tf_idf_prediction_with_sublinear_tf

February 27, 2024

[1]:
```
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import TfidfTransformer
from scipy.spatial.distance import cosine
```

[2]:
```
def cosine_similarity(data):
    #given a processed data, compute cosine similarity
    final_sample=data
    transformer = TfidfTransformer(sublinear_tf = "True")
    df_new=final_sample.values.tolist()
    tfidf_sparse = transformer.fit_transform(df_new)
    matrix_result=tfidf_sparse.todense()
    tf_idf=pd.DataFrame(matrix_result)
    tf_idf.columns=final_sample.columns
    columns=tf_idf.columns
    d={}
    for sample in columns:
        d[sample]=[1-cosine(tf_idf["Ceramic"], tf_idf[sample])]
    similarity=pd.DataFrame.from_dict(d)
    similarity.index=['score']
    similarity.sort_values(by='score', ascending=False, axis=1, inplace=True)
    return similarity
```

[3]:
```
def tf_idf_score(df):
    df_removed_0=df[df["Blank_pipe-APW.raw filtered Peak height"]==0]
    df_new=df_removed_0.drop(["row ID", "row m/z","row retention time",
                               'Gshallon_pipe-APW.raw filtered Peak height'],␣
 ↪axis=1)

    final_df=pd.DataFrame()
    for column in df_new.columns:
        final_df[column.split('_')[0]]=df_new[column]
    final_df.drop('Blank',axis=1, inplace=True)
    result=cosine_similarity(final_df)
    return result
```

1 Result of CP1

```
[4]: cp1=pd.read_csv("Blind_Pipes-APW-CP1.csv")

     tf_idf_score(cp1).transpose()
```

```
[4]:                    score
     Ceramic         1.000000
     Ntabacum        0.053926
     Nglauca         0.047536
     Nattenuata      0.035220
     Nobtusifolia    0.027709
     AmericanSpirit  0.026672
     Nquadrivalvis   0.022905
     Csericea        0.020871
     Auvaursi        0.016269
     Aludoviciana    0.014621
     Nrustica        0.014063
     Linflata        0.011610
     Vthapsus        0.010825
     Ssonomensis     0.006982
     Rglabra         0.006372
     Tbrevifolia     0.005672
```

2 Result of CP2

```
[5]: cp2=pd.read_csv("Blind_Pipes-APW-CP2.csv")
     tf_idf_score(cp2).transpose()
```

```
[5]:                    score
     Ceramic         1.000000
     Nquadrivalvis   0.140089
     Nattenuata      0.117366
     Nglauca         0.098412
     Ntabacum        0.079915
     Nobtusifolia    0.046013
     Nrustica        0.045940
     Aludoviciana    0.029226
     Linflata        0.025831
     AmericanSpirit  0.025738
     Csericea        0.020844
     Vthapsus        0.014874
     Auvaursi        0.013997
     Ssonomensis     0.010224
     Tbrevifolia     0.008880
     Rglabra         0.007746
```

3 Result of CP3

[6]:

```python
cp2=pd.read_csv("Blind_Pipes-APW-CP3.csv")
tf_idf_score(cp2).transpose()
```

[6]:

```
                   score
Ceramic         1.000000
Nglauca         0.149992
Nobtusifolia    0.141751
Nattenuata      0.103822
Nrustica        0.097000
Ntabacum        0.096762
Nquadrivalvis   0.092443
Aludoviciana    0.066759
Linflata        0.045957
AmericanSpirit  0.038635
Vthapsus        0.031770
Csericea        0.030258
Auvaursi        0.026825
Ssonomensis     0.023370
Rglabra         0.014591
Tbrevifolia     0.012406
```

4 Result of CP4

[7]:

```python
cp2=pd.read_csv("Blind_Pipes-APW-CP4.csv")
tf_idf_score(cp2).transpose()
```

[7]:

```
                   score
Ceramic         1.000000
Ntabacum        0.133657
Nglauca         0.130043
Nattenuata      0.095447
Nquadrivalvis   0.072615
Nobtusifolia    0.072444
Nrustica        0.064679
Aludoviciana    0.052159
Linflata        0.043817
AmericanSpirit  0.042720
Csericea        0.039657
Vthapsus        0.028350
Auvaursi        0.024333
Rglabra         0.021756
Ssonomensis     0.015056
Tbrevifolia     0.012670
```

4.1 Results of CP5

[8]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP5.csv")
tf_idf_score(cp2).transpose()
```

[8]:

```
                   score
Ceramic         1.000000
Nglauca         0.110390
Nattenuata      0.057399
Ntabacum        0.052216
Nquadrivalvis   0.041151
Nobtusifolia    0.038312
Nrustica        0.025226
Aludoviciana    0.022778
Linflata        0.019542
AmericanSpirit  0.016505
Csericea        0.016469
Vthapsus        0.008360
Auvaursi        0.007800
Tbrevifolia     0.006838
Ssonomensis     0.005515
Rglabra         0.005469
```

4.2 Results of CP6

[9]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP6.csv")
tf_idf_score(cp2).transpose()
```

[9]:

```
                   score
Ceramic         1.000000
Nglauca         0.145223
Ntabacum        0.089415
Linflata        0.082072
Nattenuata      0.077398
Nobtusifolia    0.077021
Nrustica        0.061274
Nquadrivalvis   0.057972
Aludoviciana    0.053164
AmericanSpirit  0.036266
Vthapsus        0.034587
Csericea        0.033000
Auvaursi        0.025864
Ssonomensis     0.021332
Rglabra         0.017730
Tbrevifolia     0.014526
```

4.3 Results of CP7

```
[10]: cp2=pd.read_csv("Blind_Pipes-APW-CP7.csv")
      tf_idf_score(cp2).transpose()
```

```
[10]:                    score
      Ceramic         1.000000
      Csericea        0.112851
      Auvaursi        0.108767
      Ntabacum        0.057662
      Nglauca         0.044460
      Rglabra         0.038461
      AmericanSpirit  0.030405
      Nquadrivalvis   0.028489
      Nattenuata      0.027024
      Linflata        0.023078
      Nrustica        0.020558
      Nobtusifolia    0.019601
      Tbrevifolia     0.017785
      Vthapsus        0.013762
      Ssonomensis     0.012753
      Aludoviciana    0.012332
```

4.4 Results of CP8

```
[11]: cp2=pd.read_csv("Blind_Pipes-APW-CP8.csv")
      tf_idf_score(cp2).transpose()
```

```
[11]:                    score
      Ceramic         1.000000
      Csericea        0.133637
      Auvaursi        0.095179
      Ntabacum        0.078784
      Nglauca         0.072376
      AmericanSpirit  0.057446
      Nattenuata      0.045322
      Nquadrivalvis   0.039472
      Rglabra         0.038748
      Linflata        0.037552
      Nobtusifolia    0.032485
      Nrustica        0.024287
      Aludoviciana    0.020679
      Vthapsus        0.018337
      Ssonomensis     0.018328
      Tbrevifolia     0.017761
```

4.5 Results of CP9

[12]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP9.csv")
tf_idf_score(cp2).transpose()
```

[12]:

```
                   score
Ceramic         1.000000
Nglauca         0.033209
Nattenuata      0.032648
Ntabacum        0.031809
AmericanSpirit  0.023003
Nquadrivalvis   0.022525
Csericea        0.014601
Nrustica        0.014248
Nobtusifolia    0.013188
Auvaursi        0.011891
Linflata        0.011234
Rglabra         0.004603
Ssonomensis     0.004542
Aludoviciana    0.004388
Tbrevifolia     0.002761
Vthapsus        0.002417
```

4.6 Results of CP10

[13]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP10.csv")
tf_idf_score(cp2).transpose()
```

[13]:

```
                   score
Ceramic         1.000000
Ntabacum        0.053926
Nglauca         0.047536
Nattenuata      0.035220
Nobtusifolia    0.027709
AmericanSpirit  0.026672
Nquadrivalvis   0.022905
Csericea        0.020871
Auvaursi        0.016269
Aludoviciana    0.014621
Nrustica        0.014063
Linflata        0.011610
Vthapsus        0.010825
Ssonomensis     0.006982
Rglabra         0.006372
Tbrevifolia     0.005672
```

4.7 Results of CP11

[14]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP11.csv")
tf_idf_score(cp2).transpose()
```

[14]:

```
                   score
Ceramic         1.000000
Nquadrivalvis   0.140089
Nattenuata      0.117366
Nglauca         0.098412
Ntabacum        0.079915
Nobtusifolia    0.046013
Nrustica        0.045940
Aludoviciana    0.029226
Linflata        0.025831
AmericanSpirit  0.025738
Csericea        0.020844
Vthapsus        0.014874
Auvaursi        0.013997
Ssonomensis     0.010224
Tbrevifolia     0.008880
Rglabra         0.007746
```

4.8 Results of CP12

[15]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP12.csv")
tf_idf_score(cp2).transpose()
```

[15]:

```
                   score
Ceramic         1.000000
Nglauca         0.149992
Nobtusifolia    0.141751
Nattenuata      0.103822
Nrustica        0.097000
Ntabacum        0.096762
Nquadrivalvis   0.092443
Aludoviciana    0.066759
Linflata        0.045957
AmericanSpirit  0.038635
Vthapsus        0.031770
Csericea        0.030258
Auvaursi        0.026825
Ssonomensis     0.023370
Rglabra         0.014591
Tbrevifolia     0.012406
```

4.9 Results of CP13

[16]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP13.csv")
tf_idf_score(cp2).transpose()
```

[16]:

```
                  score
Ceramic         1.000000
Ntabacum        0.133657
Nglauca         0.130043
Nattenuata      0.095447
Nquadrivalvis   0.072615
Nobtusifolia    0.072444
Nrustica        0.064679
Aludoviciana    0.052159
Linflata        0.043817
AmericanSpirit  0.042720
Csericea        0.039657
Vthapsus        0.028350
Auvaursi        0.024333
Rglabra         0.021756
Ssonomensis     0.015056
Tbrevifolia     0.012670
```

4.10 Results of CP14

[17]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP14.csv")
tf_idf_score(cp2).transpose()
```

[17]:

```
                  score
Ceramic         1.000000
Nglauca         0.110390
Nattenuata      0.057399
Ntabacum        0.052216
Nquadrivalvis   0.041151
Nobtusifolia    0.038312
Nrustica        0.025226
Aludoviciana    0.022778
Linflata        0.019542
AmericanSpirit  0.016505
Csericea        0.016469
Vthapsus        0.008360
Auvaursi        0.007800
Tbrevifolia     0.006838
Ssonomensis     0.005515
Rglabra         0.005469
```

4.11 Results of CP15

[18]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP15.csv")
tf_idf_score(cp2).transpose()
```

[18]:

```
                   score
Ceramic         1.000000
Nglauca         0.104477
Nobtusifolia    0.070489
Ntabacum        0.070029
Nattenuata      0.066803
Nrustica        0.059555
Nquadrivalvis   0.049422
Aludoviciana    0.043340
Linflata        0.041785
Csericea        0.033434
AmericanSpirit  0.026909
Vthapsus        0.022304
Ssonomensis     0.016773
Tbrevifolia     0.013903
Auvaursi        0.012887
Rglabra         0.011655
```

4.12 Results of CP16

[19]:

```
cp2=pd.read_csv("Blind_Pipes-APW-CP16.csv")
tf_idf_score(cp2).transpose()
```

[19]:

```
                   score
Ceramic         1.000000
Nglauca         0.145223
Ntabacum        0.089415
Linflata        0.082072
Nattenuata      0.077398
Nobtusifolia    0.077021
Nrustica        0.061274
Nquadrivalvis   0.057972
Aludoviciana    0.053164
AmericanSpirit  0.036266
Vthapsus        0.034587
Csericea        0.033000
Auvaursi        0.025864
Ssonomensis     0.021332
Rglabra         0.017730
Tbrevifolia     0.014526
```

4.13 Results of CP17

```
[20]: cp2=pd.read_csv("Blind_Pipes-APW-CP17.csv")
      tf_idf_score(cp2).transpose()
```

```
[20]:                    score
      Ceramic         1.000000
      Csericea        0.112851
      Auvaursi        0.108767
      Ntabacum        0.057662
      Nglauca         0.044460
      Rglabra         0.038461
      AmericanSpirit  0.030405
      Nquadrivalvis   0.028489
      Nattenuata      0.027024
      Linflata        0.023078
      Nrustica        0.020558
      Nobtusifolia    0.019601
      Tbrevifolia     0.017785
      Vthapsus        0.013762
      Ssonomensis     0.012753
      Aludoviciana    0.012332
```

4.14 Results of CP18

```
[21]: cp2=pd.read_csv("Blind_Pipes-APW-CP18.csv")
      tf_idf_score(cp2).transpose()
```

```
[21]:                    score
      Ceramic         1.000000
      Csericea        0.133637
      Auvaursi        0.095179
      Ntabacum        0.078784
      Nglauca         0.072376
      AmericanSpirit  0.057446
      Nattenuata      0.045322
      Nquadrivalvis   0.039472
      Rglabra         0.038748
      Linflata        0.037552
      Nobtusifolia    0.032485
      Nrustica        0.024287
      Aludoviciana    0.020679
      Vthapsus        0.018337
      Ssonomensis     0.018328
      Tbrevifolia     0.017761
```

```
[ ]:
```

[ ]:

```
```