Лабораторна робота №02

«Закони Ціпфа та Парето для слів у текстах»

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.

Теоретичні відомості

1. конспект лекцій
2. стаття Zipf's and Heaps' laws for the natural and some related random texts2018.pdf
3. стаття Statistical regularities of the linguistics of computer programs2016.pdf

Порядок виконання роботи та вказівки до оформлення звіту

1. Звіт має містити титульну сторінку, текст завдання, коротку теоретичну частину, опис виконання завдання та результатів і висновок.

2. У теоретичній частині сформулювати закони, які Ви досліджуєте в цій лабораторній роботі: виписати формули і взаємні зв'язки між показниками степенів, які стосуються першого і другого законів Ціпфа, а також закону Парето та закону Гіпса.

3. Обрати для дослідження два тексти різними природними мовами.

Зауваження 1. Як і у всіх решті лабораторних роботах, чітко зазначити в звіті, які саме тексти або бази текстів було обрано!

Зауваження 2. Основні завдання цього лабораторного практикуму типово передбачають вивчення одного або кількох текстів різними мовами, одного малого текстового корпусу та одного великого корпусу. *Ідеальною ситуацією* є обрання Вами *базових* текстів і корпусів одразу перед початком виконання першої ж лабораторної роботи. Далі в наступних роботах Ви будете досліджувати ті самі тексти, поступово вивчаючи їхні різні властивості. Зазначимо, що для виконання окремих робіт (найперше роботи №19) бажано обрати тексти, зміст яких Ви хоч трохи знаєте, хоча загалом це не обов'язкова вимога. Зрештою, для роботи №19 Ви можете обрати окремий текст, проте його не повинні вивчати інші студенти!

Приклад: Ви обираєте один англійський (або німецький, французький чи іспанський тощо) текст, один український (або китайський, японський, французький тощо) текст і один текст шведською (словацькою, білоруською тощо) мовою.

Один із цих текстів Ви можете також обрати для рандомізації в лабораторних роботах №№9–11, порівнюючи властивості вихідного природного і рандомізованого текстів.

Зазначимо, що лише в основному архіві +main text corpora2023-24.zip для Вас підготовлено тексти понад 70 мовами світу із сумарним обсягом майже 3 ГБ.

Тому неприпустимо обрання різними студентами текстів, обраних іншими студентами! Для уникнення такої ситуації просимо погоджувати обрання текстів зі старостами.

4. Ознайомитися із інтерфейсом та функціональними можливостями програми +proj6stats&plots.

5. Зобразити графічно залежності частоти від рангу F(r) (перший закон Ціпфа або рангову залежність); ймовірності випадання слова певної частоти від частоти pmf(F) (другий закон Ціпфа або розподіл імовірності частоти слів), а також залежність кумулятивної ймовірності від частоти cmf(F) (закон Парето) в лінійному масштабі.

6. Зобразити ті ж залежності в різних (лінійному або логарифмічному) масштабах по осях абсцис і ординат; обрати масштаб, у якому залежності лінійні або найближчі до лінійних. На цій основі встановити, які математичні функції описують дані графіки.

   Рекомендація: в пунктах 4 і 5 використовувати графіки з точками, а не з'єднувальними лініями, наприклад, plt.scatter у matplotlib.pyplot або Plot -> Symbol в Origin.

7. Графіки в логарифмічному масштабі апроксимувати прямою лінією; її нахил (slope) приблизно відповідає показнику степеня відповідного закону. Якщо весь графік помітно відрізняється від лінійного, то провести лінійну апроксимацію на ділянці осі абсцис, для якої графік найближчий до прямої лінії.

8. У звіті вказати значення цих показників, а також загальну статистичну інформацію про апроксимацію: коефіцієнт кореляції залежної та незалежної змінних, стандартна похибка тощо; прямі, отримані при апроксимації, теж навести на графіках з пункту 5.

9. Перевірити виконання теоретичних співвідношень між показниками степеня α, β, k.

10. Висновок повинен містити короткий аналіз особливостей ваших графіків, апроксимації, отриманих показників степенів, співвідношень між ними і їхньої природи, порівняння отриманих характеристик для двох Ваших тексту. [Як приклад, дивіться останні абзаци статей].

Додаток
Деякі приклади незваженої лінійної апроксимації в програмному пакеті Origin

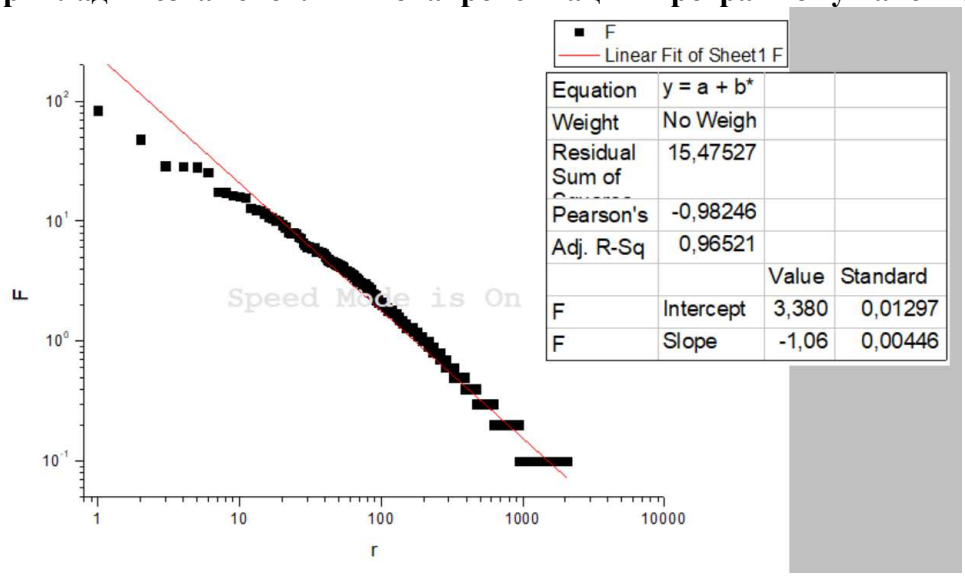


Рис. 1. Перевірка 1-го закону Ціпфа (залежність значень частоти слова від його рангу). Осям абсцис і ординат відповідають логарифмічні шкали (основа 10). За нахилом лінійної апроксимації (slope) знаходять значення коефіцієнта α. У нашому випадку маємо α = 1.06.

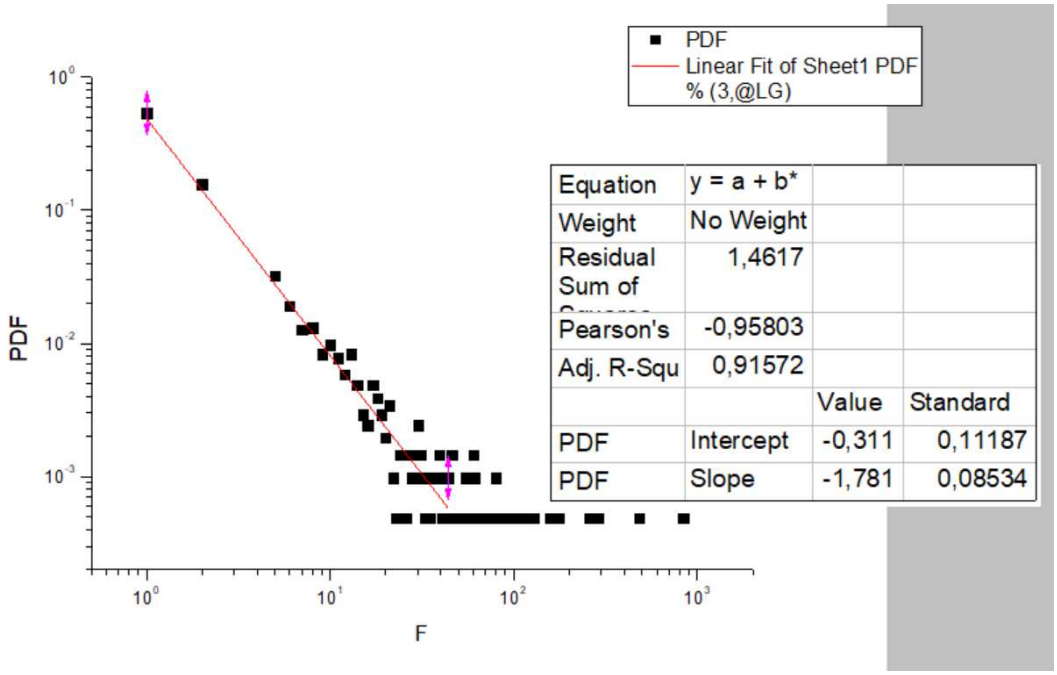


Рис. 2. Перевірка 2-го закону Ціпфа (залежності ймовірності випадання слова певної частоти від цієї частоти, що в наших даних записано як залежність pmf від частоти F). Шкали по осях абсцис і ординат логарифмічні, як і вище. Значення коефіцієнта β визначають з нахилу лінійної апроксимації: він дорівнює – 1.78, тобто маємо функцію pmf ~ $F^{-1.78}$.

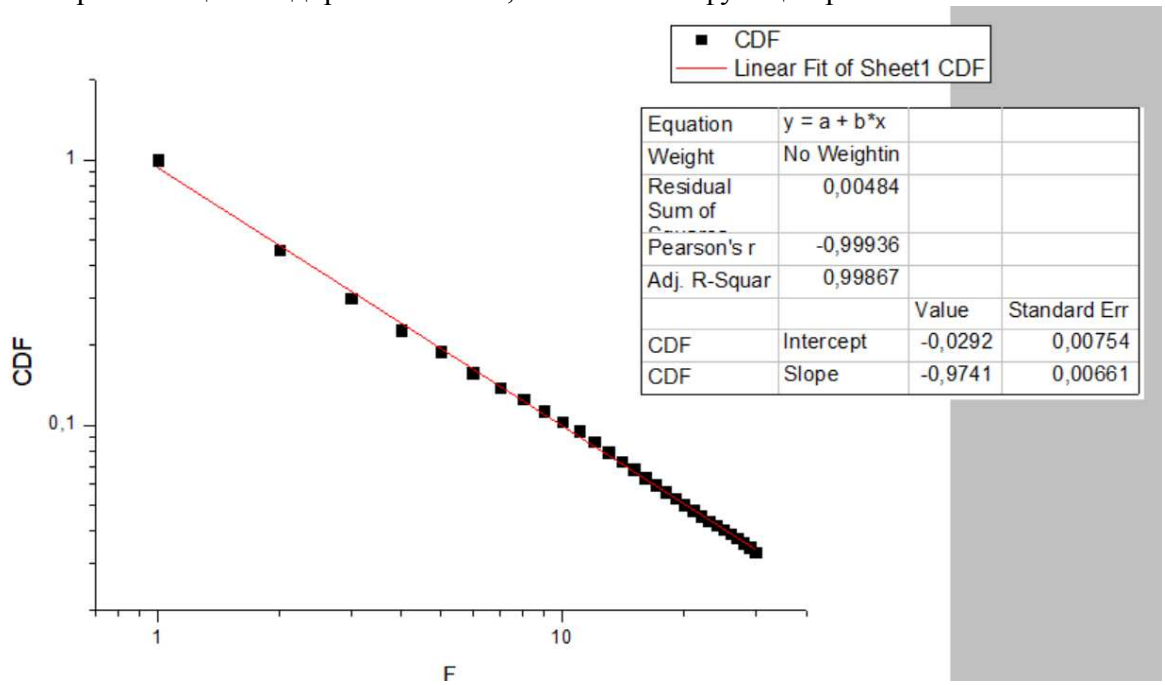
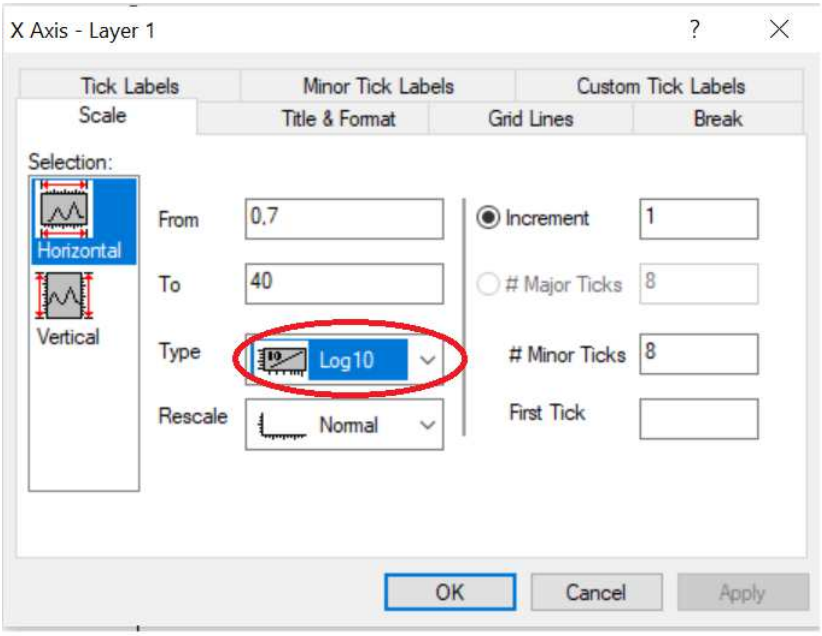


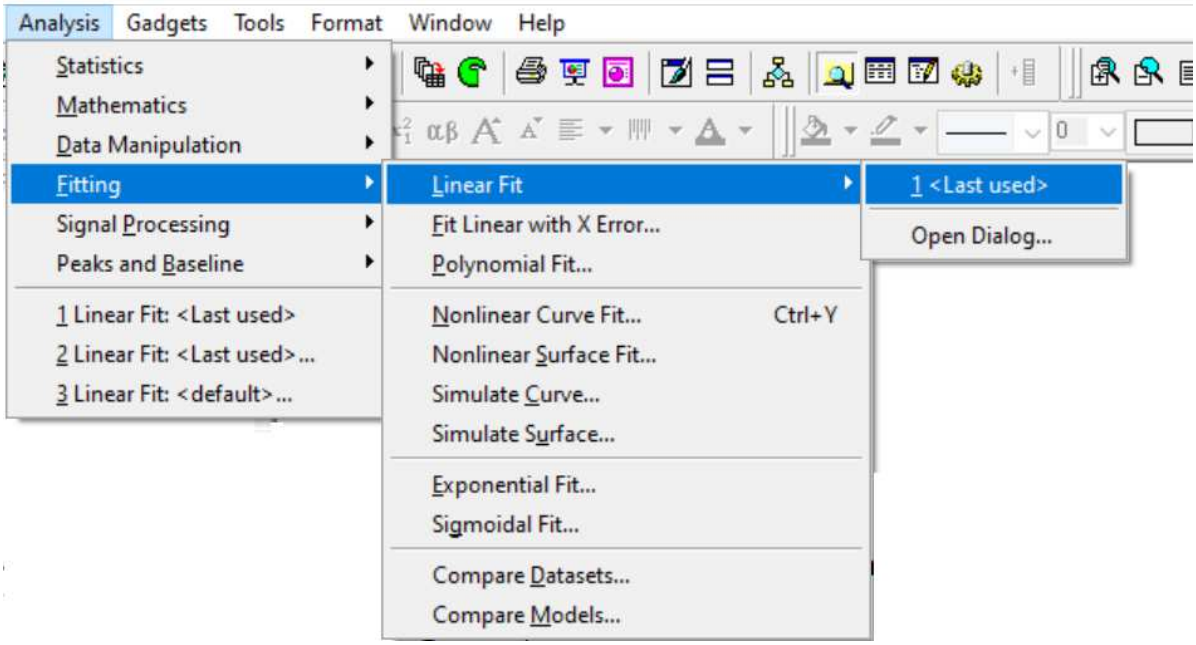
Рис. 3. Перевірка закону Парето (залежності кумулятивної функції розподілу від частоти cdf(F). Коефіцієнт k (або slope) наближено дорівнює 0.97.

