network-analysis-code-andreou

October 12, 2023

[1]:
```
#necessary imports
import pandas as pd
import editdistance
import scipy
import numpy as np
```

[2]:
```
## Reading the data and removing columns that are not important.
df = pd.read_csv("AppNetSci_ELP_Data.csv", sep = ',', encoding = 'latin-1',␣
↪usecols = ['Word','I_Zscore','I_Mean_Accuracy','I_NMG_Zscore',␣
↪'I_NMG_Mean_Accuracy'])
```

[3]:
```
#conversion and creation of word list to build dictionary set
objNames = df.iloc[:,0]
names = []
for x in range(len(objNames)):
    names.append(str(objNames[x]))
names[0:10]
```

[3]:
```
['zenith',
 'zephyr',
 'zeppelin',
 'zero',
 'zeroed',
 'zeros',
 'zest',
 'zigzag',
 'zigzagging',
 'zimbabwe']
```

[4]:
```
def buildOrthogonalityNetwork(nameList):
    #initializing dictionary
    orthoNetwork = dict(zip(nameList, [None]*len(nameList)))
    #for each name
    for name in nameList:
        for compareName in nameList:
            if editdistance.eval(name,compareName) == 1:
                if orthoNetwork[name] == None:
```

```python
                        orthoNetwork[name] = [compareName]
                    else:
                        orthoNetwork[name] += [compareName]
    return orthoNetwork
```

[5]:
```python
#AVOID RUNNNING: takes >30 minutes to process the (43,000)^2 entries with␣
↪NP-hard edit distance
#Load pickle file in next section if you are interested in raw data
#pickle file represents dict of all words that have at least a single connection
network = buildOrthogonalityNetwork(names)
```

[6]:
```python
#verify edges and nodes are approximately what they should be
totalEdges = 0
totalHermits = 0
for word in network.keys():
    if network[word] == None:
        totalHermits += 1
        continue
    totalEdges += len(network[word])
print((len(network.keys()),totalEdges/2, totalHermits/len(network.keys())))
```

```
(40468, 41517.0, 0.40748245527330235)
```

[7]:
```python
#removing out the hermits
hermitlessDict = {}
for word in network.keys():
    if network[word] != None:
        hermitlessDict[word] = network[word]
len(hermitlessDict.keys())
```

[7]: 23978

[31]:
```python
#saving either hermitless dict
import pickle
# create a binary pickle file
f = open("hermitlessSimilarDict.pkl","wb")
# write the python object (dict) to pickle file
pickle.dump(hermitlessSimilarDict,f)
# close file
f.close()
```

[2]:
```python
#loading either pickle if needed
import pickle
#load pickle file
file = open('hermitlessDict.pkl', 'rb')
#get data back
hermitlessDict = pickle.load(file)
```

```
file.close()
```

[3]:
```
#using bfs with a known common word to generate LCC for the undirected graph
LCCDict = {'top':hermitlessDict['top']}
#intializer
queue = hermitlessDict['top'].copy()
#cycle through
while len(queue) > 0:
    #remove first word in list
    potentialWord = queue.pop(0)
    #add to LCC
    LCCDict[potentialWord] = hermitlessDict[potentialWord]
    #find all words connected to this one
    potentialNewWords = hermitlessDict[potentialWord]
    #cycle through those words
    for potentialWord in potentialNewWords:
        #if we haven't added it in yet
        if potentialWord not in LCCDict.keys() and potentialWord not in queue:
            #add to queue
            queue.append(potentialWord)
len(LCCDict.keys())
```

[3]: 11363

[4]:
```
#finding degree of each node and average degree in LCC
numEdges = 0
for word in LCCDict.keys():
    numEdges += len(LCCDict[word])
numEdges/len(LCCDict.keys())
```

[4]: 5.7671389597817475

[5]:
```
#finding mean clustering coefficient
import networkx as nx
nxDict = nx.Graph(LCCDict)
```

[6]:
```
nx.transitivity(nxDict)
```

[6]: 0.27377735607763865

[7]:
```
#takes a few minutes, avoid if possible
nx.average_shortest_path_length(nxDict)
```

[7]: 8.782726505453184

[22]:
```
#back to data, grab data from original datasource file
recog = df.iloc[:,0:5]
```

```
print(recog)
```

```
                Word  I_Zscore  I_Mean_Accuracy  I_NMG_Zscore  \
0             zenith     -0.01             0.79          0.02
1             zephyr      0.36             0.39          0.15
2           zeppelin      0.51             0.61          0.29
3               zero     -0.76             0.97         -0.53
4             zeroed      0.48             0.76          0.00
…                …        …                …              …
40463  agglutination      1.14             0.29          2.31
40464     aggravated     -0.10             0.94         -0.17
40465     aggravates      0.53             0.77         -0.04
40466    aggravation      0.13             0.91          0.16
40467      aggregate     -0.08             0.84          0.08

       I_NMG_Mean_Accuracy
0                     0.79
1                     0.69
2                     0.96
3                     1.00
4                     1.00
…                    …
40463                 0.18
40464                 0.96
40465                 0.88
40466                 0.96
40467                 0.96

[40468 rows x 5 columns]
```

[15]:
```
#make large dataset with [word] x [recog. stats: first level features: ortho␣
↪measures]
masterOrthoData = np.zeros((len(LCCDict.keys())-1,11))
```

[18]:
```
#get recog stats from dataset
RTrow = 0
row = 0
masterWordList = []
for word in recog['Word']:
    if word in LCCDict.keys():
        #get RT
        masterWordList.append(word)
        RT = recog['I_Zscore'][RTrow]
        Acc = recog['I_Mean_Accuracy'][RTrow]
        RT_Name = recog['I_NMG_Zscore'][RTrow]
        Acc_Name = recog['I_NMG_Mean_Accuracy'][RTrow]
        masterOrthoData[row][0] = RT
```

```
        masterOrthoData[row][1] = Acc
        masterOrthoData[row][2] = RT_Name
        masterOrthoData[row][3] = Acc_Name
        row += 1
    RTrow += 1
masterOrthoData
```

[18]:
```
array([[-0.76,  0.97, -0.53, …,  0.  ,  0.  ,  0.  ],
       [-0.61,  1.  , -0.31, …,  0.  ,  0.  ,  0.  ],
       [-0.39,  0.94, -0.39, …,  0.  ,  0.  ,  0.  ],
       …,
       [-0.54,  1.  , -0.63, …,  0.  ,  0.  ,  0.  ],
       [ 0.13,  0.59,  0.13, …,  0.  ,  0.  ,  0.  ],
       [ 0.69,  0.25,  0.44, …,  0.  ,  0.  ,  0.  ]])
```

[26]:
```
#download first level predictors
## Reading the data and removing columns that are not important.
firstLevelArray = pd.read_csv("FirstLevelPredictorStorage.csv", sep = ',',
 ↪encoding = 'latin-1')
```

[27]:
```
#flipping to match other datasets
firstLevelArray = firstLevelArray.iloc[::-1]
```

[22]:
```
#now add in the predictors like we did the RT stats
row = 0
#for each word in established order from before
for word in masterWordList:
    #find row in data set
    nextRow = firstLevelArray.loc[firstLevelArray['Word'] == word]
    #capitalize if necessary and rerun
    if nextRow['Length'].values.size == 0:
        nextRow = firstLevelArray.loc[firstLevelArray['Word'] == word.
 ↪capitalize()]
    #fill out master
    masterOrthoData[row][4] = nextRow['Length'].values[0]
    masterOrthoData[row][7] = nextRow['Log_Freq_HAL'].values[0]
    if nextRow['NPhon'].values not in ['#']:
        masterOrthoData[row][5] = nextRow['NPhon'].values[0]
        masterOrthoData[row][6] = nextRow['NSyll'].values[0]
    else:
        masterOrthoData[row][5] = 0
        masterOrthoData[row][6] = 0
    row = row + 1
masterOrthoData
```

[22]:
```
array([[-0.76,  0.97, -0.53, …,  0.  ,  0.  ,  0.  ],
       [-0.61,  1.  , -0.31, …,  0.  ,  0.  ,  0.  ],
```

```
       [-0.39,  0.94, -0.39, …,  0.  ,  0.  ,  0.  ],
       …,
       [-0.54,  1.  , -0.63, …,  0.  ,  0.  ,  0.  ],
       [ 0.13,  0.59,  0.13, …,  0.  ,  0.  ,  0.  ],
       [ 0.69,  0.25,  0.44, …,  0.  ,  0.  ,  0.  ]])
```

[23]:

```
#now to generate the ortho network stats
row = 0
for word in masterWordList:
    wordDegree = len(LCCDict[word])
    wordCloseCentrality = nx.closeness_centrality(nxDict,u=word)
    wordClustering = nx.clustering(nxDict,nodes=word)
    masterOrthoData[row][8] = wordDegree
    masterOrthoData[row][9] = wordClustering
    masterOrthoData[row][10] = wordCloseCentrality
    row = row + 1
```

[24]:

```
import statsmodels.api as sm
```

[63]:

```
print(masterOrthoData[0])
print(masterOrthoData[0,4:8])
```

```
[-0.76        0.97       -0.53        1.          4.          4.
  2.          9.965       3.          0.33333333  0.1228311 ]
[4.    4.    2.    9.965]
```

[65]:

```
firstLevel = masterOrthoDataCopy[:,5:9].copy()
secondLevel = masterOrthoDataCopy[:,5:12].copy()
firstLevel = sm.tools.tools.add_constant(firstLevel, prepend=True)
secondLevel = sm.tools.tools.add_constant(secondLevel, prepend=True)
for x in range(4):
    mod = sm.OLS(masterOrthoData[:,x],firstLevel)
    res = mod.fit()
    print(res.summary())
    mod = sm.OLS(masterOrthoData[:,x],secondLevel)
    res = mod.fit()
    print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.437
Model:                            OLS   Adj. R-squared:                  0.437
Method:                 Least Squares   F-statistic:                     2202.
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                -292.07
No. Observations:               11362   AIC:                             594.1
Df Residuals:                   11357   BIC:                             630.8
```

```
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4653      0.016     29.537      0.000       0.434       0.496
x1            -0.0273      0.003     -9.129      0.000      -0.033      -0.021
x2            -0.0198      0.004     -5.424      0.000      -0.027      -0.013
x3             0.1178      0.005     22.038      0.000       0.107       0.128
x4            -0.0964      0.001    -86.848      0.000      -0.099      -0.094
==============================================================================
Omnibus:                     1589.327   Durbin-Watson:                   1.620
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3152.524
Skew:                           0.872   Prob(JB):                         0.00
Kurtosis:                       4.902   Cond. No.                         70.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.443
Model:                            OLS   Adj. R-squared:                  0.443
Method:                 Least Squares   F-statistic:                     1292.
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                -225.09
No. Observations:               11362   AIC:                             466.2
Df Residuals:                   11354   BIC:                             524.9
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7243      0.039     18.799      0.000       0.649       0.800
x1            -0.0488      0.004    -12.459      0.000      -0.056      -0.041
x2            -0.0179      0.004     -4.832      0.000      -0.025      -0.011
x3             0.1026      0.006     18.517      0.000       0.092       0.113
x4            -0.0950      0.001    -84.825      0.000      -0.097      -0.093
x5            -0.0051      0.001     -7.598      0.000      -0.006      -0.004
x6             0.0179      0.009      1.963      0.050    2.51e-05       0.036
x7            -0.9302      0.190     -4.895      0.000      -1.303      -0.558
==============================================================================
Omnibus:                     1519.117   Durbin-Watson:                   1.624
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2958.057
Skew:                           0.845   Prob(JB):                         0.00
Kurtosis:                       4.841   Cond. No.                     1.01e+03
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.339
Model:                            OLS   Adj. R-squared:                  0.339
Method:                 Least Squares   F-statistic:                     1458.
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 5393.2
No. Observations:               11362   AIC:                        -1.078e+04
Df Residuals:                   11357   BIC:                        -1.074e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2727      0.010     28.554      0.000       0.254       0.291
x1             0.0417      0.002     23.041      0.000       0.038       0.045
x2             0.0191      0.002      8.628      0.000       0.015       0.023
x3            -0.0539      0.003    -16.626      0.000      -0.060      -0.048
x4             0.0493      0.001     73.249      0.000       0.048       0.051
==============================================================================
Omnibus:                     3054.934   Durbin-Watson:                   1.622
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7959.990
Skew:                          -1.456   Prob(JB):                         0.00
Kurtosis:                       5.886   Cond. No.                         70.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.343
Model:                            OLS   Adj. R-squared:                  0.343
Method:                 Least Squares   F-statistic:                     848.2
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 5428.6
No. Observations:               11362   AIC:                        -1.084e+04
Df Residuals:                   11354   BIC:                        -1.078e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
==============================================================================
const          0.1944      0.023      8.298      0.000       0.148       0.240
x1             0.0488      0.002     20.532      0.000       0.044       0.054
x2             0.0190      0.002      8.451      0.000       0.015       0.023
x3            -0.0483      0.003    -14.345      0.000      -0.055      -0.042
x4             0.0485      0.001     71.295      0.000       0.047       0.050
x5             0.0028      0.000      6.816      0.000       0.002       0.004
x6            -0.0102      0.006     -1.843      0.065      -0.021       0.001
x7             0.1929      0.116      1.670      0.095      -0.034       0.419
==============================================================================
Omnibus:                     3007.604   Durbin-Watson:                   1.624
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7749.110
Skew:                          -1.438   Prob(JB):                         0.00
Kurtosis:                       5.845   Cond. No.                     1.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.257
Model:                            OLS   Adj. R-squared:                  0.257
Method:                 Least Squares   F-statistic:                     982.1
