

Міністерство освіти і науки України НТУУ «Київський політехнічний інститут» Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту.

Виконали:

студенти 3 курсу ФТІ групи ФБ-72

Башкиров I. О. Давидюк П. П.

Перевірив:

Чорний О. М.

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. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.
- 3. За допомогою програми CoolPinkProgram оцінити значення (10) Н, (20) Н, (30) Н.
- 4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Результати:

Монограми:

3 пробілом

'':0.15744762916879637 o:0.0933214293902421 a:0.07242164773409253 e:0.06865626150553412 и:0.05818546684990108 н:0.05420564780961603 T:0.05085698949818676 л:0.04358785204405718 c:0.04282317654563575 p:0.040193340175403346 в:0.03970108875225727 ĸ:0.030495312822407823 м:0.02526126412862732 y:0.024944335130163413 п:0.02403265852181618 д:0.023575471583606546 я:0.016104039024074465 г:0.015905789820780024 ь:0.015177527441331048 з:0.015010296820864986 ы:0.014670441043788798 ч:0.013274604816511597 $6 \hspace{-0.05cm}:\hspace{-0.05cm} 0.012995437571056156$ й:0.009900322458398012 ж:0.0075267265550087595 ш:0.007459294853207929 x:0.006984575672530078 ю:0.00449364860800738 ш:0.0031719872527110914 ц:0.0028671959605713355 э:0.0026042123235480944 ф:0.0021443281172664266

 $H_1 = 4.379945305657554$

Без пробіла

o:0.1107603902392176 a:0.08595506966842471 e:0.08148604630689321 и:0.06905857589896677 н:0.06433504869986954 т:0.06036062713587143 л:0.05173310710769994 c:0.05082553681902216 p:0.04770426333944249 в:0.047120024970187836 к:0.03619396713859255 м:0.02998183258769578 y:0.02960567911707977 п:0.028523637644158818 д:0.02798101625463189 я:0.019113398266492727 г:0.018878102265724415 b:0.018013749609840816 з:0.017815268629600877 ы:0.017411904056855197 ч:0.015755228133078295 6:0.015423892948322917 й:0.011750394160817614 ж:0.008933244763863656 ш:0.008853212110541102 x:0.00828978223115031 ю:0.005333376017415106 щ:0.0037647360122930154 ц:0.0034029884192750643 э:0.0030908610713170975 $\varphi \mathpunct{:} 0.002545038375657268$

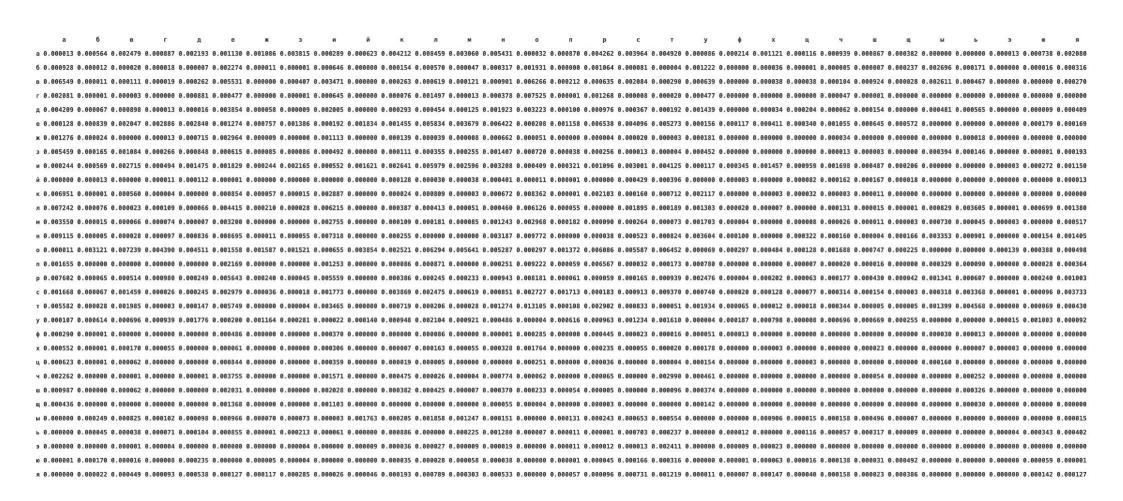
 $H_1 = 4.452870917974479$

Біграми:

Перетинаються, з пробілом – H_2 = 3.991013097121348 Не перетинаються, з пробілом – H_2 = 3.990569680311937 Перетинаються, без пробіла – H_2 = 4,148671405718143 Не перетинаються, без пробіла – H_2 = 4.147853620317564

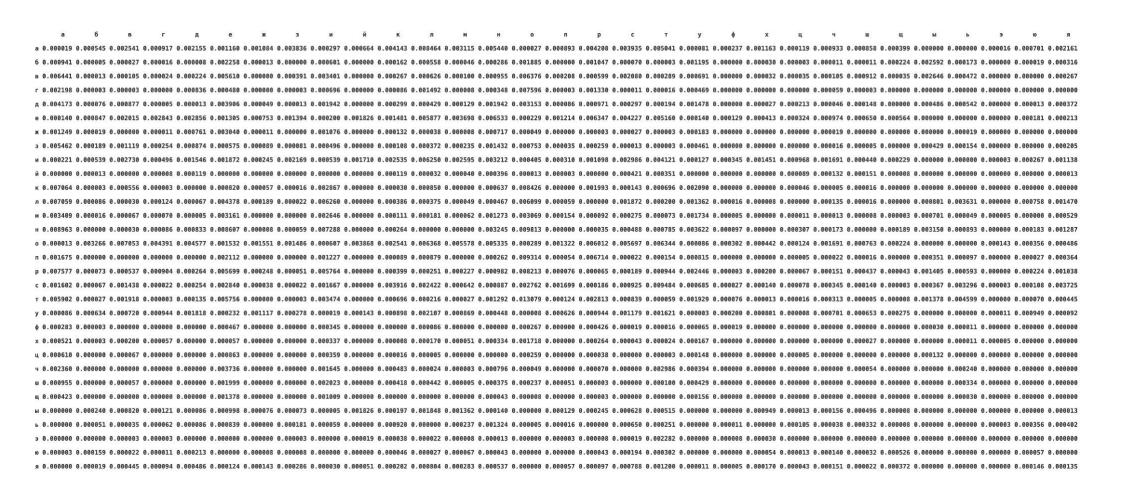
Перетинаються, з пробілом

 $H_2 = 3.991013097121348$



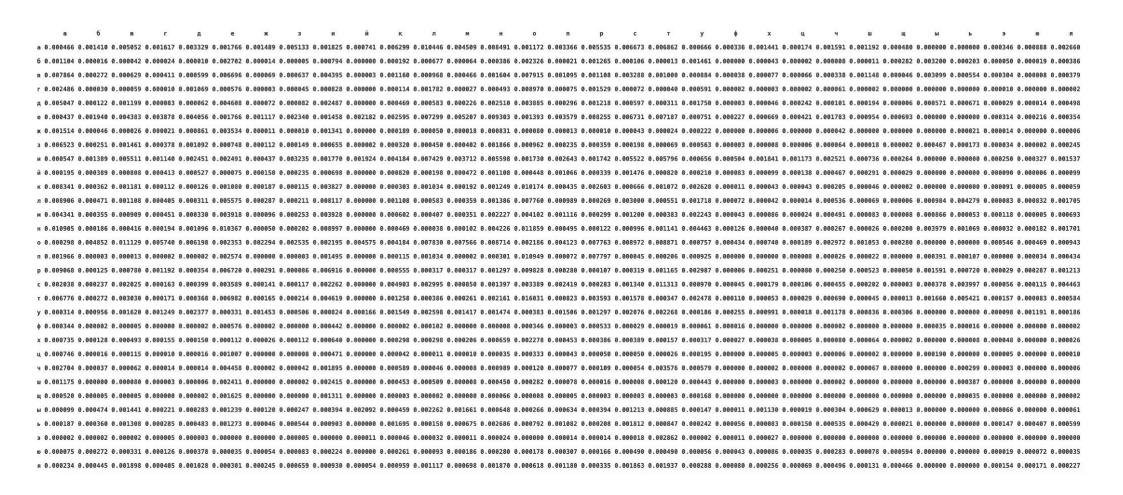
Не перетинаються, з пробілом

 $H_2 = 3.990569680311937$



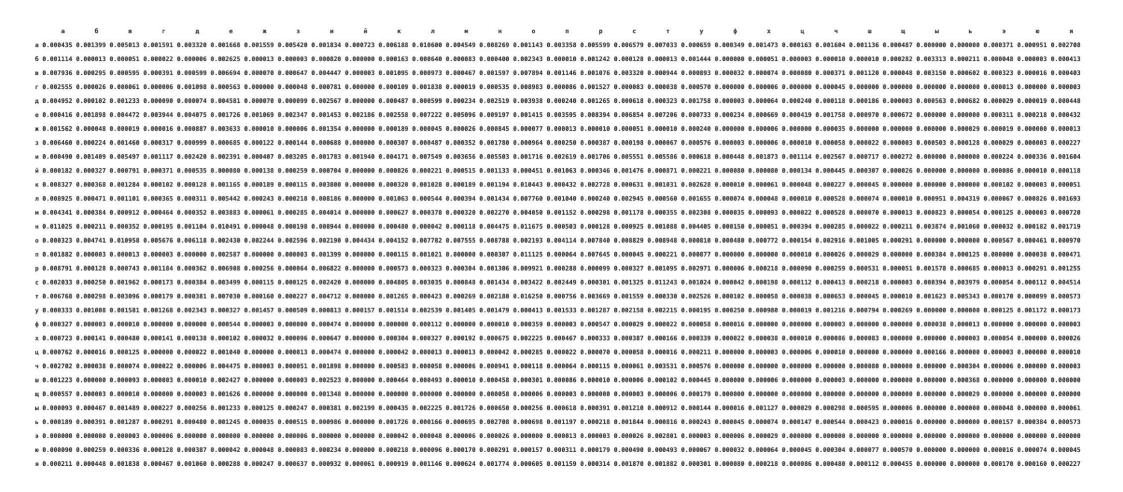
Перетинаються, без пробіла

 $H_2 = 4,148671405718143$



Не перетинаються, без пробіла

 $H_2 = 4.147853620317564$



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

$$H0 = log_2 32 = 5$$

Без пробіла, Н :	R =1-	4,452870917974479	= 0,109425816
1	1	5	
3 пробілом, Н :	R =1-	4,379945305657554	= 0,124010939
1	1	5	
Перетинаються, з пробілом, Н :	R =1-	3,991013097121348	= 0,201797381
2	2	5	
Не перетинаються, з пробілом, Н:	R =1-	3,990569680311937	= 0,201886064
2	2	5	
Перетинаються, без пробіла, Н :	R =1-	4,148671405718143	= 0,170265719
2	2	5	
Не перетинаються, без пробіла, Н	R =1-	4,147853620317564	= 0,170429276
2	2	5	

Код програми

lab1_symbol_enthropy.py:

```
import math
#symbol enthropy without space
f = open('text.txt','r',encoding = 'utf-8')
mytext = f.read().replace('-','').replace('?','').replace(''','').replace('\n',
").lower().replace('!',").replace('(',").replace(')',").replace('.',").replace(',",").replace(';',")
f.close()
lenght = len(mytext)
frequency_dictionary = { } #dictionary of symbols(=keys) and their counts
for symbol in mytext:
  count = frequency_dictionary.get(symbol,0)
 frequency_dictionary[symbol] = count + 1
frequency_symbols = frequency_dictionary.keys() #symbols(=keys), набор ключей словаря
enthropy = 0
for symbols in frequency symbols:
  print (symbols, frequency_dictionary[symbols]/lenght)
  enthropy -= (frequency_dictionary[symbols]/lenght)*math.log(frequency_dictionary[symbols]/lenght,2)
print (enthropy)
print('-----')
#symbol enthropy with space
f2 = open('text.txt','r',encoding = 'utf-8')
mytext2 = f2.read().replace('-',").replace('?',").replace('"",").replace('\n',
").lower().replace('!',").replace('(',").replace(')',").replace('.',").replace(',",").replace(';',")
f2.close()
lenght2 = len(mytext2)
frequency_dictionary2 = { }
for symbol in mytext2:
  count2 = frequency_dictionary2.get(symbol,0)
 frequency_dictionary2[symbol] = count2 + 1
frequency_symbols2 = frequency_dictionary2.keys()
enthropy2 = 0
for symbols in frequency_symbols2:
  print (symbols, frequency_dictionary2[symbols]/lenght2)
  enthropy2 -= (frequency_dictionary2[symbols]/lenght2)*math.log(frequency_dictionary2[symbols]/
lenght2,2)
print (enthropy2)
lab1_bigrams without spaces.py:
import re
import math
frequency = {}
f = open('text.txt', 'r',encoding = 'utf-8')
mytext = f.read().replace('?','').replace(''','').replace('\n',
").lower().replace('!',").replace('(',").replace(',",").replace(',",").replace(',",").replace(',",").replace(';',")
f.close()
match_pattern = re.findall('\w\w', mytext)
```

```
for bigram in match_pattern:
  count = frequency.get(bigram,0)
 frequency[bigram] = count + 1
frequency_list = frequency.keys()
lenght = len(mytext)
enthropy = 0
for bigrams in frequency_list:
  print (bigrams, frequency[bigrams]/lenght)
 enthropy -= (frequency[bigrams]/lenght)*math.log(frequency[bigrams]/lenght,2)
print (enthropy)
print('-----')
lab1_bigrams with spaces.py:
import re
import math
frequency = {}
f = open('text.txt', 'r',encoding = 'utf-8')
mytext = f.read().replace('?','').replace('"','').replace('-', ").replace('\n',
").lower().replace('!',").replace('(',").replace(')',").replace('.',").replace(',",").replace(';',")
match_pattern = re.findall('\w.|.\w', mytext)
for bigram in match_pattern:
 count = frequency.get(bigram,0)
 frequency[bigram] = count + 1
frequency_list = frequency.keys()
lenght = len(mytext)
enthropy = 0
for bigrams in frequency_list:
  print (bigrams, frequency[bigrams]/lenght)
  enthropy -= (frequency[bigrams]/lenght)*math.log(frequency[bigrams]/lenght,2)
print (enthropy)
print('-----')
lab1_alternation bigrams without spaces.py
import re
import math
f = open('text.txt', 'r',encoding = 'utf-8')
mytext = f.read().replace('-','').replace('?','').replace(''','').replace('''','').replace('\n',
").lower().replace('!',").replace('(',").replace(')',").replace('.',").replace(',",").replace(';',")
f.close()
output = open("text_output.txt", "r+")
for i in range(1,len(mytext)):
  output.write(mytext[i-1])
  output.write(mytext[i])
```

```
output.close()
frequency = {}
f1 = open('text_output.txt', 'r')
mytext = f1.read()
f1.close()
match_pattern = re.findall('\w\w', mytext)
for bigram in match_pattern:
  count = frequency.get(bigram,0)
  frequency[bigram] = count + 1
frequency_list = frequency.keys()
lenght = len(mytext)
enthropy = 0
for bigrams in frequency_list:
  print (bigrams, frequency[bigrams]/lenght)
  enthropy -= (frequency[bigrams]/lenght)*math.log(frequency[bigrams]/lenght,2)
print (enthropy)
print('-----')
lab1_alternation bigrams with spaces.py
import re
import math
f = open('text.txt', 'r',encoding = 'utf-8')
mytext = f.read().replace('-','').replace('?','').replace('''','').replace('\n',
").lower().replace('!',").replace('(',").replace(',",").replace(',",").replace(',",").replace(',",").replace(';',")
f.close()
output = open("text_output2.txt", "r+")
for i in range(1,len(mytext)):
  output.write(mytext[i-1])
  output.write(mytext[i])
output.close()
frequency = {}
f1 = open('text_output2.txt', 'r')
mytext = f1.read()
f1.close()
match_pattern = re.findall('\w.|.\w', mytext)
for bigram in match_pattern:
  count = frequency.get(bigram,0)
  frequency[bigram] = count + 1
frequency_list = frequency.keys()
lenght = len(mytext)
enthropy = 0
for bigrams in frequency_list:
  print (bigrams, frequency[bigrams]/lenght)
```

епштору -=	(frequency[bigrams]/lenght)*math.log(frequency[bigrams]/lenght,2)	
print (enthrop	у)	
print('	')	