Додаткові деталі

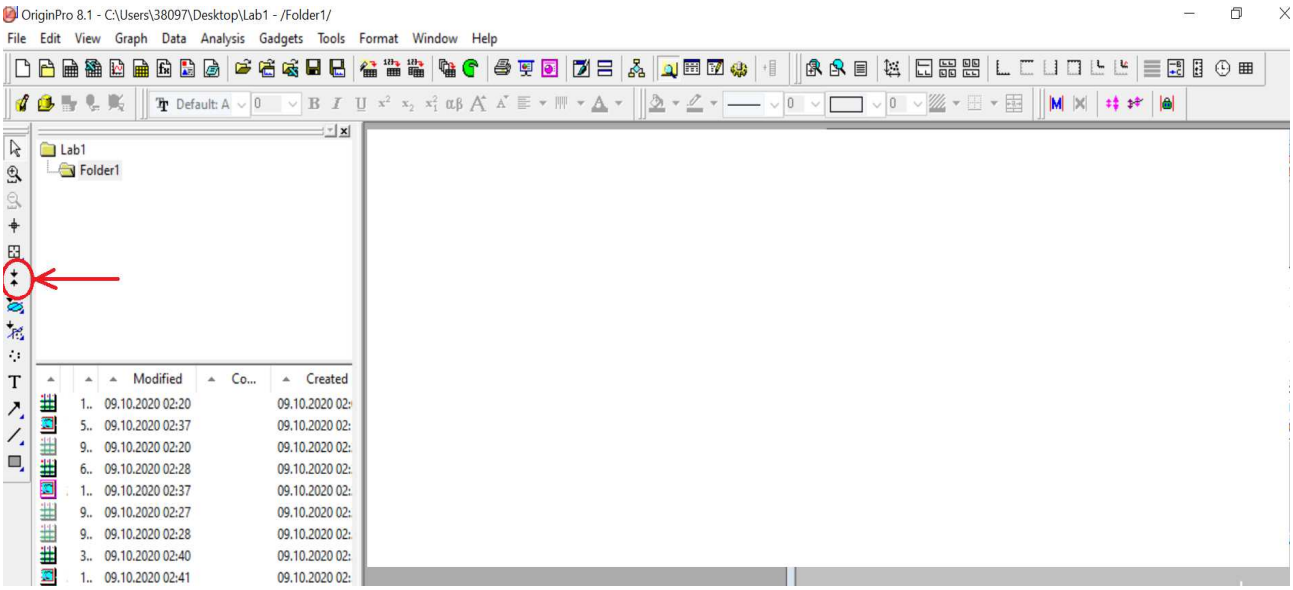
Після побудови графіка як функції y(x) в лінійному масштабі приведіть обидві шкали до логарифмічного масштабу. Для цього двічі клацніть по кожній із осей і оберіть для них логарифмічну шкалу:



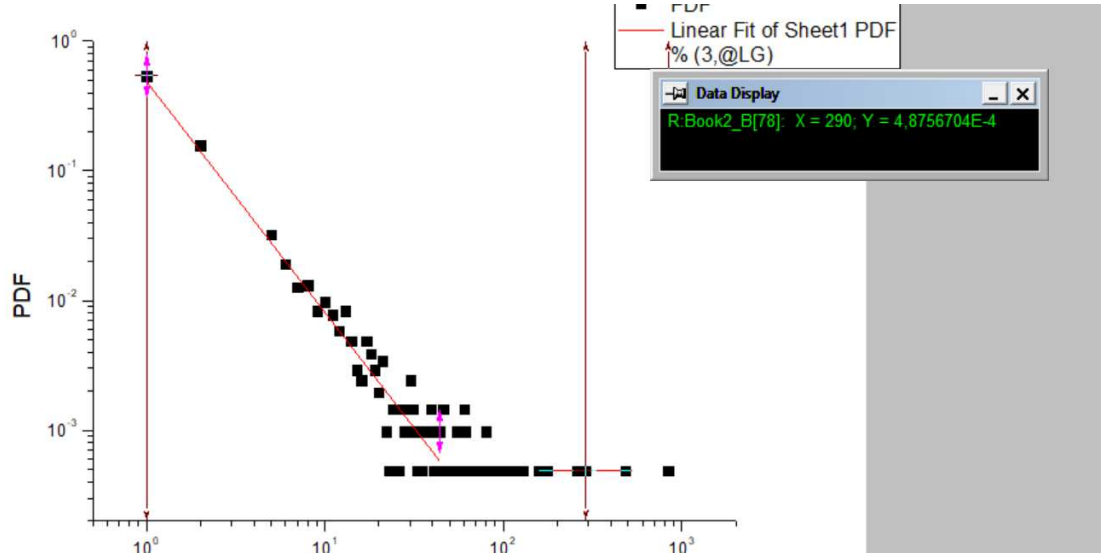
Лінійну апроксимацію в програмі реалізують так:



Якщо на «хвості» залежності помітні сильні флуктуації, то під час апроксимації відкидаємо дані, які відповідають «хвостові» (див. рис. 2). Це означає, що спочатку слід виставити межі апроксимацією. Це здійснюють так, як показано на рисунку внизу:



З’являться межі на активному графіку. Тепер слід належно виставити ці межі:



Після цього можна виконувати лінійну апроксимацію.

Нарешті, для того аби апроксимація виконувалася саме в подвійному логарифмічному масштабі, а не у початковому лінійному масштабі, в меню апроксимації (Analysis -> Fitting -> Linear Fit) при першому використанні слід обрати «Open Dialog» та поставити «пташку» в полі «Apparent Fit». У подальшому використанні програми можна використовувати опцію «Last Used», а не щоразу відкривати діалогове вікно.