієї лабораторної роботи я обрав програму **++V(L)binning2023_2** та завантажив текстові файли

Clemencia Novela de costumbres by Fernán Caballero

DONALD J. TRUMP January 20, 2017

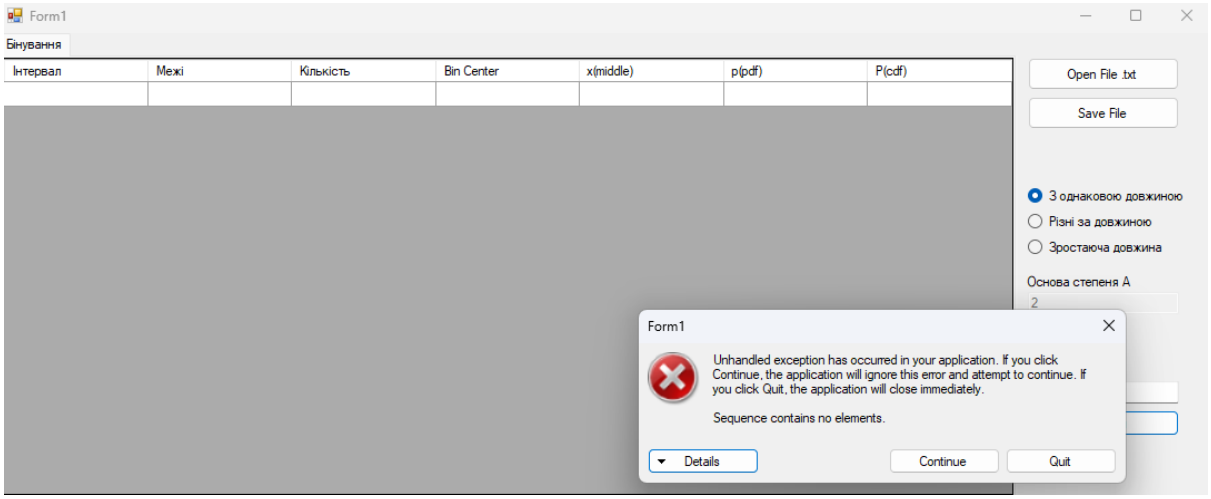
Після запуску програми та завантаження текстів, я отримав такі результати:

