tf_idf_prediction_with_idf

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()
    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
```

[6]:
```
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
```

[5]:
```
df = pd.read_csv("Blind_Pipes-APW-CP1.csv")
df.columns
```

[5]:
```
Index(['row ID', 'row m/z', 'row retention time',
       'Aludoviciana_pipe-APW.raw filtered Peak height',
       'AmericanSpirit_pipe-APW.raw filtered Peak height',
       'Auvaursi_pipe-APW.raw filtered Peak height',
       'Blank_pipe-APW.raw filtered Peak height',
       'Ceramic_Pipe_10-APW.raw filtered Peak height',
       'Csericea_pipe-APW.raw filtered Peak height',
       'Gshallon_pipe-APW.raw filtered Peak height',
       'Linflata_pipe-APW.raw filtered Peak height',
       'Nattenuata_pipe-APW.raw filtered Peak height',
       'Nglauca_pipe-APW.raw filtered Peak height',
       'Nobtusifolia_pipe-APW.raw filtered Peak height',
       'Nquadrivalvis_pipe-APW.raw filtered Peak height',
       'Nrustica_pipe-APW.raw filtered Peak height',
       'Ntabacum_pipe-APW.raw filtered Peak height',
       'Rglabra_pipe-APW.raw filtered Peak height',
       'Ssonomensis_pipe-APW.raw filtered Peak height',
       'Tbrevifolia_pipe-APW.raw filtered Peak height',
       'Vthapsus_pipe-APW.raw filtered Peak height'],
      dtype='object')
```

1 Result of CP1

[7]:
```
cp1=pd.read_csv("Blind_Pipes-APW-CP1.csv")

tf_idf_score(cp1).transpose()
```

[7]:

	score
Ceramic	1.000000
Ntabacum	0.037006
Nglauca	0.028811
Nattenuata	0.023394
AmericanSpirit	0.018668
Nobtusifolia	0.016883
Nquadrivalvis	0.012917
Auvaursi	0.012554
Csericea	0.008902
Nrustica	0.008153
Aludoviciana	0.006667
Vthapsus	0.005830
Linflata	0.005694
Ssonomensis	0.004060
Rglabra	0.003146
Tbrevifolia	0.000994

2 Result of CP2

[8]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP2.csv")
tf_idf_score(cp2).transpose()
```

[8]:

	score
Ceramic	1.000000
Nquadrivalvis	0.110212
Nattenuata	0.098550
Nglauca	0.062635
Ntabacum	0.059834
Nrustica	0.034703
Nobtusifolia	0.029563
Aludoviciana	0.019011
AmericanSpirit	0.012512
Linflata	0.010881
Auvaursi	0.010237
Vthapsus	0.009461
Csericea	0.008916
Ssonomensis	0.005778
Tbrevifolia	0.003489
Rglabra	0.003397

3 Result of CP3

[10]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP3.csv")
tf_idf_score(cp2).transpose()
```

[10]:

	score
Ceramic	1.000000
Nobtusifolia	0.113238
Nglauca	0.103138
Nattenuata	0.081673
Nrustica	0.072272
Ntabacum	0.069496
Nquadrivalvis	0.066033
Aludoviciana	0.052063
Linflata	0.026995
Vthapsus	0.025931
AmericanSpirit	0.022625
Auvaursi	0.019761
Ssonomensis	0.015183
Csericea	0.013054
Rglabra	0.009535
Tbrevifolia	0.005545

4 Result of CP4

[11]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP4.csv")
tf_idf_score(cp2).transpose()
```

[11]:

	score
Ceramic	1.000000
Ntabacum	0.092870
Nattenuata	0.068380
Nglauca	0.067825
Nobtusifolia	0.048839
Aludoviciana	0.039670
Nquadrivalvis	0.038390
Nrustica	0.037763
Vthapsus	0.023118
Linflata	0.022231
AmericanSpirit	0.019355
Csericea	0.015168
Auvaursi	0.014648
Rglabra	0.013675
Ssonomensis	0.008040
Tbrevifolia	0.004034

4.1 Results of CP5

[13]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP5.csv")
tf_idf_score(cp2).transpose()
```

[13]:

	score
Ceramic	1.000000
Nglauca	0.087402
Nattenuata	0.046711
Ntabacum	0.034881
Nobtusifolia	0.026851
Nquadrivalvis	0.025126
Nrustica	0.017060
Aludoviciana	0.013867
AmericanSpirit	0.011135
Linflata	0.009787
Csericea	0.008823
Vthapsus	0.004581
Auvaursi	0.004015
Ssonomensis	0.002910
Tbrevifolia	0.002869
Rglabra	0.002088

4.2 Results of CP6

[14]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP6.csv")
tf_idf_score(cp2).transpose()
```

[14]:

	score
Ceramic	1.000000
Nglauca	0.084446
Ntabacum	0.065593
Nattenuata	0.059270
Nobtusifolia	0.052418
Linflata	0.048934
Nrustica	0.042476
Aludoviciana	0.039476
Nquadrivalvis	0.034331
Vthapsus	0.029556
Auvaursi	0.018198
AmericanSpirit	0.017311
Ssonomensis	0.014078
Csericea	0.013004
Rglabra	0.012061
Tbrevifolia	0.005327

4.3 Results of CP7

[15]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP7.csv")
tf_idf_score(cp2).transpose()
```

[15]:

	score
Ceramic	1.000000
Auvaursi	0.079210
Csericea	0.072618
Ntabacum	0.035645
Rglabra	0.025491
Nglauca	0.022239
AmericanSpirit	0.016685
Nattenuata	0.015586
Nrustica	0.012374
Linflata	0.011509
Nquadrivalvis	0.011223
Nobtusifolia	0.008920
Tbrevifolia	0.008323
Vthapsus	0.007673
Ssonomensis	0.007507
Aludoviciana	0.004317

4.4 Results of CP8

[16]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP8.csv")
tf_idf_score(cp2).transpose()
```

[16]:

	score
Ceramic	1.000000
Csericea	0.085393
Auvaursi	0.066258
Ntabacum	0.051235
Nglauca	0.042103
AmericanSpirit	0.041725
Nattenuata	0.030325
Rglabra	0.026787
Linflata	0.023667
Nquadrivalvis	0.016822
Nobtusifolia	0.015802
Nrustica	0.013388
Vthapsus	0.012094
Ssonomensis	0.011077
Tbrevifolia	0.009964
Aludoviciana	0.009807

4.5 Results of CP9

[20]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP9.csv")
tf_idf_score(cp2).transpose()
```

[20]:

	score
Ceramic	1.000000
Nattenuata	0.031645
Nglauca	0.024707
Ntabacum	0.022955
AmericanSpirit	0.020391
Nquadrivalvis	0.015371
Auvaursi	0.011324
Nobtusifolia	0.010507
Nrustica	0.010149
Linflata	0.009476
Csericea	0.008842
Rglabra	0.003904
Ssonomensis	0.003276
Aludoviciana	0.002355
Vthapsus	0.002024
Tbrevifolia	0.001122

4.6 Results of CP10

[21]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP10.csv")
tf_idf_score(cp2).transpose()
```

[21]:

	score
Ceramic	1.000000
Ntabacum	0.037006
Nglauca	0.028811
Nattenuata	0.023394
AmericanSpirit	0.018668
Nobtusifolia	0.016883
Nquadrivalvis	0.012917
Auvaursi	0.012554
Csericea	0.008902
Nrustica	0.008153
Aludoviciana	0.006667
Vthapsus	0.005830
Linflata	0.005694
Ssonomensis	0.004060
Rglabra	0.003146
Tbrevifolia	0.000994

4.7 Results of CP11

[22]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP11.csv")
tf_idf_score(cp2).transpose()
```

[22]:

	score
Ceramic	1.000000
Nquadrivalvis	0.110212
Nattenuata	0.098550
Nglauca	0.062635
Ntabacum	0.059834
Nrustica	0.034703
Nobtusifolia	0.029563
Aludoviciana	0.019011
AmericanSpirit	0.012512
Linflata	0.010881
Auvaursi	0.010237
Vthapsus	0.009461
Csericea	0.008916
Ssonomensis	0.005778
Tbrevifolia	0.003489
Rglabra	0.003397

4.8 Results of CP12

[23]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP12.csv")
tf_idf_score(cp2).transpose()
```

[23]:

	score
Ceramic	1.000000
Nobtusifolia	0.113238
Nglauca	0.103138
Nattenuata	0.081673
Nrustica	0.072272
Ntabacum	0.069496
Nquadrivalvis	0.066033
Aludoviciana	0.052063
Linflata	0.026995
Vthapsus	0.025931
AmericanSpirit	0.022625
Auvaursi	0.019761
Ssonomensis	0.015183
Csericea	0.013054
Rglabra	0.009535
Tbrevifolia	0.005545

4.9 Results of CP13

[24]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP13.csv")
tf_idf_score(cp2).transpose()
```

[24]:

	score
Ceramic	1.000000
Ntabacum	0.092870
Nattenuata	0.068380
Nglauca	0.067825
Nobtusifolia	0.048839
Aludoviciana	0.039670
Nquadrivalvis	0.038390
Nrustica	0.037763
Vthapsus	0.023118
Linflata	0.022231
AmericanSpirit	0.019355
Csericea	0.015168
Auvaursi	0.014648
Rglabra	0.013675
Ssonomensis	0.008040
Tbrevifolia	0.004034

4.10 Results of CP14

[25]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP14.csv")
tf_idf_score(cp2).transpose()
```

[25]:

	score
Ceramic	1.000000
Nglauca	0.087402
Nattenuata	0.046711
Ntabacum	0.034881
Nobtusifolia	0.026851
Nquadrivalvis	0.025126
Nrustica	0.017060
Aludoviciana	0.013867
AmericanSpirit	0.011135
Linflata	0.009787
Csericea	0.008823
Vthapsus	0.004581
Auvaursi	0.004015
Ssonomensis	0.002910
Tbrevifolia	0.002869
Rglabra	0.002088

4.11 Results of CP15

[26]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP15.csv")
tf_idf_score(cp2).transpose()
```

[26]:

	score
Ceramic	1.000000
Nglauca	0.065783
Nobtusifolia	0.053351
Nattenuata	0.049387
Ntabacum	0.045184
Nrustica	0.045099
Aludoviciana	0.034634
Nquadrivalvis	0.030282
Linflata	0.024530
Vthapsus	0.018139
Csericea	0.017521
AmericanSpirit	0.014545
Ssonomensis	0.011170
Tbrevifolia	0.007516
Rglabra	0.007171
Auvaursi	0.006079

4.12 Results of CP16

[27]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP16.csv")
tf_idf_score(cp2).transpose()
```

[27]:

	score
Ceramic	1.000000
Nglauca	0.084446
Ntabacum	0.065593
Nattenuata	0.059270
Nobtusifolia	0.052418
Linflata	0.048934
Nrustica	0.042476
Aludoviciana	0.039476
Nquadrivalvis	0.034331
Vthapsus	0.029556
Auvaursi	0.018198
AmericanSpirit	0.017311
Ssonomensis	0.014078
Csericea	0.013004
Rglabra	0.012061
Tbrevifolia	0.005327

4.13 Results of CP17

[28]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP17.csv")
tf_idf_score(cp2).transpose()
```

[28]:

	score
Ceramic	1.000000
Auvaursi	0.079210
Csericea	0.072618
Ntabacum	0.035645
Rglabra	0.025491
Nglauca	0.022239
AmericanSpirit	0.016685
Nattenuata	0.015586
Nrustica	0.012374
Linflata	0.011509
Nquadrivalvis	0.011223
Nobtusifolia	0.008920
Tbrevifolia	0.008323
Vthapsus	0.007673
Ssonomensis	0.007507
Aludoviciana	0.004317

4.14 Results of CP18

[29]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP18.csv")
tf_idf_score(cp2).transpose()
```

[29]:

	score
Ceramic	1.000000
Csericea	0.085393
Auvaursi	0.066258
Ntabacum	0.051235
Nglauca	0.042103
AmericanSpirit	0.041725
Nattenuata	0.030325
Rglabra	0.026787
Linflata	0.023667
Nquadrivalvis	0.016822
Nobtusifolia	0.015802
Nrustica	0.013388
Vthapsus	0.012094
Ssonomensis	0.011077
Tbrevifolia	0.009964
Aludoviciana	0.009807

[ ]:

[ ]: