Eksploracja danych internetowych

Laboratorium 4

Prowadzący: pracownik UR
Wykonał: Piotr Rojek, pr125159

Zadanie 1

Powtórz kroki wykonane na wykładzie związane z tworzeniem i weryfikacją nowych cech:

a) **Stwórz nowe cechy - długość wiadomości oraz procentowy udział znaków interpunkcyjnych.**
b) **Ustal, które cechy należy poddać transformacji.**
c) **Wskaż, która transformacja Boxa Coxa będzie najlepsza - odpowiedź uzasadnij.**

```python
import pandas as pd
import numpy as np
import string
import plotly.graph_objects as go

from matplotlib import pyplot
from plotly.subplots import make_subplots

data = pd.read_csv( filepath_or_buffer: "SMSSpamCollection.tsv", sep='\t', header=None)
data.columns = ['label', 'body_text']

pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', 35)

# Długość wiadomości oraz procentowy udział znaków interpunkcyjnych

def count_punctuation(text):  1 usage
    count = sum([1 for char in text if char in string.punctuation])
    return round(count / (len(text) - text.count(" ")), 3) * 100

data['body_len'] = data['body_text'].apply(lambda x: len(x) - x.count(" "))
data['punctuation_%'] = data['body_text'].apply(lambda x: count_punctuation(x))

print("Długość wiadomości oraz procentowy udział znaków interpunkcyjnych:")
print(data.head(), "\n")
```

```python
# Ewaluacja długości wiadomości
bins1 = np.linspace( start: 0,  stop: 200,  num: 50)
pyplot.hist(data[data['label'] == 'spam']['body_len'], bins1, alpha=0.5, density=True, label='spam')
pyplot.hist(data[data['label'] == 'ham']['body_len'], bins1, alpha=0.5, density=True, label='ham')
pyplot.legend(loc='upper left')
pyplot.show()

# Ewaluacja procentowej zawartości znaków interpunkcyjnych
bins2 = np.linspace( start: 0,  stop: 50,  num: 50)
pyplot.hist(data[data['label'] == 'spam']['punctuation_%'], bins2, alpha=0.5, density=True, label='spam')
pyplot.hist(data[data['label'] == 'ham']['punctuation_%'], bins2, alpha=0.5, density=True, label='ham')
pyplot.legend(loc='upper right')
pyplot.show()
```

```python
# Ustalenie, które cechy należy poddać transformacji

pyplot.hist(data['body_len'], bins1)
pyplot.title("Body length distribution")
pyplot.show()
# Rozkład długości wiadomości jest względnie normalny, nie ma dużo przypadków odstających.
# Nie ma potrzeby transformacji.

pyplot.hist(data['punctuation_%'], bins2)
pyplot.title("Punctuation % distribution")
pyplot.show()
# Rozkład procentowej zawartości znaków interpunkcyjnych jest silnie skośny, ma większą ilość przypadków odstających.
# Warto poddać tę cechę transformacji.

print("Które cechy należy poddać transformacji:")
print("Transformacji należy poddać cechę z procentową zawartością znaków interpunkcyjnych.\n")
```

```python
# Wskazanie, która transformacja Boxa Coxa będzie najlepsza

def box_cox_transform(x, lambd):  7 usages
    if lambd == 0:
        return np.log(x)
    else:
        return (np.power(x, lambd) - 1) / lambd

data['punctuation_%_safe'] = data['punctuation_%'] + 1

box_cox_minus_2 = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x, -2))
box_cox_minus_1 = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x, -1))
box_cox_minus_pol = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x, -0.5))
box_cox_0 = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x,  lambd: 0))
box_cox_pol = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x,  lambd: 0.5))
box_cox_1 = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x,  lambd: 1))
box_cox_2 = data['punctuation_%_safe'].apply(lambda x: box_cox_transform(x,  lambd: 2))

fig = make_subplots(rows=3, cols=3)
fig.add_trace(go.Histogram(x=box_cox_minus_2, name='lambda = -2'), row=1, col=1)
fig.add_trace(go.Histogram(x=box_cox_minus_1, name='lambda = -1'), row=1, col=2)
fig.add_trace(go.Histogram(x=box_cox_minus_pol, name='lambda = -0.5'), row=1, col=3)
fig.add_trace(go.Histogram(x=box_cox_0, name='lambda = 0'), row=2, col=2)
fig.add_trace(go.Histogram(x=box_cox_pol, name='lambda = 0.5'), row=3, col=1)
fig.add_trace(go.Histogram(x=box_cox_1, name='lambda = 1'), row=3, col=2)
fig.add_trace(go.Histogram(x=box_cox_2, name='lambda = 2'), row=3, col=3)
fig.show()

print("Która transformacja Boxa Coxa będzie najlepsza:")
print("Najlepsza transformacja Boxa Coxa to będzie z lambda = 0.5.")
print("Dzieje się tak, ponieważ rozkład ten jest najbardziej symetryczny i przypomina normalny.")
print("Pozostałe wartości lambda dają rozkłady silnie skośne lub z dużą koncentracją przy jednym końcu.")
```

