tf_idf_prediction_without_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(use_idf = False )
    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	0.030175
Nattenuata	0.022658
Nglauca	0.022380
AmericanSpirit	0.017095
Nobtusifolia	0.014895
Auvaursi	0.013736
Nquadrivalvis	0.011674
Nrustica	0.006714
Csericea	0.006684
Aludoviciana	0.005507
Ssonomensis	0.005069
Vthapsus	0.005011
Linflata	0.004498
Rglabra	0.002424
Tbrevifolia	0.000834

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	0.106603
Nattenuata	0.095079
Nglauca	0.054254
Ntabacum	0.051664
Nrustica	0.032152
Nobtusifolia	0.026637
Aludoviciana	0.017705
AmericanSpirit	0.010170
Linflata	0.008807
Auvaursi	0.008626
Vthapsus	0.008086
Csericea	0.006769
Ssonomensis	0.005878
Tbrevifolia	0.004036
Rglabra	0.003042

3 Result of CP3

[6]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP3.csv")
tf_idf_score(cp2).transpose()
```

[6]:

	score
Ceramic	1.000000
Nobtusifolia	0.110721
Nglauca	0.094013
Nattenuata	0.079505
Nrustica	0.067964
Ntabacum	0.059730
Nquadrivalvis	0.058625
Aludoviciana	0.055032
Vthapsus	0.026291
Linflata	0.025112
AmericanSpirit	0.019306
Auvaursi	0.018657
Ssonomensis	0.017666
Csericea	0.010671
Rglabra	0.009742
Tbrevifolia	0.005763

4 Result of CP4

[7]:
```
cp2=pd.read_csv("Blind_Pipes-APW-CP4.csv")
tf_idf_score(cp2).transpose()
```

[7]:

	score
Ceramic	1.000000
Ntabacum	0.087554
Nattenuata	0.065355
Nglauca	0.057487
Nobtusifolia	0.044973
Aludoviciana	0.037640
Nrustica	0.033478
Nquadrivalvis	0.032453
Linflata	0.020422
Vthapsus	0.019822
AmericanSpirit	0.016909
Auvaursi	0.014393
Rglabra	0.014054
Csericea	0.012061
Ssonomensis	0.008191
Tbrevifolia	0.003816

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	0.079989
Nattenuata	0.051915
Ntabacum	0.033098
Nobtusifolia	0.025135
Nquadrivalvis	0.021829
Nrustica	0.016886
Aludoviciana	0.012651
AmericanSpirit	0.009644
Linflata	0.008869
Csericea	0.006443
Auvaursi	0.004715
Ssonomensis	0.004117
Vthapsus	0.003591
Tbrevifolia	0.002858
Rglabra	0.002046

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	0.074996
Ntabacum	0.060168
Nattenuata	0.057292
Linflata	0.046881
Nobtusifolia	0.046822
Nrustica	0.040765
Aludoviciana	0.037863
Nquadrivalvis	0.028751
Vthapsus	0.027021
Auvaursi	0.016817
Ssonomensis	0.014211
AmericanSpirit	0.014188
Rglabra	0.011992
Csericea	0.010035
Tbrevifolia	0.005463

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
Auvaursi	0.080098
Csericea	0.068447
Ntabacum	0.032097
Rglabra	0.024630
Nglauca	0.018227
Nattenuata	0.016358
AmericanSpirit	0.014451
Nrustica	0.011873
Linflata	0.010167
Nquadrivalvis	0.010028
Ssonomensis	0.008535
Nobtusifolia	0.007777
Tbrevifolia	0.007732
Vthapsus	0.006960
Aludoviciana	0.004111

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	0.077874
Auvaursi	0.064339
Ntabacum	0.046189
AmericanSpirit	0.036907
Nglauca	0.034116
Nattenuata	0.031091
Rglabra	0.024565
Linflata	0.019922
Nquadrivalvis	0.014509
Nobtusifolia	0.013571
Ssonomensis	0.012145
Vthapsus	0.011699
Nrustica	0.011657
Tbrevifolia	0.009674
Aludoviciana	0.008848

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
Nattenuata	0.032485
Ntabacum	0.018802
Nglauca	0.017506
AmericanSpirit	0.016251
Nquadrivalvis	0.014231
Nrustica	0.009313
Nobtusifolia	0.009275
Auvaursi	0.009255
