# Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського" Фізико-технічний інститут

# Криптографія

Комп'ютерний практикум No1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: Студенти групи ФБ-32 Коптєва Ганна, Чупріна Вікторія **Мета роботи**: Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

#### Порядок виконання роботи:

- 1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30)
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

#### Хід роботи:

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.

Текст: Детская сказка: «Леди и Бродяга»

# Завдання 1

Написати програми для підрахунку частот букв і частот біграм у тексті, а також підрахунки  $H^1$  та  $H^2$  за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення  $H^1$  та  $H^2$  на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1МБ), де ймовірність замінити відповідними частотами. Також отримати значення  $H^1$  та  $H^2$  на тому ж тексті, у якому вилучено всі пробіли.

**Частоти літер** Без пробілів (unigrams\_no\_spaces.csv):

char	count	freq
0	65358	0.1133082300
е	48944	0.0848520152
a	46206	0.0801052675
И	38015	0.0659048986
Т	37491	0.0649964633
н	36914	0.0639961444
С	30368	0.0526476381
В	26551	0.0460302766
л	26436	0.0458309062
р	25143	0.0435892902
К	19184	0.0332584394
Д	18285	0.0316998835
М	18047	0.0312872736
у	17034	0.0295310810
п	15915	0.0275911209
Ь	13423	0.0232708524
я	12315	0.0213499626
ч	10074	0.0174648415
б	9931	0.0172169288
Γ	9807	0.0170019556
ы	9481	0.0164367840
3	9143	0.0158508086
ж	6440	0.0111647388
й	5816	0.0100829381
x	5044	0.0087445563
ш	4882	0.0084637042
ю	3363	0.0058302821
э	1969	0.0034135669
ц	1775	0.0030772378
щ	1736	0.0030096253
ф	904	0.0015672242
ë	695	0.0012048903
ъ	127	0.0002201742

# 3 пробілами (unigrams\_with\_spaces.csv):

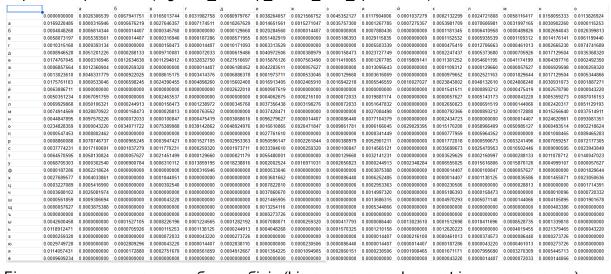
char	count	freq			
	117309	0.1690027012			
0	65358	0.0941588331			
е	48944	0.0705117954			
a	46206	0.0665672609			
И	38015	0.0547667927			
Т	37491	0.0540118855			
н	36914	0.0531806231			
С	30368	0.0437500450			
В	26551	0.0382510355			
л	26436	0.0380853593			
р	25143	0.0362225824			
К	19184	0.0276376733			
Д	18285	0.0263425176			
М	18047	0.0259996398			
у	17034	0.0245402485			
п	15915	0.0229281469			
Ь	13423	0.0193380155			
я	12315	0.0177417612			
Ч	10074	0.0145132361			
6	9931	0.0143072213			
Γ	9807	0.0141285791			
ы	9481	0.0136589231			
3	9143	0.0131719791			
ж	6440	0.0092778678			
й	5816	0.0083788943			
x	5044	0.0072667027			
ш	4882	0.0070333153			
ю	3363	0.0048449487			
э	1969	0.0028366649			
ц	1775	0.0025571763			
щ	1736	0.0025009905			
ф	904	0.0013023591			
ë	695	0.0010012606			
ъ	127	0.0001829642			

