Author Identification Writeup

Introduction:
The program uses the 3 different types of distancing.

Note: My program has very “different” outputs so all conclusions made in this should not be taken as fact or even remotely true. Everything here is based on the outputs I got from my program.

Manhattan:
$ui = u'/|u'|$ $vi = v'/|v'|$
$MD = \Sigma|ui - vi|$

Euclidean:
$ui = u'/|u'|$ $vi = v'/|v'|$
$MD = \Sigma|ui - vi|^2$

Cosine:
$ui = u'/|u'|$ $vi = v'/|v'|$
$MD = 1 - \Sigma|ui \bullet vi|$

For All Below the imputed file is ‘gildas.txt’

Euclidean:

Noise 100:	Noise 50:	Noise 0:
gildas	gildas	gildas
[0.000003249984275]	[0.000004040074145]	[0.000000812299447]
j.-m.-synge	j.-m.-synge	ellis
[0.035423047840595]	[0.041072659194469]	[0.021004948765039]
ellis	ellis	j.-m.-synge
[0.038063898682594]	[0.042131986469030]	[0.022783061489463]
Anonymous	Anonymous	Anonymous
[0.070288650691509]	[0.073764465749264]	[0.058673623949289]
wilhelm	wilhelm	wilhelm
[0.099032945930958]	[0.099910452961922]	[0.063665315508842]
wilhelm	wilhelm	wilhelm
[0.099032945930958]	[0.099910452961922]	[0.063665315508842]

By having the noise filtered there was a difference of around 0.06 to the 2nd closest, the one with the greatest distance was increased by 0.009. However the overall order still stayed in tact.
This is likely due to the most common words being in the top 50.
For noises 100, 50 and 0. The order is maintained throughout.

Manhattan:

Noise 100:	Noise 50:	Noise 0:
gildas [0.000136106740683]	gildas [0.000154846144142]	gildas [0.000022051390260]
j.-m.-synge [0.966570734977722]	j.-m.-synge [0.966570258140564]	j.-m.-synge [0.160988688468933]
ellis [0.997731328010559]	ellis [0.997730970382690]	ellis [0.207486808300018]
wilhelm [1.483721733093262]	wilhelm [1.483722209930420]	Anonymous [0.694013118743896]
wilhelm [1.483721733093262]	wilhelm [1.483722209930420]	wilhelm [0.824191212654114]
Anonymous [1.573667645454407]	Anonymous [1.573667764663696]	wilhelm [0.824191510677338]

Manhattan from noise 100 => 50 maintains its accuracy till the it 10e-5 but its order did not change. However, Manhattan disagrees with Euclidean on which has a the greatest distance. Anonymous is the greatest distance from gildas.txt while Euclidean says it is Willheim.
From noise 50 => 0 the ordering changes. Anonymous and Ellis swap places. This means that noise has some general improvement.

Cosine:

Noise 100:	Noise: 50	Noise 0:
gildas [0.999438643455505]	gildas [0.999132513999939]	gildas [0.999964952468872]
j.-m.-synge [0.999767959117889]	j.-m.-synge [0.999545574188232]	Anonymous [0.999999105930328]
ellis [0.999853670597076]	ellis [0.999702274799347]	ellis [0.999999642372131]
Anonymous [0.999922752380371]	Anonymous [0.999840915203094]	j.-m.-synge [0.999999761581421]
wilhelm [0.999979913234711]	wilhelm [0.999971628189087]	wilhelm [1.000000000000000]
wilhelm [0.999979913234711]	wilhelm [0.999971628189087]	wilhelm [1.000000000000000]

My cosine functionality is extremely problematic and the differences are all quite small at around 0.00001. The order is maintained from 100 => 50 noise but at 0 noise j.-m.-synge and Anonymous swap.

General Conclusion for regular-sized files:
Euclidean is the most successful at getting the correct ordering out of the 3 functions.
This could also mean that of all the methods, Euclidean is the least sensitive to noise which may mean that if there is a noise.txt available Euclidean is not a great choice for which method to use.

Testing Smaller Files:
The files below are made small, all between 200 and 500 words all shortened versions of other files in the texts folder.

Euclidean: Noise 100	Manhattan: Noise 100	Cosine: Noise 100
1) Ellis Parker [0.000422004901338]	1) Ellis Parker [0.000913013936952]	1) Ellis Parker [0.997612118721008]
2) Arnold bennett [0.040285766124725]	2) Arnold bennett [0.957270026206970]	2) Alice Freeman Parker [0.999800324440002]
3) Amelia E Barr [0.043298147618771]	3) Amelia E Barr [0.968503892421722]	3) Andy Adams [0.999821007251740]
4) Andy Adams [0.044752515852451]	4) Andy Adams [1.005141019821167]	4) Arnold bennett [0.999823868274689]
5) Aanonymous [0.047149285674095]	5) Aanonymous [1.063791871070862]	5) Amelia E Barr [0.999830007553101]
6) Alice Freeman Parker [0.056828815490007]	6) Alice Freeman Parker [1.072861552238464]	6) Aanonymous [0.999893784523010]
7) Andrew Lang [0.076803535223007]	7) Andrew Lang [1.136855602264404]	7) Andrew Lang [0.999959051609039]

The outputs clearly show when the file sizes are smaller there is a lot of conflict between the 3 methods. However all of them were able to match the original file.