

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# read csv files
df = pd.read_csv('data_authors.csv')
df2 = pd.read_csv('data_authors2.csv')

# list of features which will be used to create the PCA model
features = ['av_sentences', 'av_voc', 'lex_richness', 'av_punct', 'av_num_word_ns',
            'av_num_words_sentences', 'av_num_words_sentences_ns', 'av_num_ns_sentences',
            'av_words_length', 'av_ego', 'av_modal', 'emphasis', 'comparison', 'contrast',
            'addition',
            'illustration', 'comma', 'dotcomma', 'excl', 'qmark', 'dash', 'slash', 'twodot',
            'CC', 'CD', 'DT', 'EX',
            'FW', 'IN', 'JJ', 'JJR', 'JJS', 'LS', 'MD', 'NN', 'NNS', 'NNP', 'NNPS', 'PDT',
            'POS', 'PRP', 'PRP$', 'RB', 'RBR', 'RBS', 'RP', 'TO', 'UH', 'VB', 'VBD', 'VBG',
            'VBN', 'VBP', 'VBZ',
            'WDT', 'WP', 'WP$', 'WRB']

x = df.loc[:, features].values
x2 = df2.loc[:, features].values
final = np.concatenate((x, x2)) # add the test book values

final_auth = pd.DataFrame(np.concatenate((df[['Authors']], df2[['Authors']]))) # add the un
know author of test book
final_auth.columns = ['Authors']

pca = PCA(n_components=2)
principalComponents = pca.fit_transform(final)
principalDf = pd.DataFrame(data=principalComponents
                           , columns=['principal component 1', 'principal component 2'])
finalDf = pd.concat([principalDf, final_auth], axis=1)

fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_xlabel('Principal Component 1', fontsize=15)
ax.set_ylabel('Principal Component 2', fontsize=15)
ax.set_title('2 Component PCA', fontsize=20)

# authors to take into account
targets = ['Bacon', 'Bentham', 'Berkeley', 'Burke', 'Cavendish', 'Clifford', 'Dewey',
            'Emerson', 'Fellerton', 'Godwin', 'Goldman', 'Hide', 'Hobbes',
            'Hume', 'James', 'Jordan', 'Martineau', 'Mill',
            'Russell', 'Santayana', 'Sidgwick', 'Spencer', 'Thoreau',
            'Wollstonecraft', 'unknown']
colors = ['#C0C0C0', '#808080', '#5D6D7E', '#FF0000', '#F41F53', '#F08080', '#800000', '#F3
9C12',
            '#E67E22', '#D35400', '#FFFF00', '#FFFE91', '#808000', '#00FF00', '#70E770',
            '#008000', '#00B9B9', '#00FFFF', '#00D1A0', '#008080', '#0000FF', '#000080', '#8C8
CEC',
            '#FF00FF', '#000000']

# create the graph
for target, color in zip(targets, colors):
    indicesToKeep = finalDf['Authors'] == target
    ax.scatter(finalDf.loc[indicesToKeep, 'principal component 1']
               , finalDf.loc[indicesToKeep, 'principal component 2']
               , c=color
               , s=50)
    ax.annotate(target, (
        finalDf.loc[indicesToKeep, 'principal component 1'], finalDf.loc[indicesToKeep, 'princi
pal component 2']))

ax.grid()
plt.show()

```