# Частоти біграм

#### Біграми, без пробілів (bigrams overlap matrix no spaces.csv):

	a	6	В	r	д	е	ж	3	и	й	K	л	м	н	0	п
a	0.0003571336	0.0013869265	0.0052633860	0.0012534348	0.0032575436	0.0022849614	0.0020283800	0.0048438407	0.0016556435	0.0008824320	0.0075171415	0.0088555256	0.0045803247	0.0056707957	0.0013297158	0.0029038773
6	0.0008234876	0.0000052010	0.0000710800	0.0000000000	0.0000156029	0.0024531262	0.0000017337	0.0000121356	0.0009812505	0.0000000000	0.0001542956	0.0007714779	0.0000641453	0.0003536663	0.0025398091	0.0000138693
В	0.0067525983	0.0002409785	0.0004802233	0.0002964555	0.0013453187	0.0063399877	0.0000364068	0.0007784125	0.0037949776	0.0000104019	0.0006830613	0.0008338895	0.0002617824	0.0023473731	0.0077858586	0.0011563500
r	0.0011494153	0.0000329395	0.0001404263	0.0000242712	0.0012863743	0.0004212789	0.0000034673	0.0000537434	0.0007264027	0.0000000000	0.0001248234	0.0015394884	0.0000866829	0.0004108770	0.0091363782	0.0001612302
д	0.0061804911	0.0000693463	0.0012655704	0.0000242712	0.0001352253	0.0049565285	0.0000537434	0.0000676127	0.0029298822	0.0000000000	0.0003467316	0.0006657247	0.0001126878	0.0022017458	0.0043116077	0.0001837678
e	0.0003051238	0.0024115184	0.0036025415	0.0043584165	0.0043774867	0.0022537555	0.0011303451	0.0021133292	0.0015186845	0.0023889809	0.0021462687	0.0069433007	0.0059689849	0.0095125820	0.0015793625	0.0035852050
ж	0.0014961469	0.0000693463	0.0000537434	0.0000069346	0.0007714779	0.0051021558	0.0000069346	0.0000173366	0.0016660454	0.0000000000	0.0002115063	0.0000260049	0.0000277385	0.0009327081	0.0001230897	0.0000502761
3	0.0055823791	0.0001334917	0.0011944904	0.0004420828	0.0009240398	0.0002704507	0.0000745473	0.0000658790	0.0004784896	0.0000000000	0.0002409785	0.0003293950	0.0002808526	0.0022121477	0.0007073325	0.0001959034
И	0.0003085911	0.0016608445	0.0052252455	0.0009552456	0.0029073446	0.0025034023	0.0004819570	0.0026646325	0.0018480795	0.0012308973	0.0039735444	0.0060678034	0.0035938733	0.0057592122	0.0017180552	0.0023716443
й	0.0001109541	0.0002323102	0.0005686399	0.0002600487	0.0007385384	0.0002063053	0.0001005522	0.0001664312	0.0006466545	0.0000069346	0.0005929111	0.0001976370	0.0004022087	0.0009049695	0.0003952740	0.0007836135
K	0.0081308565	0.0006327852	0.0007940154	0.0001040195	0.0002721843	0.0005790418	0.0001560292	0.0001074868	0.0028345310	0.0000052010	0.0002496468	0.0007264027	0.0002149736	0.0012274299	0.0108318958	0.0003900731
n	0.0062012950	0.0002929882	0.0007662769	0.0003051238	0.0003397970	0.0048577100	0.0004126106	0.0002167073	0.0068479495	0.0000000000	0.0006795940	0.0001577629	0.0001889687	0.0010367275	0.0073264392	0.0007541413
м	0.0035973406	0.0003016565	0.0008338895	0.0002756516	0.0004646204	0.0046410027	0.0001629639	0.0002357775	0.0039614088	0.0000034673	0.0004715550	0.0002253756	0.0002999229	0.0026126228	0.0051836377	0.0010124563
н	0.0118062117	0.0002791190	0.0006431872	0.0001508283	0.0005426350	0.0116900566	0.0000225376	0.0001490946	0.0096287371	0.0000104019	0.0005565043	0.0000797483	0.0001438936	0.0034170401	0.0115340274	0.0006310516
0	0.0002427121	0.0055321030	0.0122552291	0.0052200446	0.0071894802	0.0027218432	0.0030130978	0.0018203410	0.0023179009	0.0035071903	0.0030148314	0.0075258098	0.0071236012	0.0098766502	0.0022641575	0.0042405277
п	0.0010592651	0.0000069346	0.0000034673	0.0000017337	0.0000034673	0.0033424928	0.0000000000	0.0000052010	0.0010124563	0.0000000000	0.0000953512	0.0007194681	0.0000017337	0.0001230897	0.0107694842	0.0000641453