Linflata	0.008246
Csericea	0.006017
Rglabra	0.004923
Ssonomensis	0.004357
Aludoviciana	0.002183
Vthapsus	0.001770
Tbrevifolia	0.000924

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	0.030175
Nattenuata	0.022658
Nglauca	0.022380
AmericanSpirit	0.017095
Nobtusifolia	0.014895
Auvaursi	0.013736
Nquadrivalvis	0.011674
Nrustica	0.006714
Csericea	0.006684
Aludoviciana	0.005507
Ssonomensis	0.005069
Vthapsus	0.005011
Linflata	0.004498
Rglabra	0.002424
Tbrevifolia	0.000834

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	0.106603
Nattenuata	0.095079
Nglauca	0.054254
Ntabacum	0.051664
Nrustica	0.032152
Nobtusifolia	0.026637
Aludoviciana	0.017705
AmericanSpirit	0.010170
Linflata	0.008807
Auvaursi	0.008626
Vthapsus	0.008086
Csericea	0.006769
Ssonomensis	0.005878
Tbrevifolia	0.004036
Rglabra	0.003042

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
Nobtusifolia	0.110721
Nglauca	0.094013
Nattenuata	0.079505
Nrustica	0.067964
Ntabacum	0.059730
Nquadrivalvis	0.058625
Aludoviciana	0.055032
Vthapsus	0.026291
Linflata	0.025112
AmericanSpirit	0.019306
Auvaursi	0.018657
Ssonomensis	0.017666
Csericea	0.010671
Rglabra	0.009742
Tbrevifolia	0.005763

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	0.087554
Nattenuata	0.065355
Nglauca	0.057487
Nobtusifolia	0.044973
Aludoviciana	0.037640
Nrustica	0.033478
Nquadrivalvis	0.032453
Linflata	0.020422
Vthapsus	0.019822
AmericanSpirit	0.016909
Auvaursi	0.014393
Rglabra	0.014054
Csericea	0.012061
Ssonomensis	0.008191
Tbrevifolia	0.003816

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	0.079989
Nattenuata	0.051915
Ntabacum	0.033098
Nobtusifolia	0.025135
Nquadrivalvis	0.021829
Nrustica	0.016886
Aludoviciana	0.012651
AmericanSpirit	0.009644
Linflata	0.008869
Csericea	0.006443
Auvaursi	0.004715
Ssonomensis	0.004117
Vthapsus	0.003591
Tbrevifolia	0.002858
Rglabra	0.002046

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	0.059950
Nobtusifolia	0.050650
Nattenuata	0.047270
Nrustica	0.043362
Ntabacum	0.042210
Aludoviciana	0.035187
Nquadrivalvis	0.027982
Linflata	0.022619
Vthapsus	0.017892
Csericea	0.016059
AmericanSpirit	0.013186
Ssonomensis	0.011671
Tbrevifolia	0.007706
Rglabra	0.006990
Auvaursi	0.005513

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	0.074996
Ntabacum	0.060168
Nattenuata	0.057292
Linflata	0.046881
Nobtusifolia	0.046822
Nrustica	0.040765
Aludoviciana	0.037863
Nquadrivalvis	0.028751
Vthapsus	0.027021
Auvaursi	0.016817
Ssonomensis	0.014211
AmericanSpirit	0.014188
Rglabra	0.011992
Csericea	0.010035
Tbrevifolia	0.005463

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
Auvaursi	0.080098
Csericea	0.068447
Ntabacum	0.032097
Rglabra	0.024630
Nglauca	0.018227
Nattenuata	0.016358
AmericanSpirit	0.014451
Nrustica	0.011873
Linflata	0.010167
Nquadrivalvis	0.010028
Ssonomensis	0.008535
Nobtusifolia	0.007777
Tbrevifolia	0.007732
Vthapsus	0.006960
Aludoviciana	0.004111

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	0.077874
Auvaursi	0.064339
Ntabacum	0.046189
AmericanSpirit	0.036907
Nglauca	0.034116
Nattenuata	0.031091
Rglabra	0.024565
Linflata	0.019922
Nquadrivalvis	0.014509
Nobtusifolia	0.013571
Ssonomensis	0.012145
Vthapsus	0.011699
Nrustica	0.011657
Tbrevifolia	0.009674
Aludoviciana	0.008848

[ ]:
```

```

[ ]: