Problem3

November 18, 2021

1 Problem 3

[22]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[9]:
```
indices = []
strings = []
with open("sequencesMSAfasta", "r") as file:
    for line in file.readlines():
        if line.startswith(">"): indices.append(line[:-1])
        else: strings.append(line[:-1])
```

[14]:
```
data = pd.DataFrame(
    data=strings,
    index=indices,
    columns=["String"]
)
data.head()
```

[14]:

	String
>AF153142	TGTACAAGACCCAACAATAATACAAGAAAAAGTATAAGGATAGGAC…
>AF153143	TGTACAAGGCCCGGCAATAATACAAGGAAAAGTATGAGGATAGGAC…
>AF153144	TGTACAAGACCCAATAATAATACAAGAAAAAGCATAAGGATAGGAC…
>AF153145	TGTACAAGACCCAACAATAATACAAGAAAAAGTATAAGGATAGGAC…
>HQ906866	TGCACAAGGCCCTACGATAAGGTAAGCTACAGGACACCTATAGGAR…

[44]:
```
def wdk(s1, s2, d=None, beta=None):
    assert len(s1) == len(s2)
    L = len(s1)

    if beta is None:
        beta = [1 for _ in s1]
    if d is None:
        d = len(beta)

    assert d <= len(beta)
```

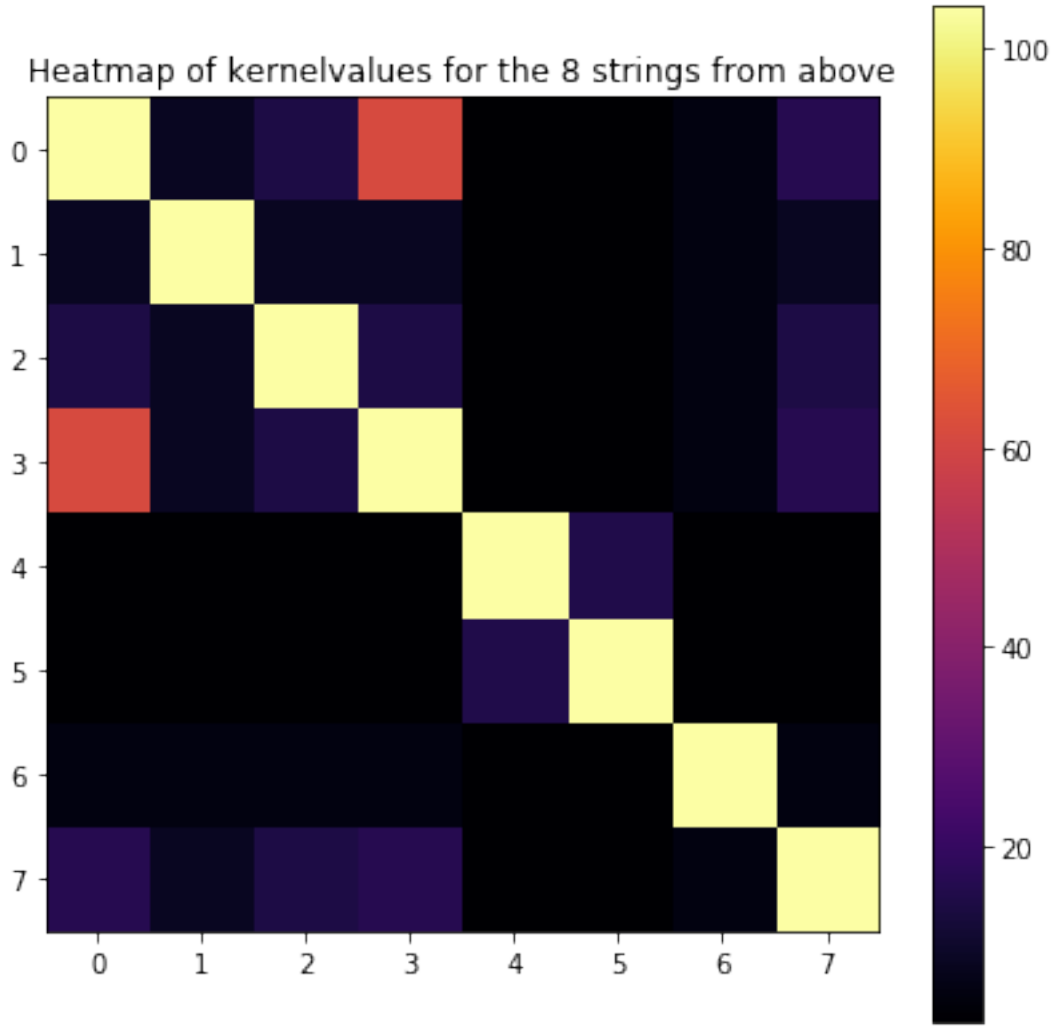
```
    return np.concatenate([[
        beta[k] * (1 if s1[k:k+l] == s2[k:k+l] else 0)
    for l in range(L-k)]
    for k in range(d)]).sum()
```

[29]:
```
def kernel(s, f, **kwargs):
    return np.array([[
        wdk(x, y, **kwargs)
    for x in s]
    for y in s])
```

[30]:
```
d = 3
beta = [
    2 * (d - k + 1) / (d*(d+1))
for k in range(1, d+1)]
```

[62]:
```
kernel_matrix = kernel(data.String, wdk, d=d, beta=beta)

plt.figure(figsize=(7, 7))
plt.imshow(kernel_matrix, cmap="inferno")
plt.colorbar()
plt.title("Heatmap of kernelvalues for the 8 strings from above")
plt.show()
```



As one can see, the largest values are on the diagonal, which makes sense, since these are the entries, where the string has been compared to itself.