р	0.0094969791	0.0001490946	0.0005409013	0.0002063053	0.0003900731	0.0060990092	0.0002808526	0.0000589444	0.0064977506	0.0000017337	0.0002583151	0.0000762810	0.0004334145	0.0009344417	0.0087601744	0.0003137921
С	0.0021323995	0.0002895209	0.0017648640	0.0001092205	0.0004802233	0.0040844985	0.0002687170	0.0001109541	0.0019573000	0.0000017337	0.0045421842	0.0031777953	0.0014545392	0.0016365733	0.0031188509	0.0025328745
т	0.0072224197	0.0004126106	0.0032991514	0.0001577629	0.0004576857	0.0068808890	0.0001248234	0.0001976370	0.0044780389	0.0000173366	0.0008304222	0.0003501989	0.0002947219	0.0019781039	0.0173625859	0.0007576086
у	0.0002271092	0.0010367275	0.0015169508	0.0017926025	0.0025952862	0.0003588672	0.0021809419	0.0005357004	0.0007055988	0.0002877872	0.0011979578	0.0019659683	0.0022537555	0.0015256191	0.0004524848	0.0014684084
ф	0.0002687170	0.0000017337	0.0000017337	0.0000381405	0.0000017337	0.0000641453	0.0000000000	0.0000000000	0.0004732887	0.0000000000	0.0000017337	0.0000138693	0.0000069346	0.0000000000	0.0002201746	0.0000017337
x	0.0005270321	0.0001456273	0.0005010272	0.0000606780	0.0002253756	0.0004958262	0.0000329395	0.0000866829	0.0010557978	0.0000000000	0.0001889687	0.0002184409	0.0002305765	0.0004056760	0.0029610880	0.0003935404
ц	0.0006622574	0.0000034673	0.0000797483	0.0000069346	0.0000086683	0.0009483110	0.0000017337	0.0000121356	0.0003848721	0.0000000000	0.0000468088	0.0000017337	0.0000104019	0.0000277385	0.0002288429	0.0000537434
ч	0.0030234997	0.0000190702	0.0000502761	0.0000086683	0.0000121356	0.0045629881	0.0000017337	0.0000069346	0.0018411449	0.0000000000	0.0006605237	0.0000260049	0.0000156029	0.0010193910	0.0001057531	0.0000502761
ш	0.0011216768	0.0000069346	0.0000104019	0.0000034673	0.0000034673	0.0025831506	0.0000017337	0.0000017337	0.0016504425	0.0000000000	0.0005998457	0.0006102477	0.0000208039	0.0004992935	0.0002340438	0.0000086683
щ	0.0004663540	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0015984328	0.0000000000	0.0000000000	0.0006449208	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000485424	0.0000000000	0.0000000000
ъ	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000329395	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
ы	0.0000572107	0.0003883394	0.0015030816	0.0002496468	0.0004299472	0.0009881851	0.0000693463	0.0002115063	0.0004420828	0.0015672269	0.0003952740	0.0022606902	0.0012603694	0.0007073325	0.0002999229	0.0005339667
ь	0.0003103248	0.0004958262	0.0013574543	0.0003085911	0.0006657247	0.0009465773	0.0001092205	0.0005582379	0.0014129314	0.0000017337	0.0023092326	0.0002011043	0.0009448437	0.0041105034	0.0009240398	0.0012499675
9	0.0000000000	0.0000000000	0.0000104019	0.0000052010	0.0000346732	0.0000000000	0.0000017337	0.0000000000	0.0000017337	0.0000277385	0.0000554771	0.0000450751	0.0000104019	0.0000346732	0.0000017337	0.0000069346
ю	0.0000936175	0.0004351482	0.0003120585	0.0000641453	0.0005148965	0.0000398741	0.0000936175	0.0000901502	0.0003120585	0.0000034673	0.0002548477	0.0000606780	0.0001959034	0.0003068575	0.0002028380	0.0003363297
я	0.0003918067	0.0004732887	0.0018151400	0.0002999229	0.0012707714	0.0004784896	0.0003016565	0.0006293179	0.0011598173	0.0001282907	0.0008130857	0.0011390134	0.0007784125	0.0020873244	0.0008442915	0.0012534348
ŏ	0.0000000000	0.0000710800	0.0001057531	0.0000121356	0.0000381405	0.0000658790	0.0000104019	0.0000329395	0.0000381405	0.0000000000	0.0000450751	0.0000173366	0.0000242712	0.0000745473	0.0000554771	0.0001317580

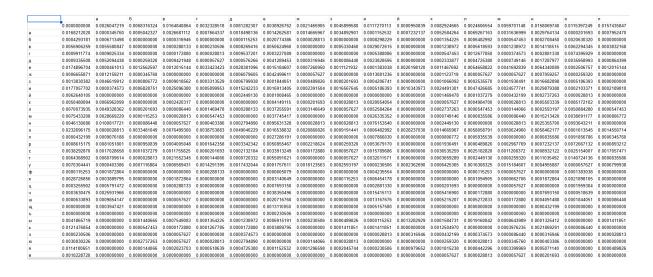
### Біграми, з пробілами (bigrams overlap matrix with spaces.csv):



#### Біграми, не перетинаються, без пробілів (bigrams\_nonoverlap\_matrix\_no\_spaces.csv):



Біграми, не перетинаються, з пробілами (bigrams\_nonoverlap\_matrix\_with\_spaces):



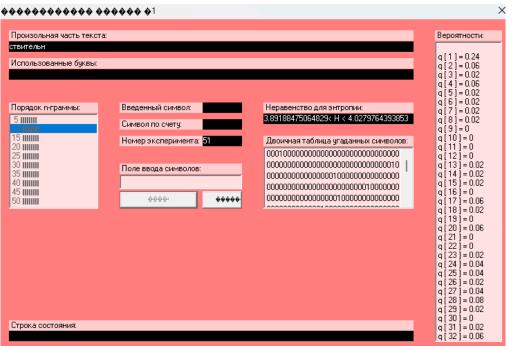
### Ентропія (entropy summary.csv):

variant	with_spaces	chars	alphabet_size	H1	R1	H2_overlap_per	R2_overlap	H2_nonoverlap_	R2_nonoverlap
with_spaces	TRUE	694125	34	4.3666326960	0.1416875499	3.9572732893	0.2221519031	3.9568901993	0.2222272039
no_spaces	FALSE	576816	33	4.4659764992	0.1146654299	4.1444044379	0.1784138313	4.1459743050	0.1781026211

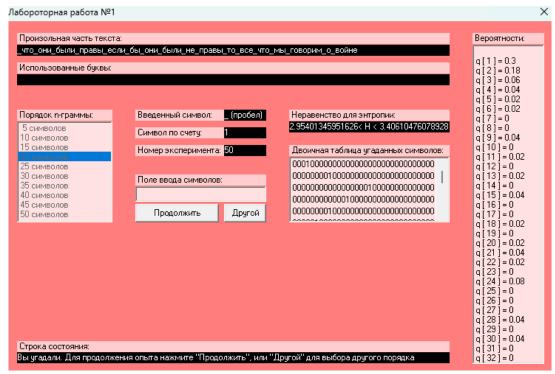
# Завдання 2

За допомогою програми CoolPinkProgram оцінити значення  $\boldsymbol{H}^{10}$ ,  $\boldsymbol{H}^{20}$ ,  $\boldsymbol{H}^{30}$ 

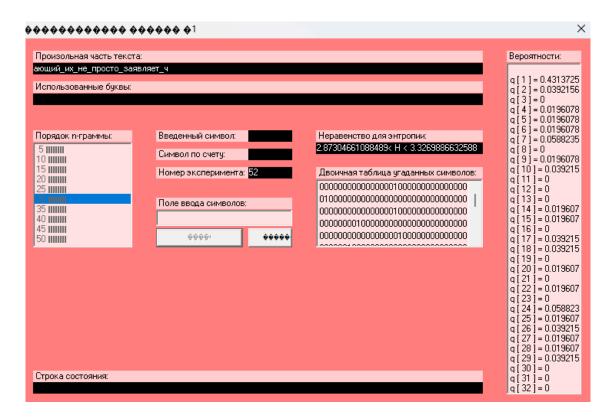
 $H^{10}$ : 3.8918 <  $H^{10}$  < 4.0279



 $H^{20}$ : 2.9540 <  $H^{20}$  < 3.4061



 $H^{30}$ : 2.8730 <  $H^{30}$  < 3.3269



## Завдання 3

Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

$$R = 1 - \frac{H_{\infty}}{H_0}$$

```
H^{10}: 

Нижня межа: R=1-\frac{3.8918}{5}\approx 1-0,78\approx 0,22 

Верхня межа: R=1-\frac{4.0279}{5}\approx 1-0,81\approx 0,19 

H^{20}: 

Нижня межа: R=1-\frac{2.9540}{5}\approx 1-0,59\approx 0,4 

Верхня межа: R=1-\frac{3.4061}{5}\approx 1-0,68\approx 0,19 

H^{30}: 

Нижня межа: R=1-\frac{2.8730}{5}\approx 1-0,57\approx 0,43 

Верхня межа: R=1-\frac{3.3269}{5}\approx 1-0,67\approx 0,33
```

## Код:

```
#!/usr/bin/env python3
import csv, math, re
from pathlib import Path
from collections import Counter
INPUT FILE = Path("lab1.txt")
OUT DIR = Path("out")
LETTERS = "абвгдеёжзийклмнопрстуфхцчшщъыьэюяіїєґ"
def read text(p: Path) -> str:
    return p.read_bytes().decode("utf-8", "ignore")
def normalize(text: str, with_spaces: bool) -> str:
    s = text.lower()
    allowed = LETTERS + (r"\s" if with_spaces else "")
    s = re.sub(f"[^{allowed}]", " " if with_spaces else "", s,
flags=re.IGNORECASE)
    return " ".join(s.split()) if with_spaces else s
def entropy(counts: Counter) -> float:
    n = sum(counts.values())
```

```
return 0.0 if n == 0 else -sum((c/n)*math.log2(c/n) for c in
counts.values() if c)
def redundancy(h: float, alphabet size: int) -> float:
    return 0.0 if alphabet_size <= 1 else max(0.0, min(1.0, 1 - h /</pre>
math.log2(alphabet_size)))
def save unigrams(path: Path, counts: Counter):
   total = sum(counts.values())
    rows = [(k, counts[k], (counts[k]/total if total else 0.0)) for k in
counts]
    rows.sort(key=lambda x: (-x[2], x[0]))
   with open(path, "w", newline="", encoding="utf-8") as f:
        w = csv.writer(f); w.writerow(["char", "count", "freq"])
        for ch, cnt, freq in rows:
            w.writerow([repr(ch).strip("'"), cnt, f"{freq:.10f}"])
def save_matrix(path: Path, freqs: dict[str, float], alphabet: list[str]):
   with open(path, "w", newline="", encoding="utf-8") as f:
        w = csv.writer(f)
        w.writerow([""] + [repr(c).strip("'") for c in alphabet])
        for a in alphabet:
            w.writerow([repr(a).strip("'")] + [f"{freqs.get(a+b, 0.0):.10f}"
for b in alphabet])
def process_variant(raw: str, name: str, with_spaces: bool) -> dict:
    text = normalize(raw, with spaces)
    alphabet = ([" "] + sorted({c for c in text if c != " "})) if (with_spaces
and " " in text) else sorted(set(text))
   uni = Counter(text)
    save_unigrams(OUT_DIR / f"unigrams_{name}.csv", uni)
   h1 = entropy(uni)
   r1 = redundancy(h1, len(alphabet))
    bi_ov = Counter(text[i:i+2] for i in range(len(text)-1))
    total_ov = sum(bi_ov.values())
    freqs ov = {k: v/total_ov for k, v in bi_ov.items()} if total_ov else {}
    save_matrix(OUT_DIR / f"bigrams_overlap_matrix_{name}.csv", freqs_ov,
alphabet)
   h2_ov_per_char = (entropy(bi_ov) / 2.0)
    r2_ov = redundancy(h2_ov_per_char, len(alphabet))
```

```
bi_no = Counter(text[i:i+2] for i in range(0, len(text)-1, 2))
    total_no = sum(bi_no.values())
    freqs no = {k: v/total no for k, v in bi no.items()} if total no else {}
    save_matrix(OUT_DIR / f"bigrams_nonoverlap_matrix_{name}.csv", freqs_no,
alphabet)
   h2_no_per_char = (entropy(bi_no) / 2.0)
    r2_no = redundancy(h2_no_per_char, len(alphabet))
   return {
        "variant": name,
        "with_spaces": with_spaces,
        "chars": len(text),
        "alphabet_size": len(alphabet),
        "H1": f"{h1:.10f}",
        "R1": f"{r1:.10f}",
        "H2_overlap_per_char": f"{h2_ov_per_char:.10f}",
        "R2_overlap": f"{r2_ov:.10f}",
        "H2 nonoverlap_per_char": f"{h2_no_per_char:.10f}",
        "R2_nonoverlap": f"{r2_no:.10f}",
def main():
    raw = read_text(INPUT_FILE)
   OUT_DIR.mkdir(parents=True, exist_ok=True)
    summaries = []
    for name, s in (("with_spaces", True), ("no_spaces", False)):
        summaries.append(process_variant(raw, name, s))
    if summaries:
        with open(OUT_DIR / "entropy_summary.csv", "w", newline="",
encoding="utf-8") as f:
            keys = list(summaries[0].keys())
            w = csv.DictWriter(f, fieldnames=keys); w.writeheader();
w.writerows(summaries)
if __name__ == "__main__":
    main()
```