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

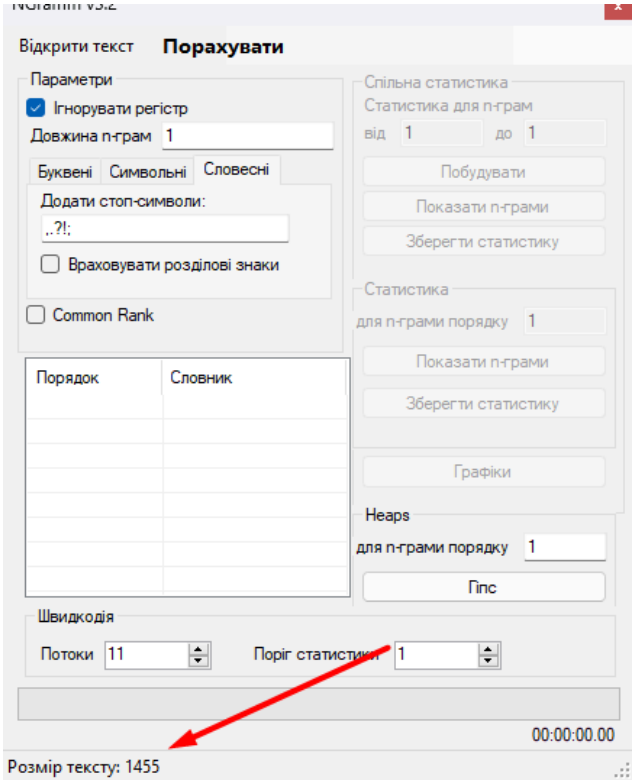
Хід виконання лабораторної роботи

Запустив програму **+LoSW_running window_single text** та завантажив тексти, які використовував в минулих лабораторних роботах, для подальшого аналізу

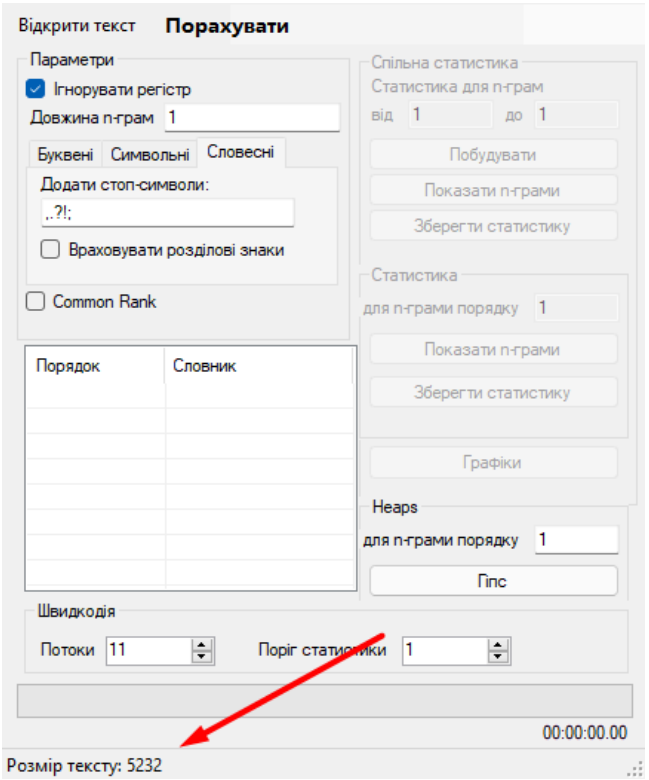
DONALD J. TRUMP January 20, 2017.txt

Людина-маятник.txt

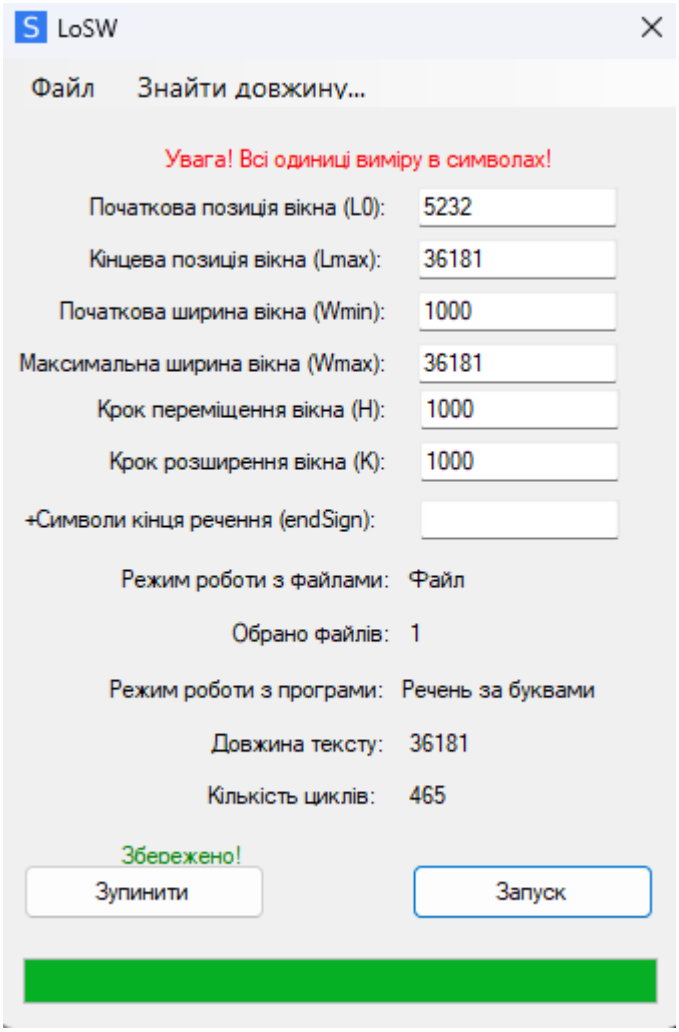
Початкову ширину вікна, крок приросту ширини вікна та крок ковзання біжучого вікна я встановив на основі довжини тексту *L*, яку було визначено у програмі в лабораторній роботі 2. Ці параметри були обрані для забезпечення оптимального аналізу текстів

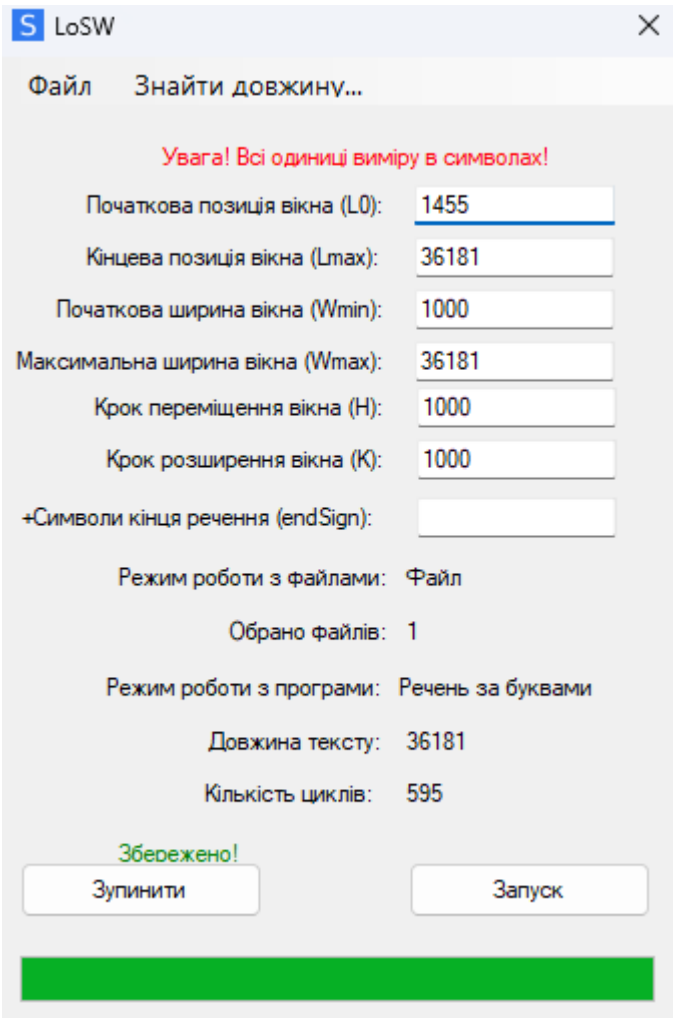


DONALD J. TRUMP January 20, 2017.txt



Людина-маятник.txt





Результати

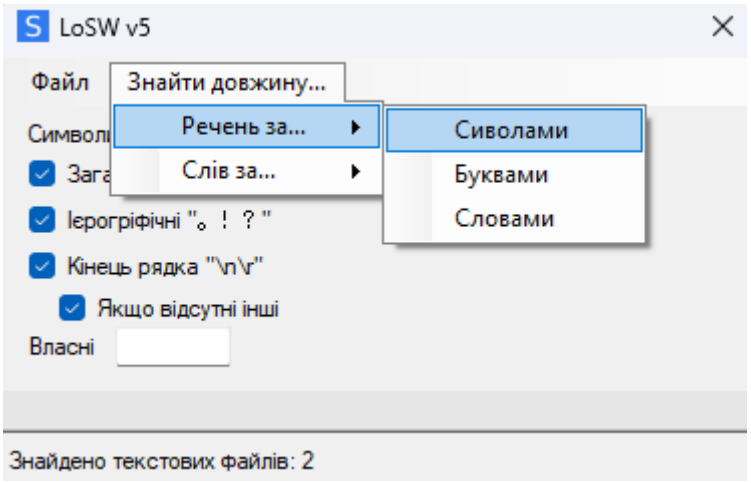
	A	B	C	D	E	F	G
	Вікно	Довжина	Середня д	СКВ речення за буквами у вікні			
2	1	1000	70,4896	49,70782			
3	2	2000	73,20745	54,35688			
4	3	3000	73,42588	53,00193			
5	4	4000	72,9069	52,40571			
6	5	5000	74,21062	53,8711			
7	6	6000	74,65737	56,55451			
8	7	7000	73,66667	55,87064			
9		Середня д	73,2235				
10		СКВ речен	53,68123				
11							

DONALDJ.TRUMPJanuary 20,2017_resultLab15.xlsx

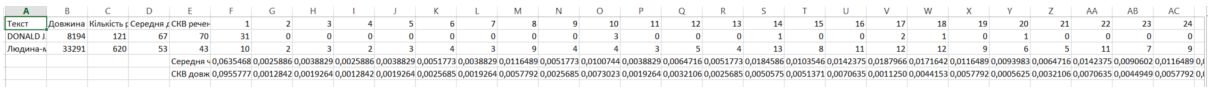
Вікно	Довжина	Середня д	СКВ речення за буквами у вікні
1	1000	39,61133	29,40064
2	2000	39,50368	29,83087
3	3000	38,8362	29,92484
4	4000	38,33877	30,0591
5	5000	38,10935	30,16461
6	6000	37,98508	30,29715
7	7000	38,07711	30,43942
8	8000	38,26044	30,62031
9	9000	38,3771	30,73688
10	10000	38,46389	30,8447
11	11000	38,5426	30,95773
12	12000	38,621	31,10337
13	13000	38,59571	31,1423
14	14000	38,49642	31,04002
15	15000	38,40346	30,984
16	16000	38,40619	30,98956
17	17000	38,50301	31,05905
18	18000	38,60363	31,17939
19	19000	38,63554	31,21767
20	20000	38,56875	31,11935
21	21000	38,50548	31,03596
22	22000	38,44355	30,96331
23	23000	38,35016	30,88544
24	24000	38,16144	30,77336
25	25000	38,00045	30,69453
26	26000	37,96688	30,66876
27	27000	37,96987	30,74925
28	28000	38,15441	30,907
29	29000	38,49313	31,26852
30	30000	38,90268	31,89913
31	31000	0	0
32	32000	0	0
33	33000	0	0
34	34000	0	0
35	35000	0	0
	Середня д	32,96821	
	СКВ речен	26,37018	

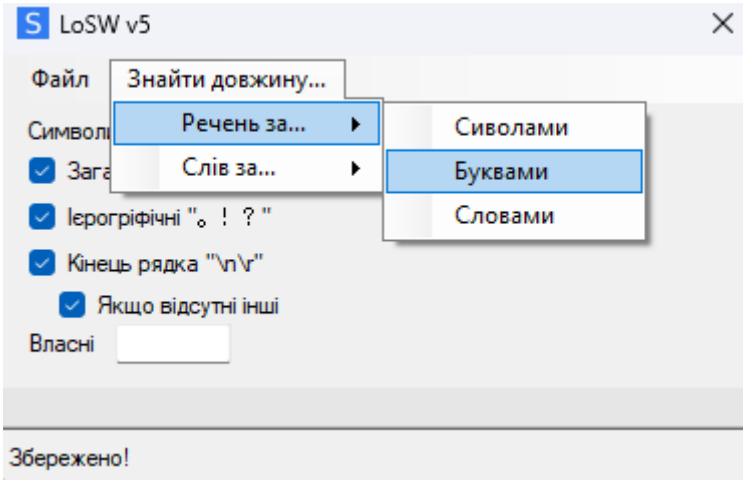
Людина-маятник_resultLab15.xlsx

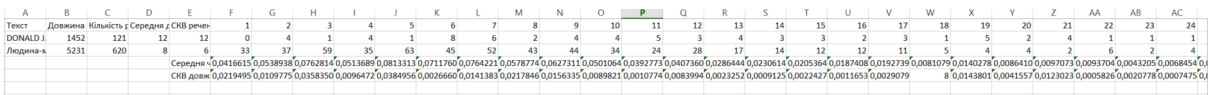
Запустивши програму **+LoSW_corpus**, я завантажив тексти для подальшого аналізу



Результати







Я написала програму, яка досліджує степінь γ, коефіцієнт 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 Δp')
    plt.title(f'{title_prefix}: Залежність СКВ ймовірності від
самої ймовірності Δp(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(Δp)')
    plt.title(f'{title_prefix}: Залежність Δp(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

```
Файл 1 (Symbols) - Степінь γ: 0.33
Файл 1 (Symbols) - Коефіцієнт кореляції R: 0.25
Файл 1 (Symbols) - Стандартна похибка: 0.18
Файл 1 (Symbols) - Коефіцієнт кореляції для експоненційного хвоста: -0.35
Файл 2 (Bykvy) - Степінь γ: 1.00
Файл 2 (Bykvy) - Коефіцієнт кореляції R: 0.50
Файл 2 (Bykvy) - Стандартна похибка: 0.32
Файл 2 (Bykvy) - Коефіцієнт кореляції для експоненційного хвоста: -0.53
```

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.