Лабораторна робота з криптографії №1

Виконав: Костюковець Остап ФБ-96

Варіант №5

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Мета: 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.

Хід роботи

Завдання 1

> Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H 1 та H 2 за безпосереднім означенням. Підрахувати частоти букв та біграм, атакож значення H 1 та H 2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H 1 та H 2 на тому ж тексті, в якому вилучено всі пробіли.

In [1]:
```
ru_RU = 'абвгдежзийклмнопрстуфхцчшщыьэюя'
text = open("text.txt").read().lower().replace("ъ","ь").replace("ё","е")

filtered_whitespaces = ' '.join(''.join([i if i in ru_RU else ' ' for i in text]).split(
filtered_no_whitespaces = filtered_whitespaces.replace(' ','')
```

Частоти букв, без пробілів

In [2]:
```
from collections import Counter
letter_frequency = Counter(filtered_no_whitespaces)
for i in ru_RU:
    letter_frequency[i] /= len(filtered_no_whitespaces)
print(sorted(letter_frequency.items(),
      key=lambda item: item[1], reverse=True))
```

```
[('о', 0.10584571020317421), ('е', 0.08401953822278149), ('а', 0.08243308288203895),
('н', 0.0648309831490383), ('и', 0.062441587146042386), ('т', 0.06075800188647887),
('л', 0.05252786009838181), ('р', 0.046900800134686824), ('с', 0.046194989391254423),
('в', 0.04187378722504139), ('к', 0.03777706309343682), ('д', 0.03151542099344481),
('у', 0.030980127218609228), ('м', 0.029898747455735737), ('п', 0.027291348746052747),
('ь', 0.022348515099385492), ('я', 0.022279444934890577), ('ы', 0.01967636311048852),
('г', 0.019454043518520515), ('з', 0.017880538833620768), ('б', 0.01760641661828158),
('ч', 0.015430706436691798), ('й', 0.011353408288851428), ('ж', 0.010215909017325818),
('ш', 0.009704358111535366), ('х', 0.008452461380065055), ('ю', 0.007101276287133308),
('щ', 0.004278033313403713), ('э', 0.004193854050425537), ('ц', 0.0033563783059247093),
('ф', 0.0013792448472578066)]
```

Частоти букв, з пробілами

In [3]:
```
letter_frequency_space = Counter(filtered_whitespaces)
for i in ru_RU+' ':
    letter_frequency_space[i] /= len(filtered_whitespaces)
```

```
print(sorted(letter_frequency_space.items(),
      key=lambda item: item[1], reverse=True))
```

```
[(' ', 0.16066190923930077), ('о', 0.0888403363171425), ('е', 0.07052079879850502),
('а', 0.06918922640172905), ('н', 0.05441511361845287), ('и', 0.05240960253922704),
('т', 0.05099650530183213), ('л', 0.04408863380672091), ('р', 0.03936562804019718),
('с', 0.03877321419836624), ('в', 0.03514626462238599), ('к', 0.03170772801139174),
('д', 0.026452093286157625), ('у', 0.026002800831191043), ('м', 0.025095157605633547),
('п', 0.022906668550796316), ('ь', 0.018757959994854877), ('я', 0.01869998677485919),
('ы', 0.01651512104627169), ('г', 0.016328519744410568), ('з', 0.015007817326383793),
('б', 0.014777736109525905), ('ч', 0.012951579679661726), ('й', 0.009529348036791254),
('ж', 0.008574601569987263), ('ш', 0.008145237409394199), ('х', 0.0070944727969723485),
('ю', 0.005960371680806697), ('щ', 0.0035907163134829404), ('э', 0.0035200614516131953),
('ц', 0.0028171361591654755), ('ф', 0.0011576527367888995)]
```

Частоти біграм, без пробілів

In [4]:

```
bigram_frequency = {}
for i,val in enumerate(filtered_no_whitespaces):
    if i +1 >= len(filtered_no_whitespaces):
        break
    temp = filtered_no_whitespaces[i] + filtered_no_whitespaces[i+1]
    if temp not in bigram_frequency:
        bigram_frequency[temp] = filtered_no_whitespaces.count(temp) / (len(filtered_no_

print(sorted(bigram_frequency.items(),
      key=lambda item: item[1], reverse=True)[:10])
```

```
[('то', 0.01430399937836852), ('на', 0.01231391526385882), ('не', 0.011586520094021762),
('ст', 0.010900135334353557), ('по', 0.010889343121151227), ('ал', 0.01087423402266796
5), ('но', 0.010606587135250175), ('ко', 0.010500823445867337), ('ра', 0.010036758278167
137), ('ен', 0.009972004998953156)]
```

Частоти біграм, з пробілами

In [5]:

```
bigram_frequency_space = {}
for i,val in enumerate(filtered_whitespaces):
    if i +1 >= len(filtered_whitespaces):
        break
    temp = filtered_whitespaces[i] + filtered_whitespaces[i+1]
    if temp not in bigram_frequency_space:
        bigram_frequency_space[temp] = filtered_whitespaces.count(temp) / (len(filtered_

print(sorted(bigram_frequency_space.items(),
      key=lambda item: item[1], reverse=True)[:10])
```

```
[('о ', 0.019442768656053943), ('а ', 0.017920971631167126), ('е ', 0.01676875388375282
5), (' н', 0.016400986269405178), ('и ', 0.0158955322575677), (' п', 0.0158339357113223
52), (' в', 0.013942559408963022), (' с', 0.013069337782777968), ('я ', 0.01190443839348
9607), ('то', 0.01160551397788684)]
```

Ентропія

In [6]:

```
import math


def entropy(dict_, n):
    sum_ = 0
    for p in dict_.values():
        sum_ += p * math.log2(p)
    return 1/n * (-sum_)
```

```
def redundancy(H, n):
    return 1 - (H/(math.log2(n)))


lfe = entropy(letter_frequency, 1)
bfe = entropy(bigram_frequency, 2)
lfse = entropy(letter_frequency_space, 1)
bfse = entropy(bigram_frequency_space, 2)

print("No spaces")
print(f"H1 entropy : {lfe}")
print(f"H2 entropy : {bfe}")
print("With spaces")
print(f"H1 entropy : {lfse}")
print(f"H2 entropy : {bfse}")
```

```
No spaces
H1 entropy : 4.490728757018316
H2 entropy : 4.179363219527201
With spaces
H1 entropy : 4.405130403820585
H2 entropy : 4.008402242541684
```

Завдання 2

> За допомогою програми CoolPinkProgram оцінити значення H (10) , H (20) , H (30)

- H10: 3.1173 < R < 3.6639
- H20: 2.375 < R < 2.7555
- H30: 2.56079 < R < 3.0895

Завдання 3

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

In [14]:

```
def redundancy(H, n):
    return 1 - (H/(math.log2(n)))


lfr = redundancy(lfe, len(ru_RU))
bfr = redundancy(bfe, len(ru_RU))
lfsr = redundancy(lfse, len(ru_RU) + 1)
bfsr = redundancy(bfse, len(ru_RU) + 1)

print("No spaces")
print(f"H1 redundancy : {lfr}")
print(f"H2 redundancy : {bfr}")
print("With spaces")
print(f"H1 redundancy : {lfsr}")
print(f"H2 redundancy : {bfsr}")
```

```
No spaces
H1 redundancy : 0.0935505023078842
H2 redundancy : 0.1563993516436103
With spaces
H1 redundancy : 0.11897391923588307
H2 redundancy : 0.19831955149166325
```

Висновок

> Під час виконання даної лабораторної роботи я навчився вимірювати частоти символів та біграм у тексті, визначати ентропію. Також, я навчився визначати надлишковість мови.