UNIVERSITÄT
ZU KÖLN

SPRACHVERARBEITUNG: ÜBUNG

SoSe 2024

Janis Pagel

Institut für Digital Humanities

2024-06-18

01

SOLUTION TO EXERCISE 07

Solution to Exercise 07

- https://lehre.idh.uni-koeln.de/site/assets/files/5151/solution07.pdf

02

TRAIN AND TEST SPLITTING IN SKLEARN

Train Test Split

Token features for spam and not-spam

```
import pandas as pd

data = {'casino': [1, 0, 1, 1, 0],
        'enlargement': [1, 1, 0, 1, 1],
        'meeting': [0, 0, 1, 1, 1],
        'profit': [0, 1, 0, 0, 0],
        'super': [1, 0, 1, 0, 0],
        'text': [1, 0, 0, 0, 1],
        'xxx': [1, 0, 0, 0, 1],
       'class': [0, 1, 0, 0, 1]}
df = pd.DataFrame(data)
print(df)
>    casino  enlargement  meeting  profit  super  text  xxx  class
> 0       1            1        0       0      1     1    1      0
> 1       0            1        0       1      0     0    0      1
> 2       1            0        1       0      1     0    0      0
> 3       1            1        1       0      0     0    0      0
> 4       0            1        1       0      0     1    1      1
```

Train Test Split

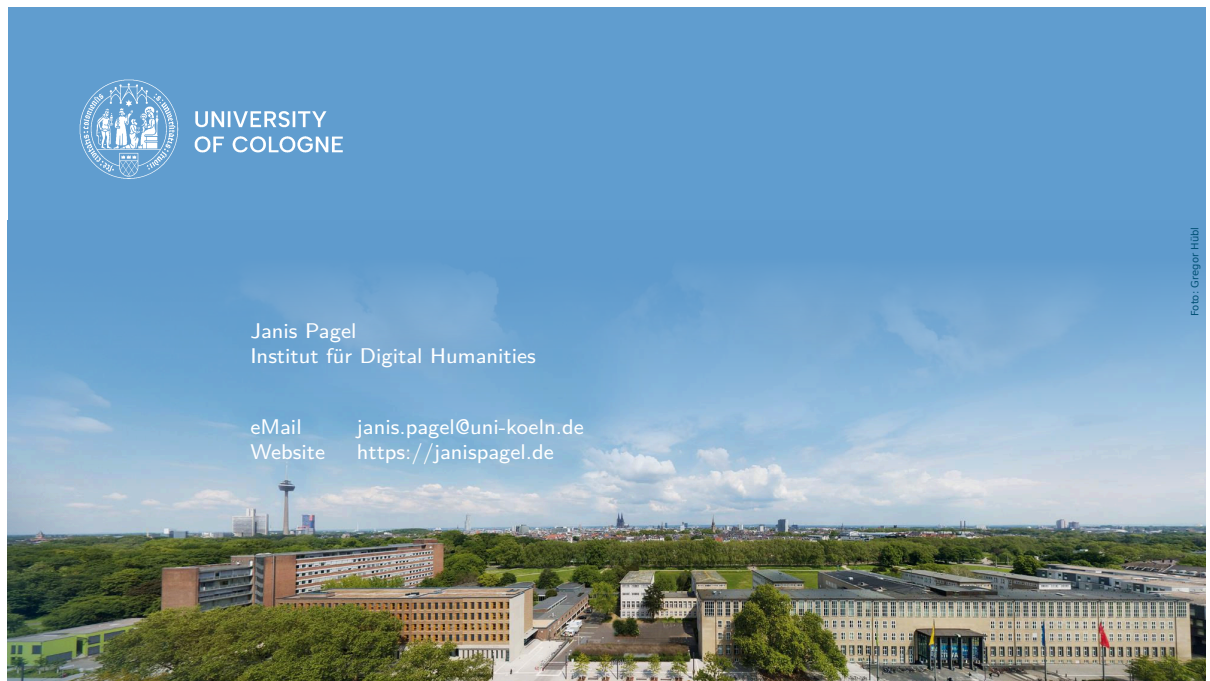
```
from sklearn.model_selection import train_test_split

train_df, test_df = train_test_split(df, train_size=0.6, random_state=42)

print(train_df)
   casino  enlargement  meeting  profit  super  text  xxx  class
2       1            0        1       0      1     0    0      0
0       1            1        0       0      1     1    1      0
3       1            1        1       0      0     0    0      0

print(test_df)
   casino  enlargement  meeting  profit  super  text  xxx  class
1       0            1        0       1      0     0    0      1
4       0            1        1       0      0     1    1      1
```

03

EXERCISE 08

Exercise 08

- https://lehre.idh.uni-koeln.de/site/assets/files/5151/exercise08.pdf


UNIVERSITY
OF COLOGNE
Janis Pagel
Institut für Digital Humanities
eMail janis.pagel@uni-koeln.de
Website https://janispagel.de
Foto: Gregor Hübl