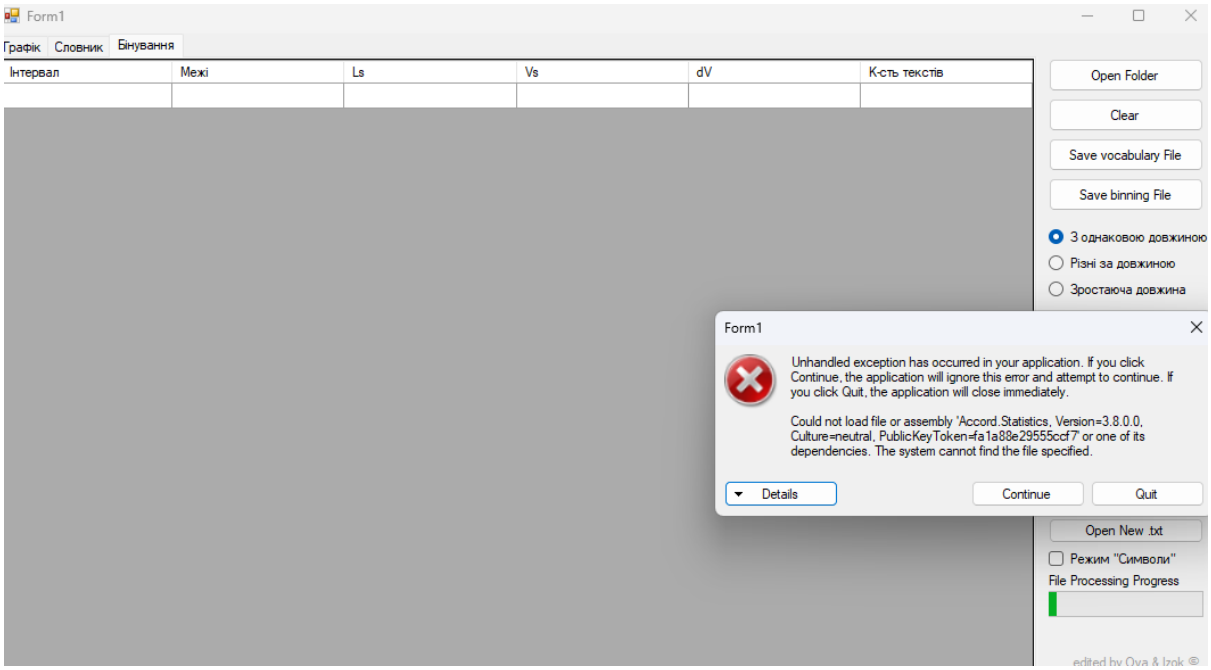


#	Назва	Кількість слів	Кількість різних слів
1	Clemencia Novela de costumbres by Fernán Caball...	95766	13138
2	DONALD J. TRUMP January 20, 2017	1455	542

Після запуску програми **+V(L)binning_DynamicBinning&x_ave_pdf&cdf(ed Klymchuk)** і завантаження відповідних текстів, я отримав такі результати

Інтервал	Межі	Ls	Vs	dV	К-сть текстів
1	1421 - 5749,65	1421	556	0	1
2	5749,65 - 10078,3	0	0	0	0
3	10078,3 - 14406,95	0	0	0	0
4	14406,95 - 18735,6	0	0	0	0
5	18735,6 - 23064,25	0	0	0	0
6	23064,25 - 27392,9	0	0	0	0
7	27392,9 - 31721,55	0	0	0	0
8	31721,55 - 36050,2	0	0	0	0
9	36050,2 - 40378,85	0	0	0	0
10	40378,85 - 44707,5	0	0	0	0
11	44707,5 - 49036,15	0	0	0	0
12	49036,15 - 53364,8	0	0	0	0
13	53364,8 - 57693,45	0	0	0	0
14	57693,45 - 62022,1	0	0	0	0
15	62022,1 - 66350,75	0	0	0	0
16	66350,75 - 70679,4	0	0	0	0
17	70679,4 - 75008,05	0	0	0	0
18	75008,05 - 79336,7	0	0	0	0
19	79336,7 - 83665,35	0	0	0	0
20	83665,35 - 87994	0	0	0	0





Оскільки в інших програмах я не зміг провести бінування, я вирішив опиратись на результати, отримані за допомогою програми **+V(L)binning_DynamicBinning&x_ave_pdf&cdf(ed Klymchuk)**. Ці результати дозволяють мені здійснити подальший аналіз і зробити висновки про структуру частотних розподілів у текстах, які я досліджую.

З результатів бінування, показаних у таблиці, можна зробити кілька висновків:

1. **Інтервали значень**: Дані розділені на 20 інтервалів, кожен з яких має свій діапазон значень (від "1421 - 5749,65" до "83665,35 - 87994"). Це може вказувати на широкий розподіл даних, який покриває діапазон від найменшого до найбільшого значення.
2. **Наявність даних в інтервалах**: З усіх інтервалів лише перший містить дані (значення в колонці "К-сть текстів" = 1). У всіх інших інтервалах кількість текстів (об'єктів) дорівнює нулю, що свідчить про те, що всі дані зосереджені в найнижчому інтервалі. Це може означати, що дані мають сильне зміщення до менших значень, або що діапазони інтервалів занадто широкі.
3. **Зміщення розподілу**: Якщо бінування налаштовано правильно, то отримані значення можуть вказувати на те, що розподіл даних є дуже асиметричним — переважна більшість значень (або навіть усі значення) потрапляють до першого інтервалу. Це може свідчити про те, що більшість значень зосереджена близько до нижньої межі розподілу.
4. **Корекція бінування**: Оскільки лише один інтервал містить дані, можливо, варто скоригувати параметри бінування, зменшити ширину інтервалів або використати адаптивне бінування, щоб отримати більше інтервалів із ненульовими значеннями. Це може надати більше інформації про розподіл даних.
5. **Колонки Vs, dV**: У таблиці ці колонки мають лише нульові значення, що може свідчити про відсутність даних для цих метрик або про те, що вони ще не були розраховані.

