In [22]:

```
import glob
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from scipy.spatial.distance import pdist, squareform
import seaborn as sns
%matplotlib inline
```

In [23]:

```
# Use the glob library to create a list of file names
filenames = glob.glob("C:/Users/SMH_2/Documents/project/Similarity Measures/1666_texts/*.tx
# Parse those filenames to create a list of file keys (ID numbers)

filekeys = [f.split('\\')[-1].split('.')[0] for f in filenames]


# Create a CountVectorizer instance with the parameters you need
vectorizer = CountVectorizer(input="filename", max_features=1000, max_df=0.7)
#setting document frequency can be a substitute for using stopwords

wordcounts = vectorizer.fit_transform(filenames).toarray()
#vectorise the count
```

In [24]:

```
metadata = pd.read_csv("1666_metadata.csv", index_col="TCP ID")
```

1 Euclidean Distance

In [25]:

```
euclidean_distances = pd.DataFrame(squareform(pdist(wordcounts)), index=filekeys, columns=f
#euclidean_distances['index'] =  [re.sub(r'[\n\r]*','', str(x)) for x in euclidean_distance
(euclidean_distances)
```

Out[25]:

	A23770	A25198	A25743	A26249	A26426	A26482	A2817
A23770	0.000000	253.136327	974.878967	250.155951	116.910222	738.046069	1381.36381
A25198	253.136327	0.000000	867.675631	282.400425	278.309181	703.720115	1333.28691
A25743	974.878967	867.675631	0.000000	921.926787	1013.308936	859.162965	1382.74256
A26249	250.155951	282.400425	921.926787	0.000000	283.407833	703.182764	1332.42560
A26426	116.910222	278.309181	1013.308936	283.407833	0.000000	770.991569	1447.08534
...	...	...	...	...	...	...	..
B05591	120.548745	280.538767	1015.319162	284.042250	24.124676	772.415691	1450.53921
B05835	117.885538	268.747093	1008.435422	280.839812	26.888659	770.310976	1445.78732
B06022	116.760439	276.566448	1013.905321	281.064050	20.663978	771.976036	1447.42288
B06375	118.806565	278.328942	1015.689913	281.898918	18.138357	772.774870	1448.44019
B06872	113.881517	271.737005	1009.249226	274.096698	35.114100	765.747347	1435.47518

142 rows × 142 columns

In [26]:

```
top5_euclidean = euclidean_distances.nsmallest(6, 'A28989')['A28989'][1:]
#find 5 books which are closest to the book given by A28989
#used 6 to avoid a[i][i] element- its a 2d matrix
print(top5_euclidean)
top5_euclidean.index
```

```
A62436      988.557029
A43020      988.622274
A29017     1000.024000
A56390     1005.630151
A44061     1012.873141
Name: A28989, dtype: float64
```

Out[26]:

```
Index(['A62436', 'A43020', 'A29017', 'A56390', 'A44061'], dtype='object')
```

In [27]:

```
metadata1=(metadata.loc[top5_euclidean.index, ['Author','Title','Keywords']])
metadata1
```

Out[27]:

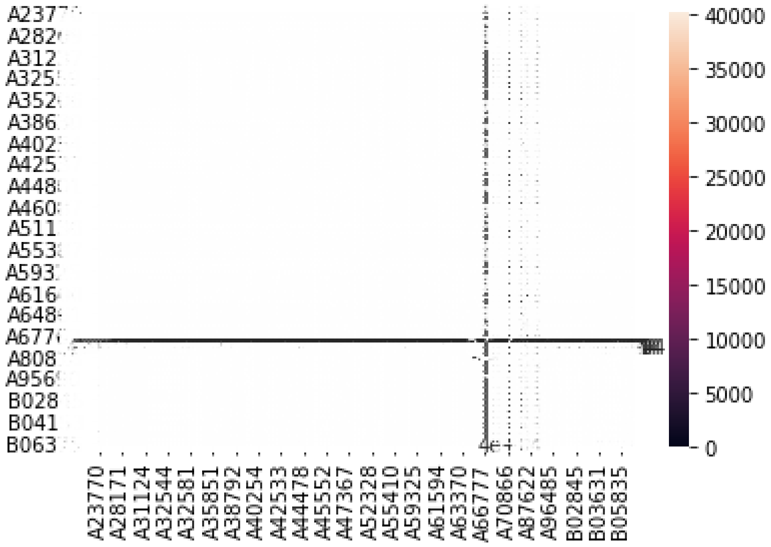
	Author	Title	Keywords
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A43020	Harvey, Gideon, 1640?-1700?	Morbus anglicus: or, The anatomy of consumptio...	Tuberculosis -- Early works to 1800.
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A56390	Parker, Samuel, 1640-1688.	A free and impartial censure of the Platonick ...	Platonists. Empiricism -- Early works to 1800.
A44061	Hodges, Nathaniel, 1629-1688.	Vindiciæ medicinæ & medicorum: or An apology f...	Medicine -- Early works to 1800. Plague -- Eng...

In [28]:

```
sns.heatmap(euclidean_distances, annot=True)
```

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20b95866dc8>
```



In [29]:

```
euclidean_distances.style.background_gradient(cmap='Blues')
```

Out[29]:

	A23770	A25198	A25743	A26249	A26426	A26482	A28171
A23770	0.000000	253.136327	974.878967	250.155951	116.910222	738.046069	1381.363819
A25198	253.136327	0.000000	867.675631	282.400425	278.309181	703.720115	1333.286916
A25743	974.878967	867.675631	0.000000	921.926787	1013.308936	859.162965	1382.742565
A26249	250.155951	282.400425	921.926787	0.000000	283.407833	703.182764	1332.425608
A26426	116.910222	278.309181	1013.308936	283.407833	0.000000	770.991569	1447.085346
A26482	738.046069	703.720115	859.162965	703.182764	770.991569	0.000000	1325.515749
A28171	1381.363819	1333.286916	1382.742565	1332.425608	1447.085346	1325.515749	0.000000
A28209	122.266103	279.512075	1012.731949	281.806671	36.152455	771.820575	1448.879222
A28989	1046.487936	1034.489729	1272.511690	1019.914212	1061.949622	1190.055041	1601.098685

2 Cosine Distance

In [30]:

```
cosine_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='cosine')), index=filek

top5_cosine = cosine_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_cosine)
```

```
A29017    0.432181
A43020    0.616269
A62436    0.629395
A57484    0.633845
A60482    0.663113
Name: A28989, dtype: float64
```

In [31]:

```
m2=(metadata.loc[top5_cosine.index, ['Author','Title','Keywords']])
m2
```

Out[31]:

	Author	Title	Keywords
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A43020	Harvey, Gideon, 1640?-1700?	Morbus anglicus: or, The anatomy of consumptio...	Tuberculosis -- Early works to 1800.
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A57484	Rochefort, César de, b. 1605.	The history of the Caribby-islands, viz, Barba...	Natural history -- West Indies. Carib Indians.
A60482	Smith, John, 1630-1679.	Gērochomia vasilikē King Solomons portraitur...	Bible. -- O.T. -- Ecclesiastes XII, 1-6 -- Par...

In [32]:

```
cosine_distances.style.background_gradient(cmap='Blues')
```

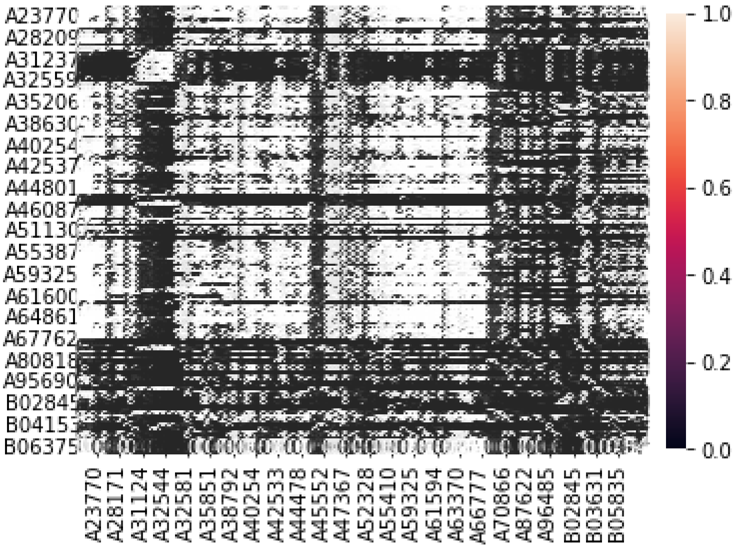
A31124	0.873058	0.829564	0.814010	0.835483	0.907176	0.810920	0.827308	0.841870	0.897653	0.789759
A31229	0.568143	0.492912	0.435313	0.516509	0.752492	0.566971	0.576743	0.855678	0.715085	0.504402
A31237	0.732072	0.694108	0.753562	0.744108	0.843147	0.772366	0.737286	0.928544	0.885323	0.788376
A32207	0.892939	0.879053	0.845474	0.900379	0.947881	0.830992	0.840664	0.974430	0.951521	0.890084
A32288	0.887352	0.871419	0.826354	0.893281	0.944894	0.820367	0.835037	0.967171	0.948556	0.883359
A32484	0.876376	0.809186	0.809437	0.815190	0.830396	0.831127	0.807914	0.962038	0.890891	0.817320
A32544	0.872494	0.866001	0.819778	0.869397	0.857545	0.835280	0.834103	0.945087	0.954347	0.882123
A32555	0.850125	0.836358	0.784664	0.856794	0.883713	0.856684	0.793697	0.964139	0.850947	0.841132
A32557	0.905855	0.847941	0.777484	0.836582	0.890293	0.826312	0.845239	0.977612	0.927167	0.862508
A32559	0.896834	0.866227	0.858452	0.906342	0.933765	0.886157	0.837849	0.974468	0.952355	0.915779

In [33]:

```
sns.heatmap(cosine_distances, annot=True)
```

Out[33]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20b98fc3d48>
```



3 Chebyshev distance

In [34]:

```python
chebyshev_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='chebyshev')), index

top5_chebyshev= chebyshev_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_chebyshev)
m3=(metadata.loc[top5_chebyshev.index, ['Author','Title','Keywords']])
m3
```

```
A29017    668.0
A57484    745.0
A65296    831.0
A25743    840.0
A30203    846.0
Name: A28989, dtype: float64
```

Out[34]:

	Author	Title	Keywords
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A57484	Rochefort, César de, b. 1605.	The history of the Caribby-islands, viz, Barba...	Natural history -- West Indies. Carib Indians.
A65296	Watson, Thomas, d. 1686.	The godly mans picture drawn with a scripture-...	Puritans -- Doctrines. Christian life.
A25743	Aranda, Emanuel d', b. 1602.	The history of Algiers and it's slavery with m...	Slavery -- Algeria -- Algiers -- Personal narr...
A30203	Bunyan, John, 1628-1688.	Sighs from hell, or, The groans of a damned so...	Hell. Future punishment.

4 Hamming distance

In [35]:

```
hamming_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='hamming')), index=fil

top5_hamming= hamming_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_hamming)
m4=(metadata.loc[top5_hamming.index, ['Author','Title','Keywords']])
m4
```

```
A26426    0.497
A31124    0.498
A95690    0.499
A42537    0.500
B03109    0.500
Name: A28989, dtype: float64
```

Out[35]:

	Author	Title	Keywords
A26426	Campbel, Agnes.	Advertisement be [sic] Agnes Campbel relict of...	Guthrie, William, 1620-1665. Presbyterian Chur...
A31124	NaN	The Case of the booksellers and printers state...	Moore, John, -- patentee. Law printing -- Pate...
A95690	Corporation of London (England)	A Table of the severall scantlings & sorts of ...	Lumber -- Law and legislation -- England -- Lo...
A42537	Gayton, Edmund, 1608-1666.	To Mr. Robert Whitehall at the wels at Astrop	Health resorts -- Humor.
B03109	NaN	Englands tryumph, and Hollands downfall; or, t...	Rupert, -- Prince, Count Palatine, 1619-1682 -...

In [36]:

```
#hamming_distances.style.background_gradient(cmap='Blues')
```

5. Jaccard distance

In [37]:

```
jaccard_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='jaccard')), index=fil

top5_jaccard= jaccard_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_jaccard)
m5=(metadata.loc[top5_jaccard.index, ['Author','Title','Keywords']])
m5
```

```
A45552    0.931357
A31229    0.935201
A41527    0.937031
A61867    0.938053
A61503    0.938806
Name: A28989, dtype: float64
```

Out[37]:

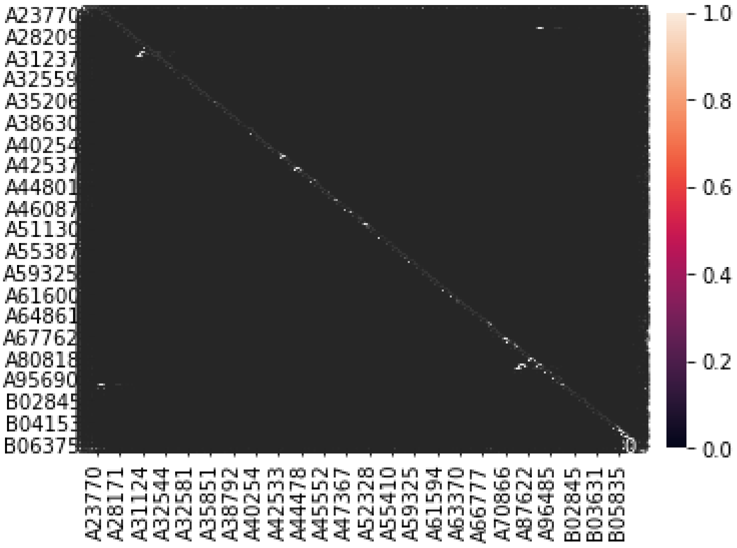
	Author	Title	Keywords
A45552	Hardy, Nathaniel, 1618-1670.	Lamentation, mourning, and woe sighed forth in...	Bible. -- N.T. -- Luke XIX, 41 -- Sermons. Fir...
A31229	Castlemaine, Roger Palmer, Earl of, 1634-1705.	An account of the present war between the Vene...	Venice (Italy) -- History -- Turkish Wars, 17t...
A41527	Goodwin, Thomas, 1600-1680.	Patience and its perfect work under sudden & s...	Patience. Conduct of life.
A61867	Sanderson, Robert, 1587-1663.	Five cases of conscience occasionally determin...	Christian life -- Early works to 1800.
A61503	Sancroft, William, 1617-1693.	Lex ignea, or, The school of righteousness a s...	London (England) -- Fire, 1666 -- Sermons.

In [38]:

```
sns.heatmap(jaccard_distances, annot=True)
```

Out[38]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20bd153f148>
```



In [39]:

```
jaccard_distances.style.background_gradient(cmap='Blues')
```

Out[39]:

	A23770	A25198	A25743	A26249	A26426	A26482	A28171	A28209	A28989	A29017
A23770	0.000000	0.946429	0.962179	0.945687	0.969977	0.966507	0.982533	0.981900	0.939344	0.969325
A25198	0.946429	0.000000	0.949555	0.917829	0.989111	0.950233	0.970670	0.989031	0.949464	0.955357
A25743	0.962179	0.949555	0.000000	0.944444	0.993579	0.950076	0.971233	0.996779	0.961988	0.961207
A26249	0.945687	0.917829	0.944444	0.000000	0.987784	0.940639	0.970793	0.991228	0.941878	0.962644
A26426	0.969977	0.989111	0.993579	0.987784	0.000000	0.993056	0.997063	0.950355	0.982213	0.993115
A26482	0.966507	0.950233	0.950076	0.940639	0.993056	0.000000	0.974719	0.996546	0.969040	0.976155
A28171	0.982533	0.970670	0.971233	0.970793	0.997063	0.974719	0.000000	0.994169	0.966851	0.972715
A28209	0.981900	0.989031	0.996779	0.991228	0.950355	0.996546	0.994169	0.000000	0.986328	0.991409
A28989	0.939344	0.949464	0.961988	0.941878	0.982213	0.969040	0.966851	0.986328	0.000000	0.944000

6. Kulsinski Distance

In [40]:

```
kulsinski_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='kulsinski')), index

top5_kulsinski= kulsinski_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_kulsinski)
m6=(metadata.loc[top5_kulsinski.index, ['Author','Title','Keywords']])
m6
```

```
A45552    0.931357
A31229    0.935201
A41527    0.937031
A61867    0.938053
A61503    0.938806
Name: A28989, dtype: float64
```

Out[40]:

	Author	Title	Keywords
A45552	Hardy, Nathaniel, 1618-1670.	Lamentation, mourning, and woe sighed forth in...	Bible. -- N.T. -- Luke XIX, 41 -- Sermons. Fir...
A31229	Castlemaine, Roger Palmer, Earl of, 1634-1705.	An account of the present war between the Vene...	Venice (Italy) -- History -- Turkish Wars, 17t...
A41527	Goodwin, Thomas, 1600-1680.	Patience and its perfect work under sudden & s...	Patience. Conduct of life.
A61867	Sanderson, Robert, 1587-1663.	Five cases of conscience occasionally determin...	Christian life -- Early works to 1800.
A61503	Sancroft, William, 1617-1693.	Lex ignea, or, The school of righteousness a s...	London (England) -- Fire, 1666 -- Sermons.

7. Minkowski Distance

In [47]:

```
minkowski_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='minkowski')), index

top5_minkowski= minkowski_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_minkowski)
m7=(metadata.loc[top5_minkowski.index, ['Author','Title','Keywords']])
m7
```

```
A62436     988.557029
A43020     988.622274
A29017    1000.024000
A56390    1005.630151
A44061    1012.873141
Name: A28989, dtype: float64
```

Out[47]:

	Author	Title	Keywords
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A43020	Harvey, Gideon, 1640?-1700?	Morbus anglicus: or, The anatomy of consumptio...	Tuberculosis -- Early works to 1800.
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A56390	Parker, Samuel, 1640-1688.	A free and impartial censure of the Platonick ...	Platonists. Empiricism -- Early works to 1800.
A44061	Hodges, Nathaniel, 1629-1688.	Vindiciæ medicinæ & medicorum: or An apology f...	Medicine -- Early works to 1800. Plague -- Eng...

8. Seuclidean distance

In [49]:

```
seuclidean_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='seuclidean')), ind

top5_seuclidean= seuclidean_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_seuclidean)
m8=(metadata.loc[top5_seuclidean.index, ['Author','Title','Keywords']])
m8
```

```
A62436    21.016932
A56390    21.391348
A39714    21.950013
A42820    22.034749
A66777    22.249879
Name: A28989, dtype: float64
```

Out[49]:

	Author	Title	Keywords
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A56390	Parker, Samuel, 1640-1688.	A free and impartial censure of the Platonick ...	Platonists. Empiricism -- Early works to 1800.
A39714	Flecknoe, Richard, d. 1678?	A farrago of several pieces being a supplement...	NaN
A42820	Glanvill, Joseph, 1636-1680.	A philosophical endeavour towards the defence ...	Witchcraft -- England -- Early works to 1800.
A66777	Wither, George, 1588-1667.	Sigh for the pitchers breathed out in a person...	Anglo-Dutch War, 1664-1667 -- Poetry. Great Br...

9. Sqeuclidean distance

In [50]:

```
sqeuclidean_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='sqeuclidean')), i

top5_sqeuclidean= sqeuclidean_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_sqeuclidean)
m9=(metadata.loc[top5_sqeuclidean.index, ['Author','Title','Keywords']])
m9
```

```
A62436     977245.0
A43020     977374.0
A29017    1000048.0
A56390    1011292.0
A44061    1025912.0
Name: A28989, dtype: float64
```

Out[50]:

	Author	Title	Keywords
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A43020	Harvey, Gideon, 1640?-1700?	Morbus anglicus: or, The anatomy of consumptio...	Tuberculosis -- Early works to 1800.
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A56390	Parker, Samuel, 1640-1688.	A free and impartial censure of the Platonick ...	Platonists. Empiricism -- Early works to 1800.
A44061	Hodges, Nathaniel, 1629-1688.	Vindiciæ medicinæ & medicorum: or An apology f...	Medicine -- Early works to 1800. Plague -- Eng...

10 Jensenshannon distance

In [51]:

```
jensenshannon_distances = pd.DataFrame(squareform(pdist(wordcounts, metric='jensenshannon')

top5_jensenshannon= jensenshannon_distances.nsmallest(6, 'A28989')['A28989'][1:]
print(top5_jensenshannon)
m5=(metadata.loc[top5_jensenshannon.index, ['Author','Title','Keywords']])
m5
```

```
A29017    0.353476
A62436    0.451909
A43020    0.455738
A57484    0.462194
A60482    0.470136
Name: A28989, dtype: float64
```

Out[51]:

	Author	Title	Keywords
A29017	Boyle, Robert, 1627-1691.	The origine of formes and qualities, (accordin...	Matter -- Constitution -- Early works to 1800....
A62436	Thomson, George, 17th cent.	Loimotomia, or, The pest anatomized in these f...	Hodges, Nathaniel, 1629-1688. -- Vindiciae med...
A43020	Harvey, Gideon, 1640?-1700?	Morbus anglicus: or, The anatomy of consumptio...	Tuberculosis -- Early works to 1800.
A57484	Rochefort, César de, b. 1605.	The history of the Caribby-islands, viz, Barba...	Natural history -- West Indies. Carib Indians.
A60482	Smith, John, 1630-1679.	Gērochomia vasilikē King Solomons portraitur...	Bible. -- O.T. -- Ecclesiastes XII, 1-6 -- Par...