in4080_2022_ex4_solutions

October 5, 2022

1 IN4080 Exercise set 4 2022, Solutions

[1]:
```
import numpy as np
```

1.1 The data

[2]:
```
man = np.array([30, 0, 120, 40 ])
woman = np.array([50, 50, 50, 50 ])
girl = np.array([84, 112, 0, 0])
```

1.2 Functions

The dot product can be calculated between numpy arrays by `@`. But we could have done everythin in pure python by defining `dot`. ### dot-product

[3]:
```
def dot(a,b):
    return sum([i*j for i,j in zip(a,b)])
```

[4]:
```
man @ woman
```

[4]: 9500

[5]:
```
dot(man, woman)
```

[5]: 9500

[6]:
```
dot(girl, woman)
```

[6]: 9800

1.2.1 Norm

To follow the steps in the calculation, we can also introduce the `norm` of a vector.

[7]:
```
def norm(v):
    return dot(v, v)**0.5
# or (v @ v)**0.5
```

[8]:
```
norm(woman)
```

[8]: 100.0

[9]:
```
norm(man)
```

[9]: 130.0

[10]:
```
norm(girl)
```

[10]: 140.0

1.2.2 The cosine between two vectors

[11]:
```
def cosi(v1, v2):
    return (v1 @ v2) / (norm(v1) * norm(v2))
```

1.3 Question a

[12]:
```
cosi(man,woman)
```

[12]: 0.7307692307692307

[13]:
```
cosi(woman,girl)
```

[13]: 0.7

1.4 Question c

We first calculate the inverse document frequency. By using numpy arrays, we can apply a function like `np.log` to a vector and it will apply to each component.

[14]:
```
idf = np.array([6/5, 6/2, 6/4, 6/3])
idf
```

[14]: array([1.2, 3. , 1.5, 2.])

[15]:
```
idf_log = np.log(idf)
idf_log
```

[15]: array([0.18232156, 1.09861229, 0.40546511, 0.69314718])

The vectors after applying tf-idf weighting.

[16]:
```
woman_tf_idf = woman * idf_log
woman_tf_idf
```

[16]:
```
array([ 9.11607784, 54.93061443, 20.27325541, 34.65735903])
```

[17]:
```
man_tf_idf = man * idf_log
man_tf_idf
```

[17]:
```
array([ 5.4696467 ,  0.        , 48.65581297, 27.72588722])
```

[18]:
```
girl_tf_idf = girl * idf_log
girl_tf_idf
```

[18]:
```
array([ 15.31501077, 123.04457633,   0.        ,   0.        ])
```

We can then calculate the cosine similarity of the adjusted vectors.

[19]:
```
# cosinus similarity after tf-df
cosi(man*idf_log, woman*idf_log)
```

[19]:
```
0.5170454343374329
```

[20]:
```
# cosinus similarity after tf-df
cosi(girl*idf_log,woman*idf_log)
```

[20]:
```
0.8104472436836814
```

[ ]:
```

```