Testy:

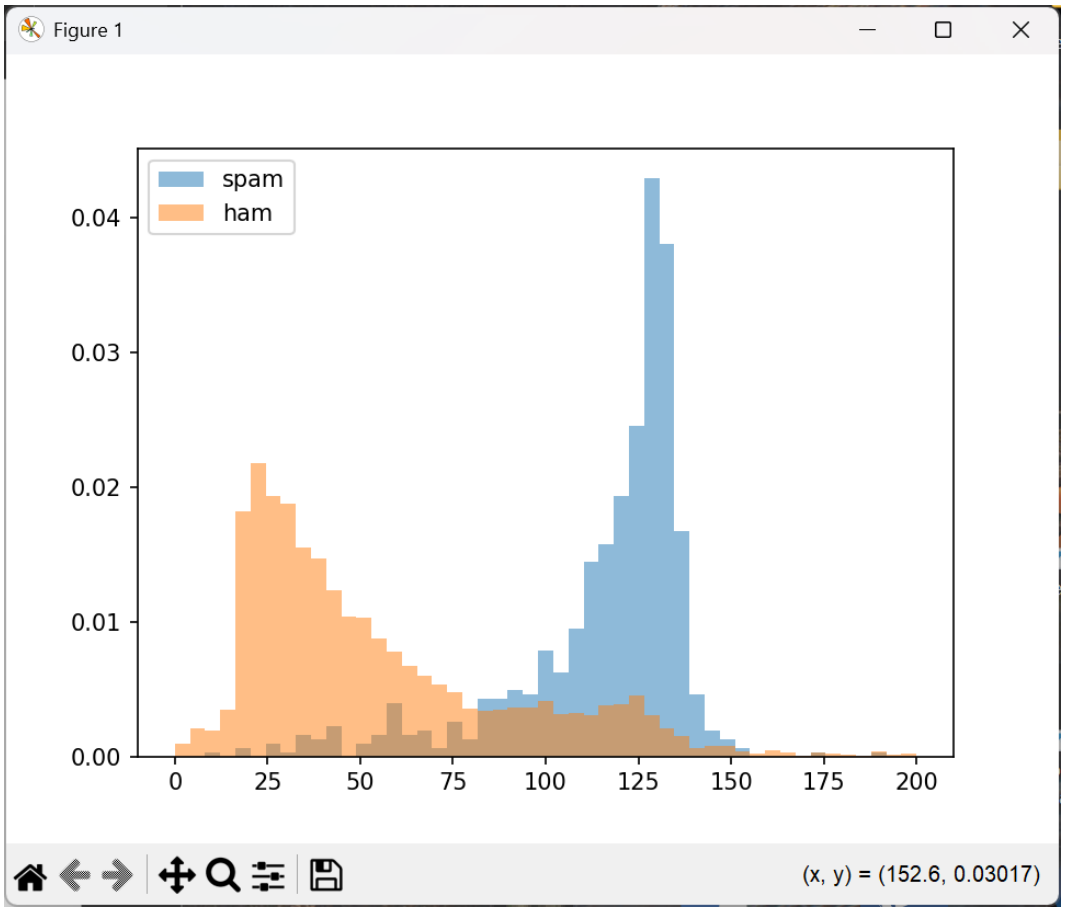


Figure 1
spam
ham
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
125
150
175
200
(x, y) = (152.6, 0.03017)

