In [12]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```

In [11]:

```
df= np.loadtxt('hw4_nolabel_train.dat')
```

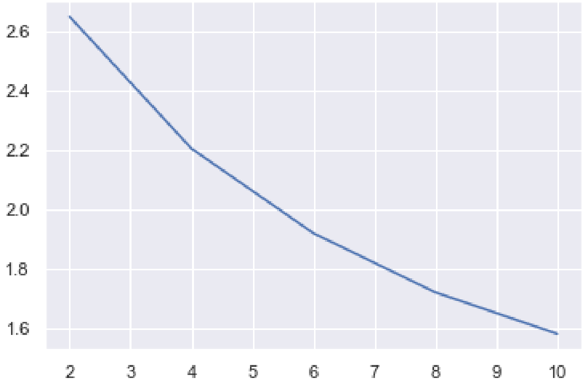
In [47]:

```
E = []
V = []
K = [2,4,6,8,10]
for k in K:
    err = 0
    sq = 0
    for T in range(500):
        kmeans = KMeans(n_clusters=k, random_state=T).fit(df)
        err += kmeans.score(df)/100/500
        sq += ((kmeans.score(df)/100)**2)/500
    E.append(-1*err)
    V.append(sq - err**2)
```

In [56]:

```
sns.set()
sns.lineplot(K,E)
print(E)
```

```
[2.6482794225946558, 2.2033466623852838, 1.9180549102254723, 1.7188274310568472, 1.57910757047811]
```

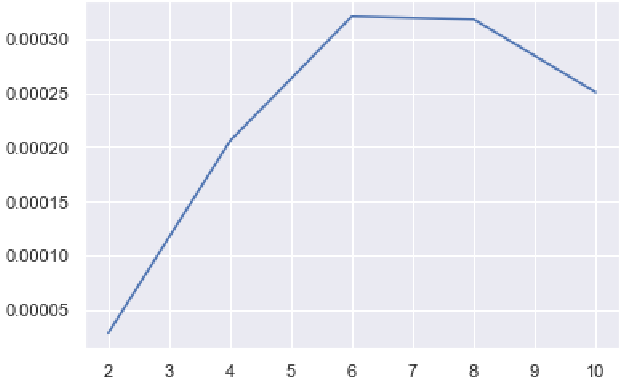


Q15: Average of err for 500 times decreases as k increases.

In [58]:

```
sns.set()
sns.lineplot(K,V)
print(V)
```

```
[2.8007285915343516e-05, 0.0002057968897553053, 0.0003212963569287907 7, 0.0003182846984906007, 0.00
025101657672488287]
```



Q16: Variance of err for 500 times increases until k = 6, then decreases.