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 48.523
No. Observations:               11362   AIC:                            -87.05
Df Residuals:                   11357   BIC:                            -50.36
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1912      0.015    -12.506      0.000      -0.221      -0.161
x1             0.0149      0.003      5.151      0.000       0.009       0.021
x2             0.0089      0.004      2.503      0.012       0.002       0.016
x3             0.0604      0.005     11.640      0.000       0.050       0.071
x4            -0.0509      0.001    -47.226      0.000      -0.053      -0.049
==============================================================================
Omnibus:                     2065.797   Durbin-Watson:                   1.213
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4471.120
Skew:                           1.063   Prob(JB):                         0.00
Kurtosis:                       5.220   Cond. No.                         70.8
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.267
Model:                            OLS   Adj. R-squared:                  0.266
Method:                 Least Squares   F-statistic:                     590.2
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 123.99
No. Observations:               11362   AIC:                            -232.0
Df Residuals:                   11354   BIC:                            -173.3
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1726      0.037     -4.620      0.000      -0.246      -0.099
x1             0.0104      0.004      2.752      0.006       0.003       0.018
x2             0.0049      0.004      1.368      0.171      -0.002       0.012
x3             0.0550      0.005     10.245      0.000       0.045       0.066
x4            -0.0487      0.001    -44.876      0.000      -0.051      -0.047
x5            -0.0078      0.001    -11.954      0.000      -0.009      -0.007
x6            -0.0063      0.009     -0.715      0.474      -0.024       0.011
x7             0.5278      0.184      2.865      0.004       0.167       0.889
==============================================================================
Omnibus:                     1967.547   Durbin-Watson:                   1.221
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4154.299
Skew:                           1.026   Prob(JB):                         0.00
Kurtosis:                       5.136   Cond. No.                     1.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.154
Model:                            OLS   Adj. R-squared:                  0.154
Method:                 Least Squares   F-statistic:                     517.0
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 13625.
No. Observations:               11362   AIC:                        -2.724e+04
Df Residuals:                   11357   BIC:                        -2.720e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8142      0.005    175.898      0.000       0.805       0.823
x1             0.0099      0.001     11.335      0.000       0.008       0.012
x2             0.0066      0.001      6.196      0.000       0.005       0.009
x3            -0.0246      0.002    -15.635      0.000      -0.028      -0.021
x4             0.0138      0.000     42.303      0.000       0.013       0.014
==============================================================================
Omnibus:                     7626.096   Durbin-Watson:                   1.670
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           111595.659
Skew:                          -3.080   Prob(JB):                         0.00
Kurtosis:                      17.063   Cond. No.                         70.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.160
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     308.2
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        18:44:04   Log-Likelihood:                 13663.
No. Observations:               11362   AIC:                        -2.731e+04
Df Residuals:                   11354   BIC:                        -2.725e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8416      0.011     74.160      0.000       0.819       0.864
x1             0.0085      0.001      7.405      0.000       0.006       0.011
x2             0.0080      0.001      7.385      0.000       0.006       0.010
x3            -0.0249      0.002    -15.281      0.000      -0.028      -0.022
x4             0.0134      0.000     40.501      0.000       0.013       0.014
x5             0.0016      0.000      8.200      0.000       0.001       0.002
x6             0.0041      0.003      1.528      0.127      -0.001       0.009
x7            -0.2718      0.056     -4.857      0.000      -0.382      -0.162
==============================================================================
Omnibus:                     7556.412   Durbin-Watson:                   1.672
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           108606.862
Skew:                          -3.047   Prob(JB):                         0.00
Kurtosis:                      16.866   Cond. No.                     1.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
```

```
specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

[4]:
```python
#building gloveDict
glove_dictionary = {}
with open('glove.6B.200d.txt',encoding="utf8") as file:
    for each_message in file:
        words_in_message, coeff_cients = each_message.split(maxsplit=1)
        coeff_cients = np.array(coeff_cients.split(),dtype = float)
        glove_dictionary[words_in_message] = coeff_cients
```

[5]:
```python
#assigning all relevant words to their vectorized form in a dict
word_to_glove_dict = {}
for word in names:
    if word in glove_dictionary.keys():
        word_to_glove_dict[word] = glove_dictionary[word]
```

[7]:
```python
#cosine similarity tester similar to orthogonality of 1 tester.
#tested ideal threshold of 0.55
similarity_network = {}
for word in word_to_glove_dict.keys():
    neighbors = []
    for otherWord in word_to_glove_dict.keys():
        cosine = np.dot(word_to_glove_dict[word],word_to_glove_dict[otherWord])/
 ↪(np.linalg.norm(word_to_glove_dict[word])*np.linalg.
 ↪norm(word_to_glove_dict[otherWord]))
        if cosine > 0.6 and cosine < 0.99:
            neighbors.append(otherWord)
    if neighbors == []:
        similarity_network[word] = []
    else:
        similarity_network[word] = neighbors
```

[8]:
```python
#verify edges and nodes are approximately what they should be
totalEdges = 0
totalHermits = 0
for word in similarity_network.keys():
    if similarity_network[word] == []:
        totalHermits += 1
        continue
    totalEdges += len(similarity_network[word])
print(totalEdges/len(similarity_network.keys()), totalHermits/
 ↪len(similarity_network.keys()))
```

```
4.3372667036245085 0.40365882632454264
```

[12]:
```python
#removing out the hermits
hermitlessSimilarDict = {}
for word in similarity_network.keys():
    if similarity_network[word] != []:
        hermitlessSimilarDict[word] = similarity_network[word]
len(hermitlessSimilarDict.keys())
```

[12]: 22590

[15]:
```python
#using bfs with a known common word to generate LCC for the undirected graph
LCCSimilarDict = {'top':hermitlessSimilarDict['top']}
#intializer
queue = hermitlessSimilarDict['top'].copy()
#cycle through
while len(queue) > 0:
    #remove first word in list
    potentialWord = queue.pop(0)
    #add to LCC
    LCCSimilarDict[potentialWord] = hermitlessSimilarDict[potentialWord]
    #find all words connected to this one
    potentialNewWords = hermitlessSimilarDict[potentialWord]
    #cycle through those words
    for potentialWord in potentialNewWords:
        #if we haven't added it in yet
        if potentialWord not in LCCSimilarDict.keys() and potentialWord not in␣
 ↪queue:
            #add to queue
            queue.append(potentialWord)
len(LCCSimilarDict.keys())
```

[15]: 17626

[16]:
```python
#finding degree of each node and average degree in LCC
numEdges = 0
for word in LCCSimilarDict.keys():
    numEdges += len(LCCSimilarDict[word])
numEdges/len(LCCSimilarDict.keys())
```

[16]: 8.921820038579371

[17]:
```python
#finding mean clustering coefficient
import networkx as nx
nxSimilarDict = nx.Graph(LCCSimilarDict)
```

[18]:
```python
nx.transitivity(nxSimilarDict)
```

[18]: 0.43026549484798804

[20]:
```
#takes a few minutes, avoid if possible
nx.average_shortest_path_length(nxSimilarDict)
```

[20]: 8.24569964583268

[24]:
```
#now that we have functioning network we try to build the testing matrix again
masterSimilarData = np.zeros((len(LCCSimilarDict.keys()),11))
RTrow = 0
row = 0
masterWordList = []
for word in recog['Word']:
    if word in LCCSimilarDict.keys():
        #get RT
        masterWordList.append(word)
        RT = recog['I_Zscore'][RTrow]
        Acc = recog['I_Mean_Accuracy'][RTrow]
        RT_Name = recog['I_NMG_Zscore'][RTrow]
        Acc_Name = recog['I_NMG_Mean_Accuracy'][RTrow]
        masterSimilarData[row][0] = RT
        masterSimilarData[row][1] = Acc
        masterSimilarData[row][2] = RT_Name
        masterSimilarData[row][3] = Acc_Name
        row += 1
    RTrow += 1
masterSimilarData
```

[24]:
```
array([[-0.39,  0.94, -0.39, …,  0.  ,  0.  ,  0.  ],
       [ 0.38,  0.91,  0.27, …,  0.  ,  0.  ,  0.  ],
       [-0.45,  0.91, -0.21, …,  0.  ,  0.  ,  0.  ],
       …,
       [ 0.  ,  0.86,  0.1 , …,  0.  ,  0.  ,  0.  ],
       [-0.1 ,  0.94, -0.17, …,  0.  ,  0.  ,  0.  ],
       [ 0.53,  0.77, -0.04, …,  0.  ,  0.  ,  0.  ]])
```

[28]:
```
#first level predictors again
row = 0
#for each word in established order from before
for word in masterWordList:
    #find row in data set
    nextRow = firstLevelArray.loc[firstLevelArray['Word'] == word]
    #capitalize if necessary and rerun
    if nextRow['Length'].values.size == 0:
        nextRow = firstLevelArray.loc[firstLevelArray['Word'] == word.
↪capitalize()]
    #fill out master
    masterSimilarData[row][4] = nextRow['Length'].values[0]
    masterSimilarData[row][7] = nextRow['Log_Freq_HAL'].values[0]
```

