**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 № 15

**«Визначення середньої довжини слів і речень»**

Виконав:

Студент групи

Фес-32с

Бойко Кирило

Львів 2024

***Завдання***

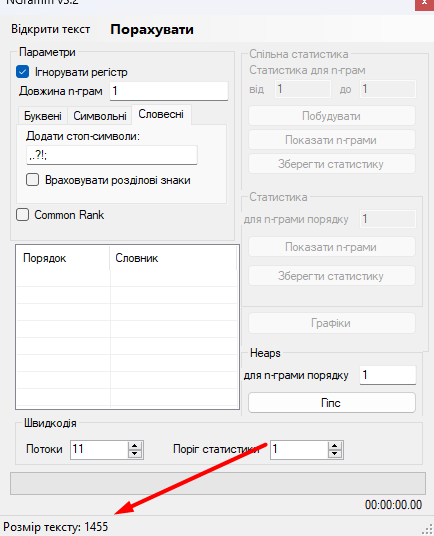
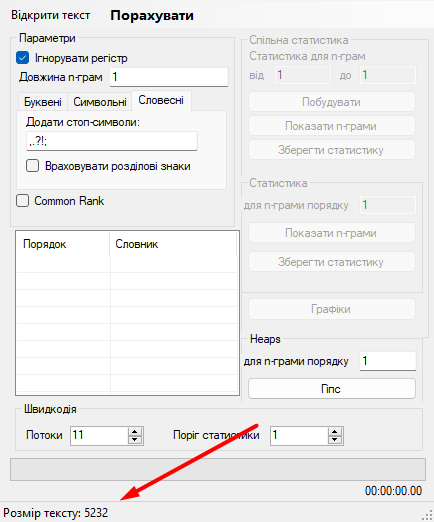
Використовуючи програми **+LoSW\_sliding window(single text)** і **+LoSW\_(corpus of texts)**, дослідити закономірності для довжин слів і речень для двох випадків: єдиного тексту англійською, українською та російською мовами, а також для корпусу текстів однією з цих мов

**Хід виконання лабораторної роботи**

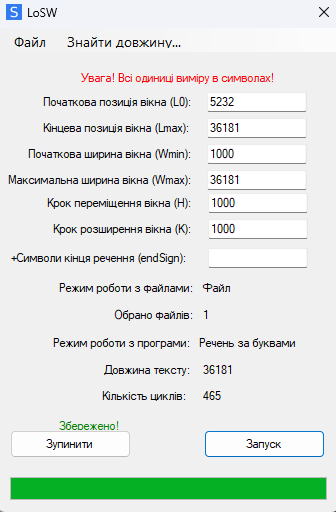
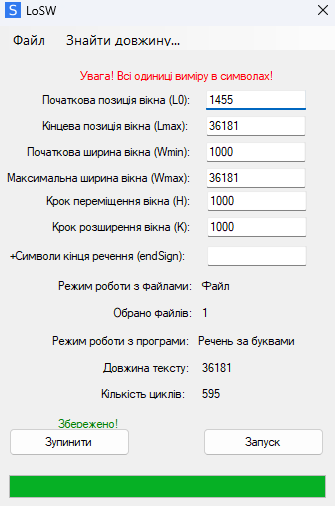
Запустив програму **+LoSW\_running window\_single text** та завантажив тексти, якi використовував в минулих лабораторних роботах, для подальшого аналізу

***DONALD J. TRUMP January 20, 2017.txt***

***Людина-маятник.txt***

Початкову ширину вікна, крок приросту ширини вікна та крок ковзання біжучого вікна я встановив на основі довжини тексту *L*, яку було визначено у програмі в лабораторній роботі 2. Ці параметри були обрані для забезпечення оптимального аналізу текстів  
  