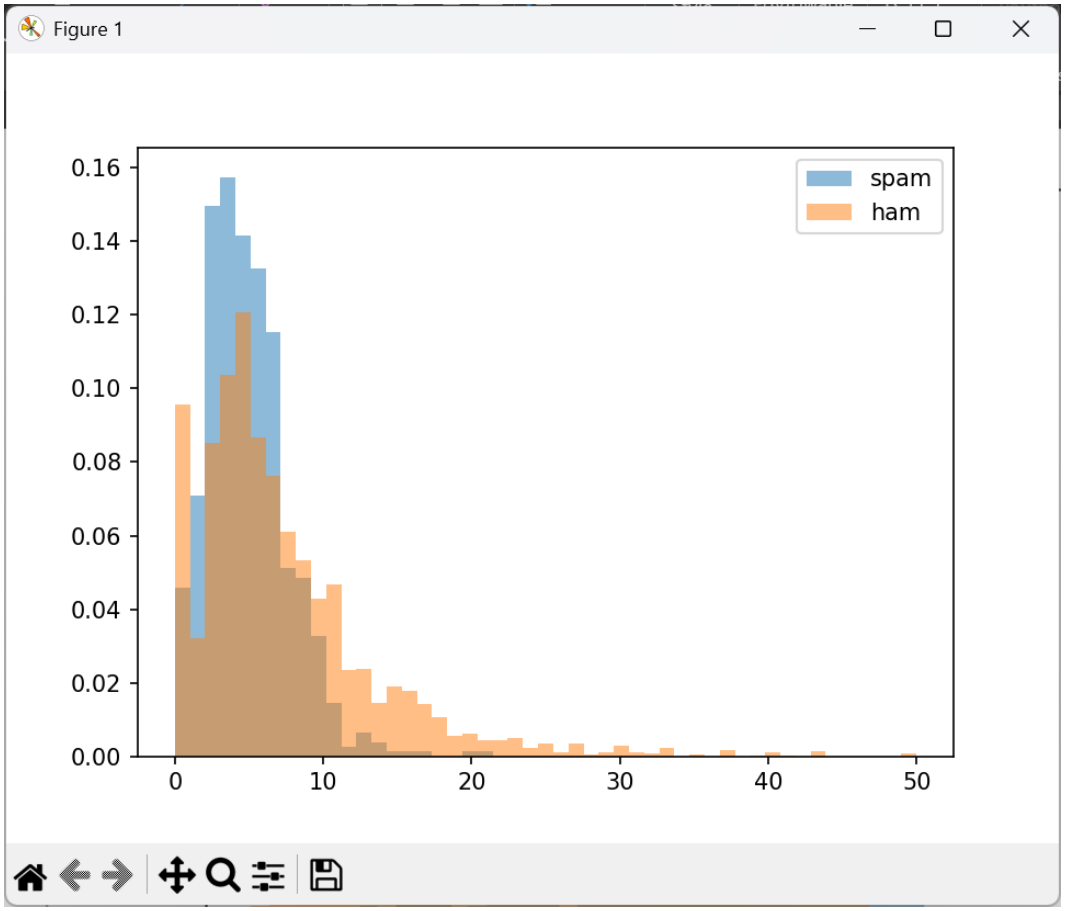


Figure 1
spam
ham
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50

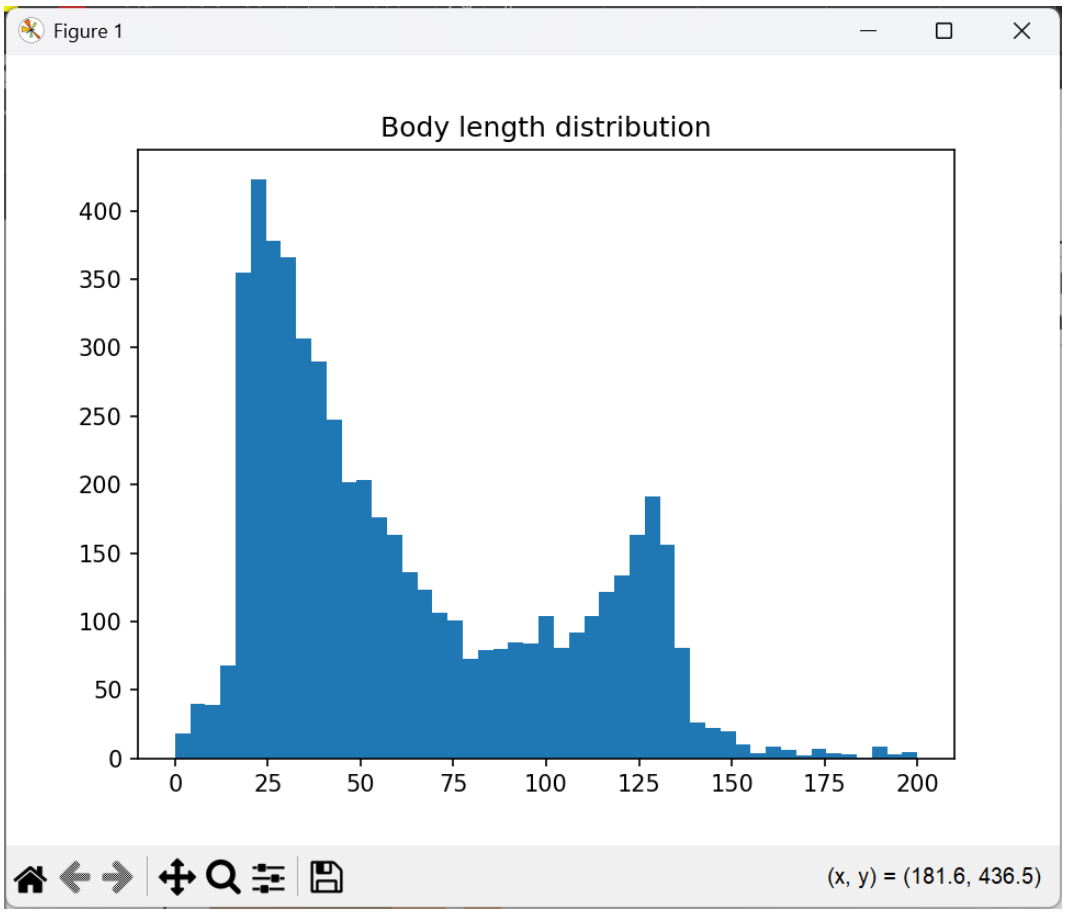


Figure 1
Body length distribution
400
350
300
250
200
150
100
50
0
0
25
50
75
100
125
150
175
200
(x, y) = (181.6, 436.5)

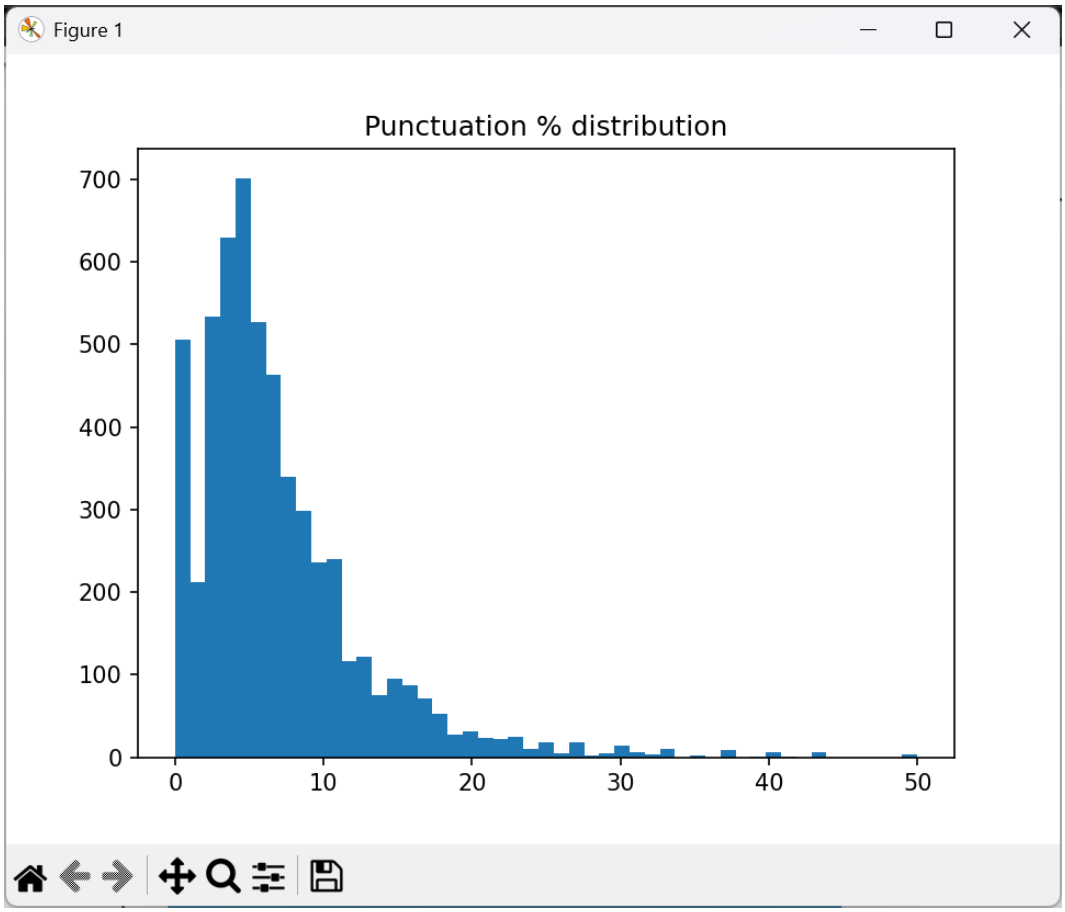
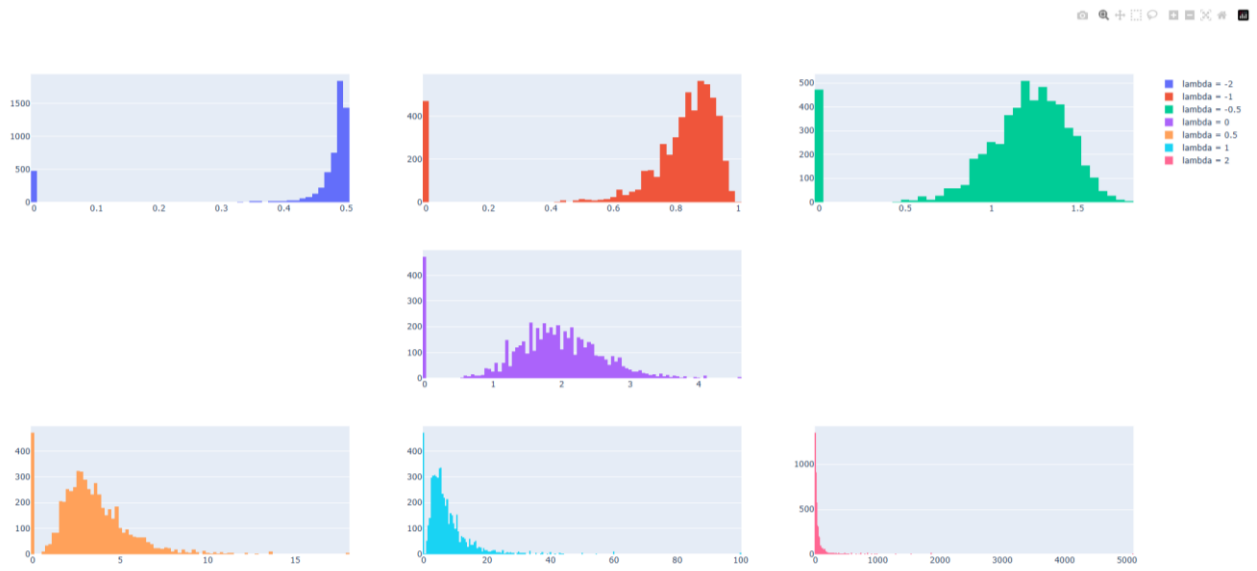


Figure 1
Punctuation % distribution
700
600
500
400
300
200
100
0
0
10
20
30
40
50



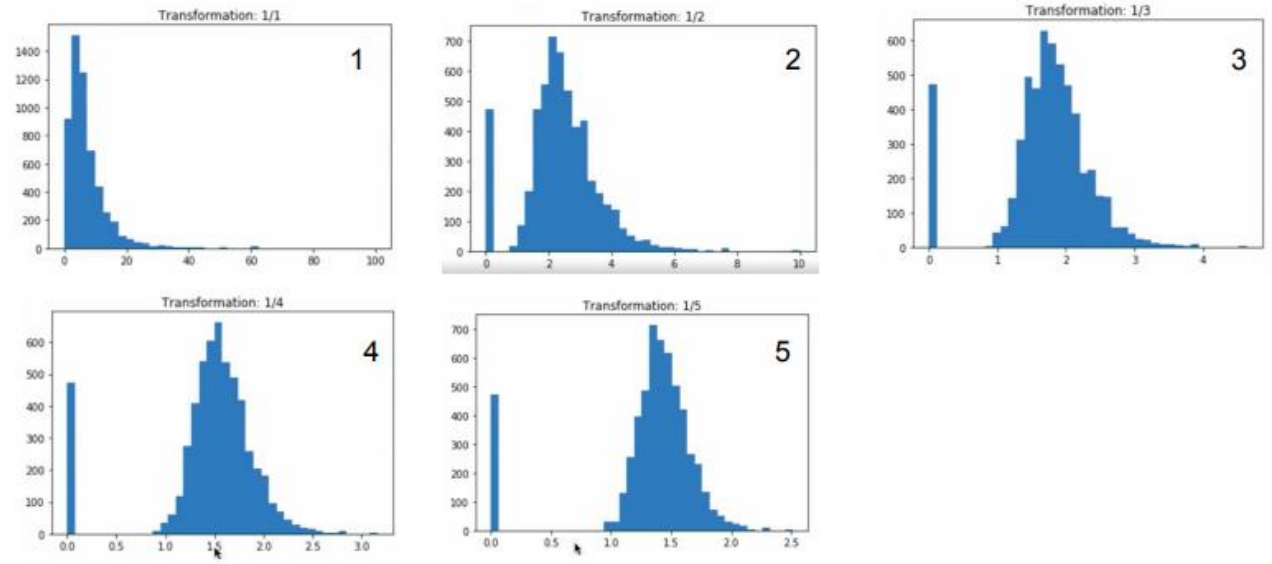

```
Długość wiadomości oraz procentowy udział znaków interpunkcyjnych:
  label                          body_text  body_len  punctuation_%
0   ham  Go until jurong point, crazy.. ...        92            9.8
1   ham      Ok lar... Joking wif u oni...        24           25.0
2  spam  Free entry in 2 a wkly comp to ...       128            4.7
3   ham  U dun say so early hor... U c a...        39           15.4
4   ham  Nah I don't think he goes to us...        49            4.1

Które cechy należy poddać transformacji:
Transformacji należy poddać cechę z procentową zawartością znaków interpunkcyjnych.

Która transformacja Boxa Coxa będzie najlepsza:
Najlepsza transformacja Boxa Coxa to będzie z lambda = 0.5.
Dzieje się tak, ponieważ rozkład ten jest najbardziej symetryczny i przypomina normalny.
Pozostałe wartości lambda dają rozkłady silnie skośne lub z dużą koncentracją przy jednym końcu.
```

Zadanie 2

Jak należy interpretować pojedynczy słupek występujący po lewej stronie wykresu na rysunkach 2-5 na slajdzie z wynikami serii przekształceń Boxa Coxa?



Pojedynczy słupek po lewej stronie wykresów 2–5 oznacza, że wiele wartości w danych było bliskich zeru. Po zastosowaniu przekształcenia Boxa-Coxa te wartości zostały przekształcone na wartości jeszcze mniejsze, co spowodowało ich „skumulowanie” w postaci słupku po lewej stronie danego wykresu. To może oznaczać, że dane należy lekko przesunąć (np.: +1), aby uniknąć takiego efektu i uzyskać bardziej równomierny rozkład.