```
    if nextRow['NPhon'].values not in ['#']:
        masterSimilarData[row][5] = nextRow['NPhon'].values[0]
        masterSimilarData[row][6] = nextRow['NSyll'].values[0]
    else:
        masterSimilarData[row][5] = 0
        masterSimilarData[row][6] = 0
    row = row + 1
masterSimilarData
```

[28]:

```
array([[-0.39,  0.94, -0.39, …,  0.  ,  0.  ,  0.  ],
       [ 0.38,  0.91,  0.27, …,  0.  ,  0.  ,  0.  ],
       [-0.45,  0.91, -0.21, …,  0.  ,  0.  ,  0.  ],
       …,
       [ 0.  ,  0.86,  0.1 , …,  0.  ,  0.  ,  0.  ],
       [-0.1 ,  0.94, -0.17, …,  0.  ,  0.  ,  0.  ],
       [ 0.53,  0.77, -0.04, …,  0.  ,  0.  ,  0.  ]])
```

[29]:

```
#finding semantic network stats
#now to generate the ortho network stats
row = 0
for word in masterWordList[0:20]:
    if word not in similarity_network.keys():
        print(word)
        row = row + 1
        continue
    wordDegree = len(similarity_network[word])
    wordCloseCentrality = nx.closeness_centrality(nxSimilarDict,u=word)
    wordClustering = nx.clustering(nxSimilarDict,nodes=word)
    masterSimilarData[row][8] = wordDegree
    masterSimilarData[row][9] = wordClustering
    masterSimilarData[row][10] = wordCloseCentrality
    row = row + 1
masterSimilarData
```

[29]:

```
array([[-0.39      ,  0.94      , -0.39      , …,  4.        ,
         0.5       ,  0.10663077],
       [ 0.38      ,  0.91      ,  0.27      , …,  7.        ,
         0.52380952,  0.1333631 ],
       [-0.45      ,  0.91      , -0.21      , …,  7.        ,
         0.47619048,  0.1079031 ],
       …,
       [ 0.        ,  0.86      ,  0.1       , …,  0.        ,
         0.        ,  0.        ],
       [-0.1       ,  0.94      , -0.17      , …,  0.        ,
         0.        ,  0.        ],
       [ 0.53      ,  0.77      , -0.04      , …,  0.        ,
         0.        ,  0.        ]])
```

[32]:
```
masterSimilarData[0]
```

[32]:
```
array([-0.39      ,  0.94      , -0.39      ,  1.        ,  4.        ,
        4.        ,  1.        ,  5.663     ,  4.        ,  0.5       ,
        0.10663077])
```

[35]:
```
import statsmodels.api as sm
firstLevel = masterSimilarData[:,4:8].copy()
secondLevel = masterSimilarData[:,4:12].copy()
firstLevel = sm.tools.tools.add_constant(firstLevel, prepend=True)
secondLevel = sm.tools.tools.add_constant(secondLevel, prepend=True)
for x in range(4):
    mod = sm.OLS(masterSimilarData[:,x],firstLevel)
    res = mod.fit()
    print(res.summary())
    mod = sm.OLS(masterSimilarData[:,x],secondLevel)
    res = mod.fit()
    print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.594
Model:                            OLS   Adj. R-squared:                  0.594
Method:                 Least Squares   F-statistic:                     6454.
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                -909.85
No. Observations:               17626   AIC:                             1830.
Df Residuals:                   17621   BIC:                             1869.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1012      0.012      8.402      0.000       0.078       0.125
x1             0.0169      0.002      8.245      0.000       0.013       0.021
x2             0.0119      0.002      5.031      0.000       0.007       0.016
x3             0.1081      0.004     30.386      0.000       0.101       0.115
x4            -0.1003      0.001    -98.759      0.000      -0.102      -0.098
==============================================================================
Omnibus:                     2388.155   Durbin-Watson:                   1.516
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4524.655
Skew:                           0.866   Prob(JB):                         0.00
Kurtosis:                       4.777   Cond. No.                         83.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
```

```
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.594
Model:                            OLS   Adj. R-squared:                  0.594
Method:                 Least Squares   F-statistic:                     3688.
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                -908.69
No. Observations:               17626   AIC:                             1833.
Df Residuals:                   17618   BIC:                             1896.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1008      0.012      8.365      0.000       0.077       0.124
x1             0.0169      0.002      8.249      0.000       0.013       0.021
x2             0.0119      0.002      5.037      0.000       0.007       0.017
x3             0.1081      0.004     30.380      0.000       0.101       0.115
x4            -0.1003      0.001    -98.718      0.000      -0.102      -0.098
x5            -0.0096      0.011     -0.895      0.371      -0.031       0.011
x6             0.2154      0.191      1.129      0.259      -0.159       0.589
x7             0.2837      0.881      0.322      0.747      -1.444       2.011
==============================================================================
Omnibus:                     2390.240   Durbin-Watson:                   1.516
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4530.614
Skew:                           0.867   Prob(JB):                         0.00
Kurtosis:                       4.778   Cond. No.                     6.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.09e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.268
Model:                            OLS   Adj. R-squared:                  0.268
Method:                 Least Squares   F-statistic:                     1616.
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                 12515.
No. Observations:               17626   AIC:                        -2.502e+04
Df Residuals:                   17621   BIC:                        -2.498e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
const          0.5371      0.006     95.464      0.000       0.526       0.548
x1             0.0205      0.001     21.473      0.000       0.019       0.022
x2             0.0015      0.001      1.330      0.184      -0.001       0.004
x3            -0.0304      0.002    -18.280      0.000      -0.034      -0.027
x4             0.0371      0.000     78.285      0.000       0.036       0.038
==============================================================================
Omnibus:                     7876.571   Durbin-Watson:                   1.617
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            44342.852
Skew:                          -2.111   Prob(JB):                         0.00
Kurtosis:                       9.523   Cond. No.                         83.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.269
Model:                            OLS   Adj. R-squared:                  0.268
Method:                 Least Squares   F-statistic:                     924.1
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                 12517.
No. Observations:               17626   AIC:                        -2.502e+04
Df Residuals:                   17618   BIC:                        -2.496e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5371      0.006     95.423      0.000       0.526       0.548
x1             0.0205      0.001     21.483      0.000       0.019       0.022
x2             0.0015      0.001      1.324      0.185      -0.001       0.004
x3            -0.0304      0.002    -18.290      0.000      -0.034      -0.027
x4             0.0371      0.000     78.256      0.000       0.036       0.038
x5             0.0046      0.005      0.921      0.357      -0.005       0.014
x6            -0.1285      0.089     -1.442      0.149      -0.303       0.046
x7             0.1690      0.411      0.411      0.681      -0.637       0.975
==============================================================================
Omnibus:                     7878.781   Durbin-Watson:                   1.616
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            44377.509
Skew:                          -2.112   Prob(JB):                         0.00
Kurtosis:                       9.526   Cond. No.                     6.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
[2] The condition number is large, 6.09e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.469
Model:                            OLS   Adj. R-squared:                  0.469
Method:                 Least Squares   F-statistic:                     3887.
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                -2994.2
No. Observations:               17626   AIC:                             5998.
Df Residuals:                   17621   BIC:                             6037.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0685      0.014     -5.054      0.000      -0.095      -0.042
x1             0.0037      0.002      1.590      0.112      -0.001       0.008
x2             0.0370      0.003     13.941      0.000       0.032       0.042
x3             0.0832      0.004     20.778      0.000       0.075       0.091
x4            -0.0773      0.001    -67.575      0.000      -0.080      -0.075
==============================================================================
Omnibus:                     2637.143   Durbin-Watson:                   1.262
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5177.385
Skew:                           0.930   Prob(JB):                         0.00
Kurtosis:                       4.896   Cond. No.                         83.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.469
Model:                            OLS   Adj. R-squared:                  0.469
Method:                 Least Squares   F-statistic:                     2221.
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                -2993.2
No. Observations:               17626   AIC:                             6002.
Df Residuals:                   17618   BIC:                             6065.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0686      0.014     -5.053      0.000      -0.095      -0.042
x1             0.0037      0.002      1.586      0.113      -0.001       0.008
x2             0.0370      0.003     13.943      0.000       0.032       0.042
```

```
x3             0.0832      0.004     20.771      0.000       0.075       0.091
x4            -0.0773      0.001    -67.553      0.000      -0.079      -0.075
x5             0.0029      0.012      0.242      0.808      -0.021       0.027
x6             0.2868      0.215      1.336      0.182      -0.134       0.708
x7            -1.0208      0.992     -1.029      0.303      -2.965       0.923
==============================================================================
Omnibus:                     2637.965   Durbin-Watson:                   1.262
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5180.339
Skew:                           0.930   Prob(JB):                         0.00
Kurtosis:                       4.896   Cond. No.                     6.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.09e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.167
Model:                            OLS   Adj. R-squared:                  0.167
Method:                 Least Squares   F-statistic:                     883.7
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                 20483.
No. Observations:               17626   AIC:                        -4.096e+04
Df Residuals:                   17621   BIC:                        -4.092e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8460      0.004    236.311      0.000       0.839       0.853
x1             0.0085      0.001     13.987      0.000       0.007       0.010
x2            -0.0011      0.001     -1.526      0.127      -0.002       0.000
x3            -0.0228      0.001    -21.584      0.000      -0.025      -0.021
x4             0.0149      0.000     49.237      0.000       0.014       0.015
==============================================================================
Omnibus:                    12223.638   Durbin-Watson:                   1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           209201.041
Skew:                          -3.168   Prob(JB):                         0.00
Kurtosis:                      18.643   Cond. No.                         83.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                      y   R-squared:                       0.167
Model:                            OLS   Adj. R-squared:                  0.167
Method:                 Least Squares   F-statistic:                     505.1
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:37:33   Log-Likelihood:                 20483.
No. Observations:               17626   AIC:                        -4.095e+04
Df Residuals:                   17618   BIC:                        -4.089e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8459      0.004    236.150      0.000       0.839       0.853
x1             0.0085      0.001     13.998      0.000       0.007       0.010
x2            -0.0011      0.001     -1.523      0.128      -0.002       0.000
x3            -0.0228      0.001    -21.588      0.000      -0.025      -0.021
x4             0.0149      0.000     49.243      0.000       0.014       0.015
x5            -0.0011      0.003     -0.350      0.727      -0.007       0.005
x6             0.0212      0.057      0.374      0.709      -0.090       0.132
x7             0.1423      0.262      0.544      0.587      -0.371       0.655
==============================================================================
Omnibus:                    12222.734   Durbin-Watson:                   1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           209159.277
Skew:                          -3.168   Prob(JB):                         0.00
Kurtosis:                      18.641   Cond. No.                     6.09e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.09e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```