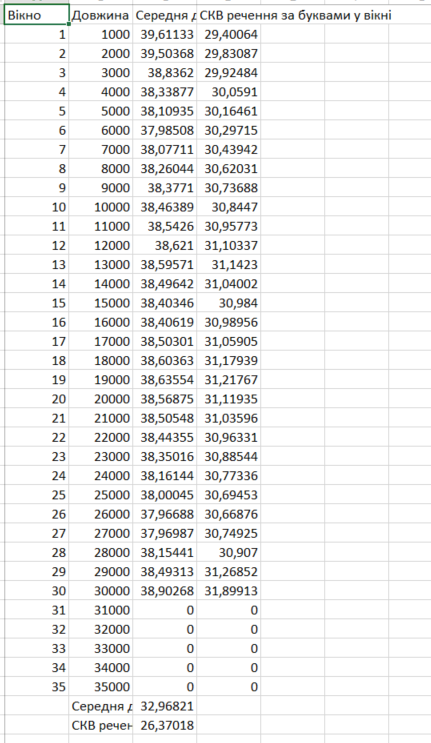
***DONALD J. TRUMP January 20, 2017.txt Людина-маятник.txt***



**Результати**

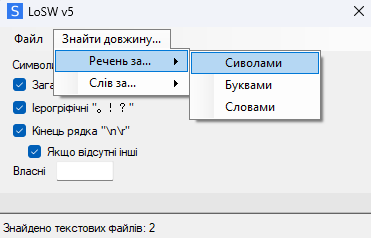
****

**DONALDJ.TRUMPJanuary 20,2017\_resultLab15.xlsx**

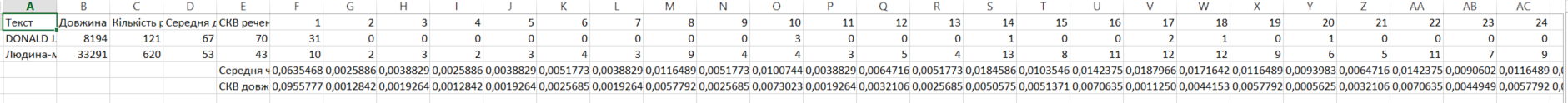
******

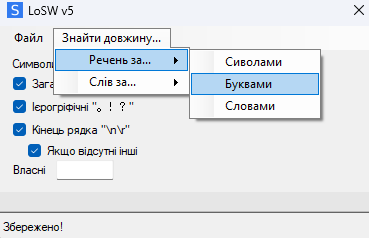
**Людина-маятник\_resultLab15.xlsx**

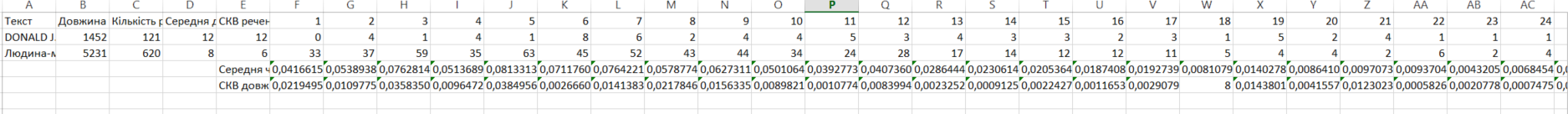
Запустивши програму **+LoSW\_corpus**, я завантажив тексти для подальшого аналізу

******

**Результати**







Я написала програму, яка досліджує степінь γ, коефіціент кореляції, стандартну похибку, коефіцієнт кореляції для експоненційного хвоста у 2-х файлах, які я отримала провівши дослідження у програмах вище  
  
**Код програми:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import linregress

corpus\_symbols\_path = 'res1 symbols.xlsx'

corpus\_bykvy\_path = 'res2 bykvy.xlsx'

def process\_and\_plot(file\_path, title\_prefix):

df = pd.read\_excel(file\_path, sheet\_name=0, header=None)

lengths = df.iloc[0, 5:].values.astype(float)

probability\_values = df.iloc[1, 5:].values.astype(float)

std\_dev\_values = df.iloc[2, 5:].values.astype(float)

nonzero\_indices = (probability\_values > 1e-10) & (std\_dev\_values > 1e-10)

lengths = lengths[nonzero\_indices]

probability\_values = probability\_values[nonzero\_indices]

std\_dev\_values = std\_dev\_values[nonzero\_indices]

plt.figure(figsize=(10, 6))

plt.plot(lengths, probability\_values, marker='o', linestyle='-', color='blue', markersize=6, linewidth=1.5)

plt.xlabel('Довжина слова/речення (l)')

plt.ylabel('Ймовірність p(l)')

plt.title(f'{title\_prefix}: Залежність ймовірності довжини від самої довжини p(l)')

plt.xlim([min(lengths)\*0.9, max(lengths)\*1.1])

plt.grid(True, which='both', linestyle='--', linewidth=0.5)

plt.yscale('log')

plt.show()

plt.figure(figsize=(10, 6))

plt.plot(probability\_values, std\_dev\_values, marker='o', linestyle='-', color='green', markersize=6, linewidth=1.5)

plt.xlabel('Ймовірність p')

plt.ylabel('СКВ ймовірності ∆p')

plt.title(f'{title\_prefix}: Залежність СКВ ймовірності від самої ймовірності ∆p(p)')

plt.grid(True, which='both', linestyle='--', linewidth=0.5)

plt.xscale('log')

plt.yscale('log')

plt.show()

log\_p = np.log10(probability\_values)

log\_std\_dev = np.log10(std\_dev\_values)

slope, intercept, r\_value, p\_value, std\_err = linregress(log\_p, log\_std\_dev)

plt.figure(figsize=(10, 6))

plt.plot(log\_p, log\_std\_dev, marker='o', label='Дані', color='purple', markersize=6)

plt.plot(log\_p, slope \* log\_p + intercept, label=f'Апроксимація (γ={slope:.2f})', color='red', linewidth=1.5)

plt.xlabel('log(p)')

plt.ylabel('log(∆p)')

plt.title(f'{title\_prefix}: Залежність ∆p(p) в подвійному логарифмічному масштабі')

plt.legend()

plt.grid(True, which='both', linestyle='--', linewidth=0.5)

plt.show()

print(f"{title\_prefix} - Степінь γ: {slope:.2f}")

print(f"{title\_prefix} - Коефіцієнт кореляції R: {r\_value:.2f}")

print(f"{title\_prefix} - Стандартна похибка: {std\_err:.2f}")

log\_prob\_values = np.log(probability\_values)

plt.figure(figsize=(10, 6))

plt.plot(lengths, log\_prob\_values, marker='o', linestyle='-', color='orange', markersize=6, linewidth=1.5)

plt.xlabel('Довжина слова/речення (l)')

plt.ylabel('log(p(l))')

plt.title(f'{title\_prefix}: Залежність p(l) в напівлогарифмічному масштабі')

plt.grid(True, which='both', linestyle='--', linewidth=0.5)

plt.show()

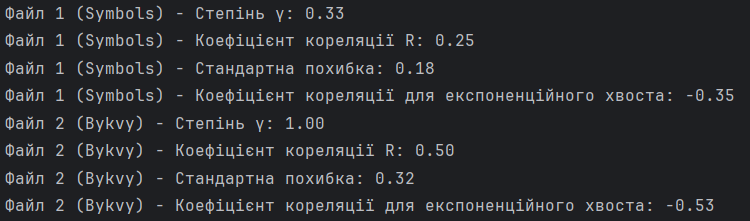
slope\_exp, intercept\_exp, r\_value\_exp, p\_value\_exp, std\_err\_exp = linregress(lengths, log\_prob\_values)

print(f"{title\_prefix} - Коефіцієнт кореляції для експоненційного хвоста: {r\_value\_exp:.2f}")

process\_and\_plot(corpus\_symbols\_path, "Файл 1 (Symbols)")

process\_and\_plot(corpus\_bykvy\_path, "Файл 2 (Bykvy)")

**Результат**

****

**Висновок:** У ході лабораторної роботи було проведено дослідження у програмах **+LoSW\_sliding window (single text)** і **+LoSW\_ (corpus of texts)**. Отримані результати показують, що для обох файлів спостерігаються позитивні значення степеня γ, що свідчить про існування певної залежності між ймовірністю та її стандартним відхиленням.