3.8 Основы кибербезопасности с Python

Spam

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

data = pd.read_csv("spam.csv")
data["Spam"] = data["Category"].apply(lambda x: 1 if x == "spam" else 0)

vect = CountVectorizer()
X = vect.fit_transform(data["Message"])

model = Pipeline([("vect", CountVectorizer()), ("NB", MultinomialNB())])

X_train, X_test, y_train, y_test = train_test_split(
    data["Message"], data["Spam"], test_size=0.3
)

model.fit(X_train, y_train)

y_predict = model.predict(X_test)

print(accuracy_score(y_predict, y_test))
# 0.986244019138756

msg = [
    "Hi! How are yoy?",
    "Free subscription",
    "Win the lottery",
    "Call me this evening",
]
```

```
print(model.predict(msg))
# [0 1 0 0]
```

Phishing

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier

data = pd.read_csv("phishing.csv")

X = data.drop(columns="class")
Y = data["class"]

X_train, X_test, y_train, y_test = train_test_split(
    X, Y, test_size=0.3
)

dt_1 = DecisionTreeClassifier()
model_1 = dt_1.fit(X_train, y_train)

dt_predict = model_1.predict(X_test)
print(f"Decision Tree Accuracy: {accuracy_score(dt_predict, y_test)}")
# Decision Tree Accuracy: 0.9415134157371119

rf_2 = RandomForestClassifier()
model_2 = rf_2.fit(X_train, y_train)

rf_predict = model_2.predict(X_test)
print(f"Random Forest Accuracy: {accuracy_score(rf_predict, y_test)}")
# Random Forest Accuracy: 0.9632197769068436
```

Парсинг файла

```
html_content = html_content = """
<html>
<title>Data Science is Fun</title>
```

```
<body>
    <h1>Data Science is Fun</h1>
    <div id='paragraphs' class='text'>
        <p id='paragraph 0'>Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
            Paragraph 0 Paragraph 0 Paragraph 0 </p>
        <p id='paragraph 1'>Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
            Paragraph 1 Paragraph 1 Paragraph 1 </p>
        <p id='paragraph 2'>Here is a link to <a href='https://www.mail.ru'>Mail ru</a></p>
    </div>
    <div id='list' class='text'>
        <h2>Common Data Science Libraries</h2>
        <ul>
            <li>NumPy</li>
            <li>SciPy</li>
            <li>Pandas</li>
            <li>Scikit-Learn</li>
        </ul>
    </div>
    <div id='empty' class='empty'></div>
</body>

</html>
"""

from bs4 import BeautifulSoup as bs

soup = bs(html_content, "lxml")

title = soup.find("title")
print(title)
# <title>Data Science is Fun</title>

print(title.text)
```

```
# Data Science is Fun

pList = soup.body.find_all("p")
for i, p in enumerate(pList):
    print(p.text)
    print("-" * 10)
# Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0
# ----------
# Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#           Paragraph 1 Paragraph 1 Paragraph 1
# ----------
# Here is a link to Mail ru
# ----------

bullet_points = [bullet.text for bullet in soup.body.find_all("li")]
print(bullet_points)
# ['NumPy', 'SciPy', 'Pandas', 'Scikit-Learn']

p2 = soup.find(id="paragraph 2")
print(p2.text)
# Here is a link to Mail ru

divAll = soup.find_all("div")
print(divAll)
# [<div class="text" id="paragraphs">
# <p id="paragraph 0">Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#           Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 </p>
# <p id="paragraph 1">Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 </p>
# <p id="paragraph 2">Here is a link to <a href="https://www.mail.ru">Mail ru</a></p>
# </div>, <div class="text" id="list">
# <h2>Common Data Science Libraries</h2>
# <ul>
# <li>NumPy</li>
# <li>SciPy</li>
# <li>Pandas</li>
# <li>Scikit-Learn</li>
# </ul>
# </div>, <div class="empty" id="empty"></div>]

divClassText = soup.find_all("div", class_="text")

for div in divClassText:
    id = div.get("id")
    print(id)
    print(div.text)

# paragraphs
#
# Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0 Paragraph 0
#            Paragraph 0 Paragraph 0 Paragraph 0
# Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1 Paragraph 1
#            Paragraph 1 Paragraph 1 Paragraph 1
# Here is a link to Mail ru
#
# list
#
# Common Data Science Libraries
#
# NumPy
# SciPy
# Pandas
# Scikit-Learn

soup.find(id="paragraph 0").decompose()
soup.find(id="paragraph 1").decompose()

print(soup.find(id="paragraphs"))
# <div class="text" id="paragraphs">
#
#
# <p id="paragraph 2">Here is a link to <a href="https://www.mail.ru">Mail ru</a></p>
# </div>

new_p = soup.new_tag("p")
print(new_p)
# <p></p>
```