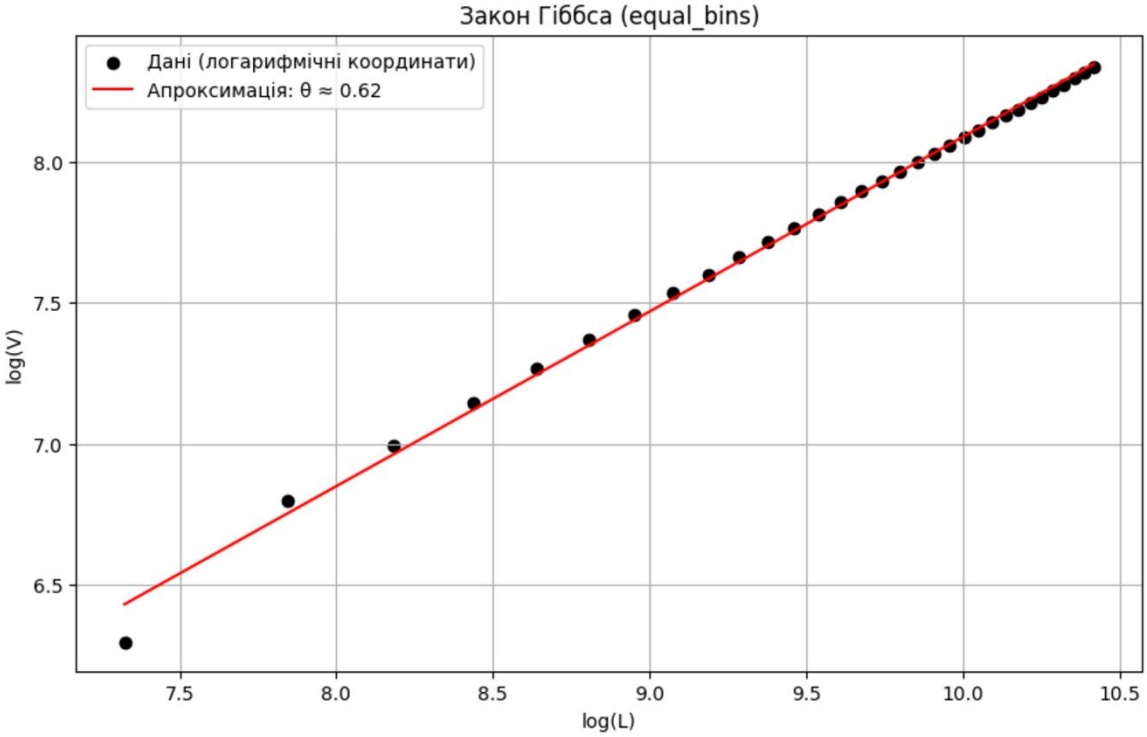
Все ж я вирішив спробувати реалізувати це через написану Python програму:

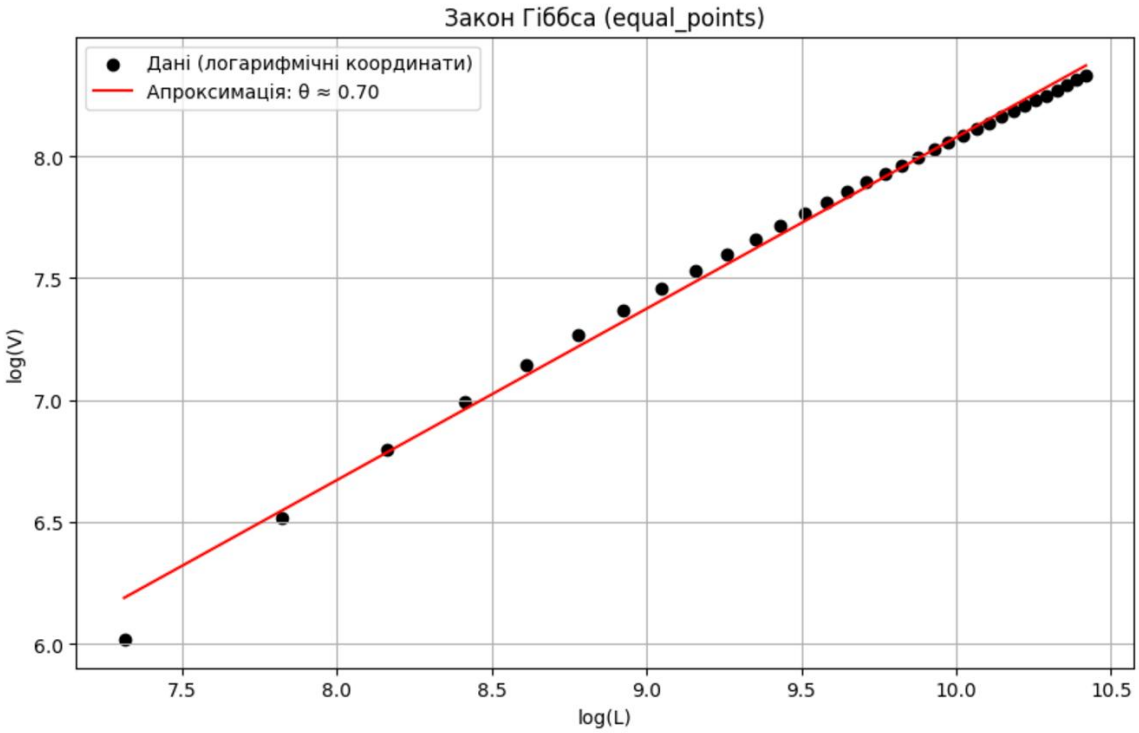
```python
import numpy as np
import matplotlib.pyplot as plt

def gibbs_law_calculation(lengths, volumes, method="equal_bins", bin_count=32):
    if method == "equal_bins":
        bins = np.linspace(np.min(lengths), np.max(lengths), bin_count + 1)
    elif method == "equal_points":
        step = len(lengths) // bin_count
        if step == 0:
            raise ValueError("Кількість точок у масиві L менша за кількість бінів")
        indices = list(range(0, len(lengths), step)) + [len(lengths) - 1]
        bins = [lengths[i] for i in indices]
    elif method == "exponential_bins":
        bins = np.logspace(np.log10(np.min(lengths)), np.log10(np.max(lengths)), bin_count + 1)
    else:
        raise ValueError("Unknown binning method")

    bin_centers, bin_averages = [], []
    for i in range(len(bins) - 1):
        mask = (lengths >= bins[i]) & (lengths < bins[i + 1])
        if np.sum(mask) > 0:
            bin_centers.append((bins[i] + bins[i + 1]) / 2)
            bin_averages.append(np.mean(volumes[mask]))

    return np.array(bin_centers), np.array(bin_averages)

def draw_graph(lengths, volumes, method, bin_count):
    centers, averages = gibbs_law_calculation(lengths, volumes, method, bin_count)

    # Логарифмічні перетворення
    log_centers = np.log(centers)
    log_averages = np.log(averages)

    # Лінійна апроксимація
    coefficients = np.polyfit(log_centers, log_averages, 1)
    slope, intercept = coefficients

    plt.figure(figsize=(10, 6))
    plt.scatter(log_centers, log_averages, color='black', label='Дані (логарифмічні координати)')
    plt.plot(log_centers, np.polyval(coefficients, log_centers), color='red', label=f"Апроксимація: θ ≈ {slope:.2f}")
    plt.title(f"Закон Гіббса ({method})")
    plt.xlabel("log(L)")
    plt.ylabel("log(V)")
    plt.legend()
    plt.grid(True)
    plt.show()

    return slope

data_file = "1.1fix.txt"
data = np.loadtxt(data_file, delimiter='\t', skiprows=1)
L, V = data[:, 0], data[:, 1]

methods = ["equal_bins", "equal_points", "exponential_bins"]
num_bins = 32
results = {}

for method in methods:
    slope = draw_graph(L, V, method, num_bins)
    results[method] = slope

print("Результати дослідження:")
for method, slope in results.items():
    print(f"Метод {method}: θ ≈ {slope:.2f}")
```

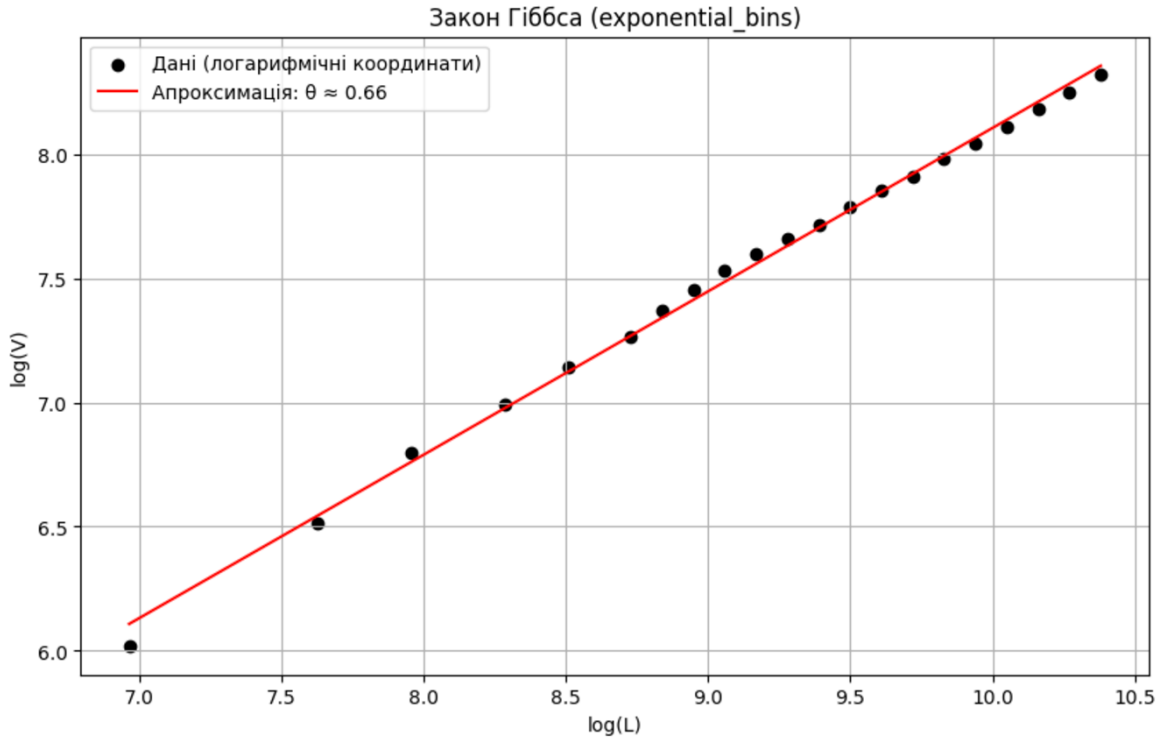
Програма виконує бінування даних трьома способами: з однаковою довжиною, де дані розділяються на біни однакової ширини; з однаковою кількістю точок, де біни містять однакову кількість емпіричних точок; та експоненційним способом, де біни зростають експоненційно.

Результат виконання програми на основі двох файлів:

Clemencia Novela de costumbres by Fernán Caballero

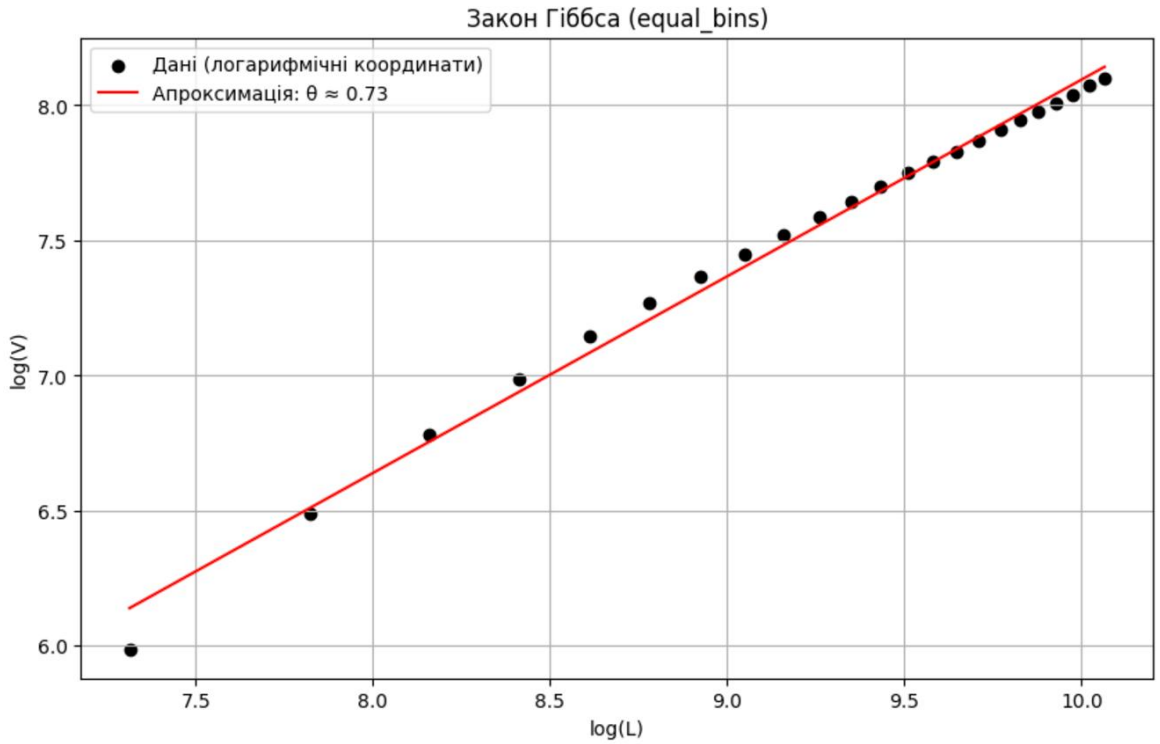


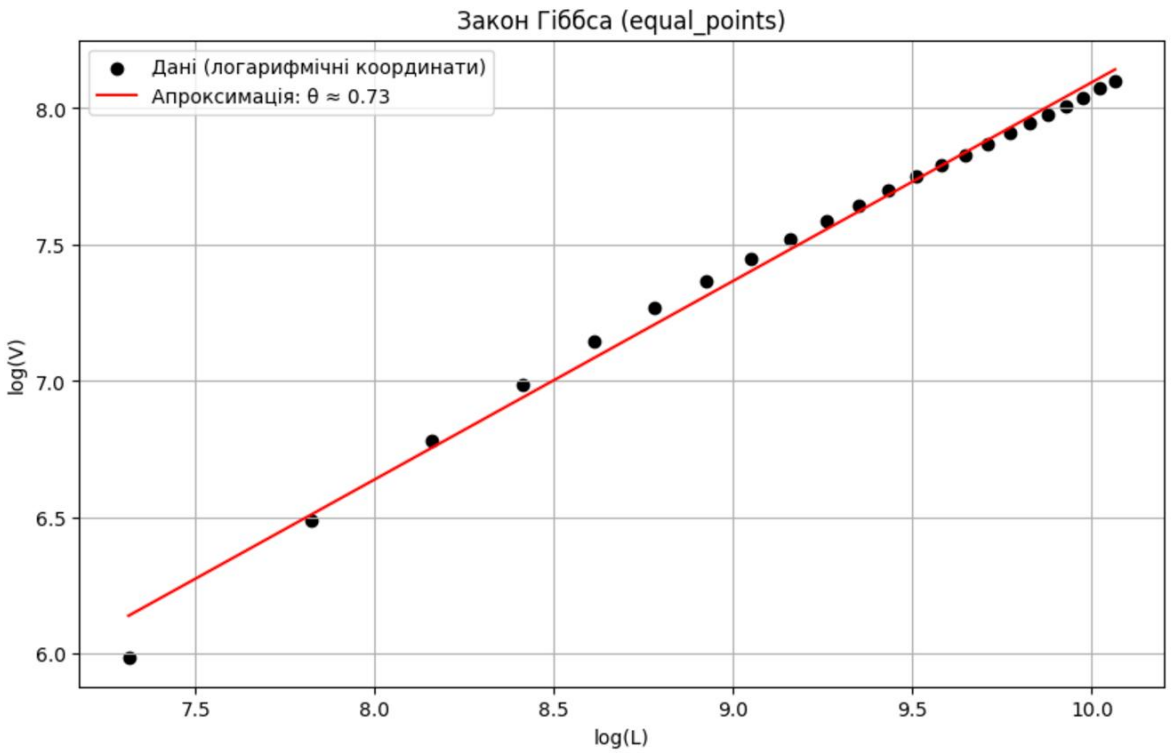


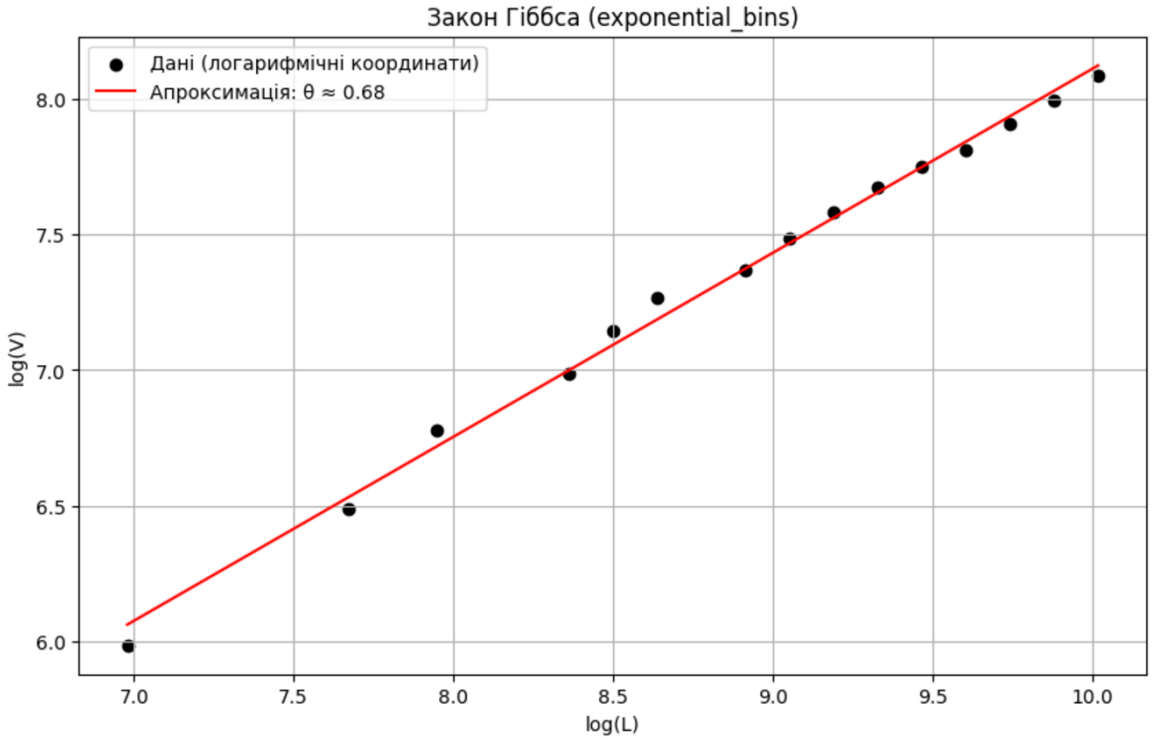



```
Результати дослідження:
Метод equal_bins: θ ≈ 0.62
Метод equal_points: θ ≈ 0.70
Метод exponential_bins: θ ≈ 0.66
```

DONALD J. TRUMP January 20, 2017








```
Результати дослідження:
Метод equal_bins: θ ≈ 0.73
Метод equal_points: θ ≈ 0.73
Метод exponential_bins: θ ≈ 0.68
```

Висновок: В процесі виконання цієї лабораторної роботи я досліджував процес розширення лексикону в межах текстового корпусу з використанням методу бінування. З аналізу таблиці можна зробити висновок, що більшість даних (або всі дані) зосереджені в першому інтервалі. Це може вказувати на нерівномірний розподіл даних або надто широкі інтервали.

А також я дослідив закон Гібса для корпусу текстів, використовуючи три методи бінування: з однаковими довжинами бінів, з однаковою кількістю емпіричних точок та експоненційне бінування. Це все з допомогою окремо написаної програми на Python для обробки даних і побудови графіків у логарифмічному масштабі, а також для визначення коефіцієнта θ за допомогою лінійної апроксимації. Проведене дослідження підтвердило ефективність закону Гібса для аналізу зростання словника у текстах природної мови, а також вплив різних методів бінування на результати.