Lucas Summers (lsumme01@calpoly.edu)
Xiuyuan Qiu (xiqiu@calpoly.edu)

CSC 466 Lab 6 Report

Experiments

For both of the two analytical methods, we vectorized the whole Reuters 50-50 document collection using TF-IDF keyword weighting schema. We used four different document preprocessing techniques when vectorizing to experiment with what tf-idf representations produce the best results:

1. No stopword removal, no stemming.
2. Stop word removal, but no stemming.
3. Stemming, but no stopword removal.
4. Both Stemming and stopword removal.

For stemming, we used the **PorterStemmer** from the **nltk** package and for the stopword list we also used **ntlk's** predefined list of english stopwords.

Unsupervised Learning

We ran sklearn's KMeans model on the TF-IDF matrixes from the 4 different representations, each using a cluster size of 50.

Supervised Learning

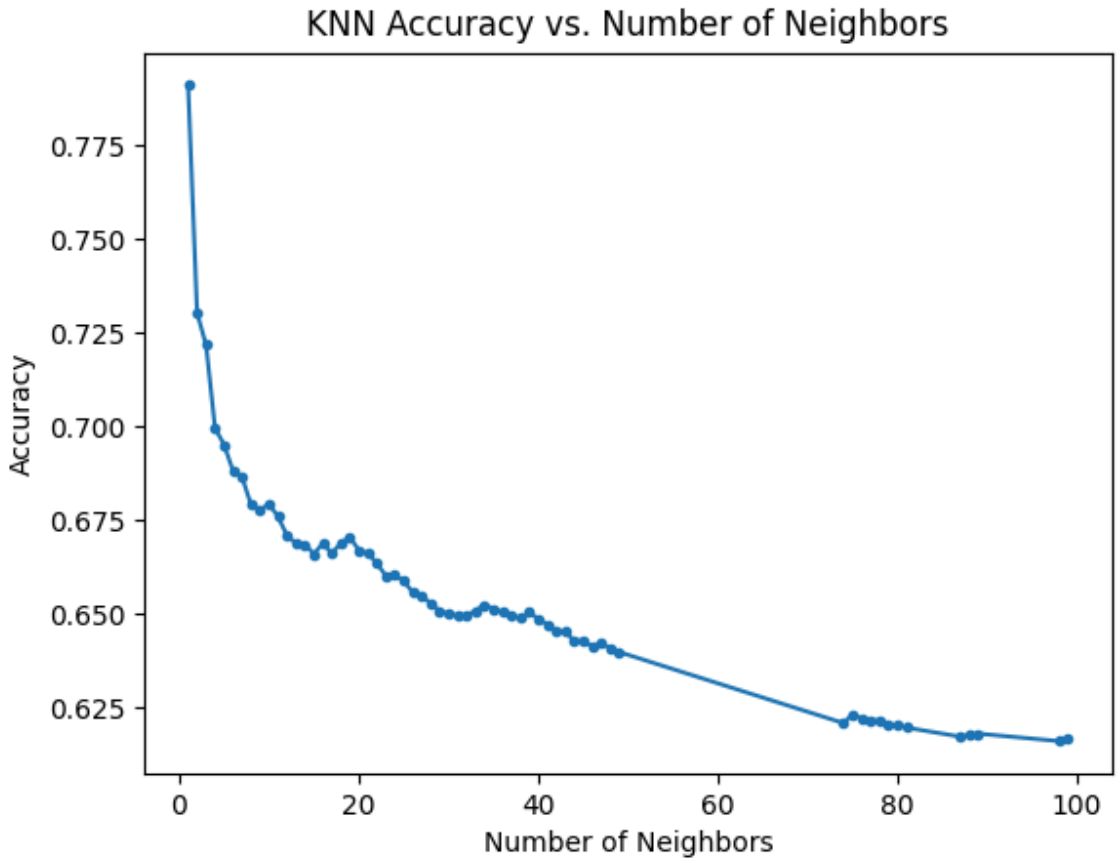
We ran sklearn's KNN model as a one-vs-world classifier (i.e. remove the target document from the dataset, train the model, and predict the target author using the trained model). Because of long grid search times, instead of running this on all 4 representations, we chose just stopword removal and stemming for document preprocessing as this produced the best results.

Results

KNN Classification

Cosine

The following grid search on N was done with cosine similarity, stemming, and stopword removal.



Turns out **n = 1** is best, with an accuracy of **0.7912**. There is a steep cliff at the first couple n, perhaps suggesting that authors may only write similarly to themselves occasionally (i.e. they write in varied styles), but in those instances where they do write similarly, it is really similar.

Top 5 authors by f1 score

MatthewBunce
precision, recall, f1: 0.94, 0.99, 0.97
KarlPenhaul
precision, recall, f1: 0.95, 0.95, 0.95
FumikoFujisaki
precision, recall, f1: 0.93, 0.96, 0.95
LynnleyBrowning
precision, recall, f1: 0.89, 0.96, 0.92
RogerFillion
precision, recall, f1: 0.90, 0.94, 0.92

Bottom 5 authors by f1 score

JaneMacartney
precision, recall, f1: 0.57, 0.63, 0.60
BenjaminKangLim
precision, recall, f1: 0.53, 0.62, 0.57
WilliamKazer
precision, recall, f1: 0.57, 0.46, 0.51
ScottHillis
precision, recall, f1: 0.45, 0.45, 0.45
MureDickie
precision, recall, f1: 0.47, 0.42, 0.44

Okapi

We were going to do another grid search with okapi, but the cluster we were using to run these evals (each run on *n* takes ~1 hour) was down for maintenance when we were ready with the okapi metric, so we only searched n from 1-3, given that the cosine similarity test already showed that smaller *n*'s are generally better (i.e. after a couple neighbors, it starts to just become which author is the most generic sounding).

In general, okapi was hard because it was slow and needed optimization, and the final implementation has us pre-computing the Okapi similarity for all 5000 documents during the pre-compute step. This had the downside of making the pre-compute step really long (20-30 minutes), but it saved time on the actual evaluation (ETAs went down to 1 hour instead of 8).

Okapi performed worse compared to cosine with best results at **n=1**, with an average accuracy of **0.6728** (n=2, 0.5872; n=3, 0.558). If it follows the trend from our cosine similarity grid search, it seems like it will not produce better results than cosine at larger n. We're not entirely sure why it's so much worse, there may be a bug in the implementation.

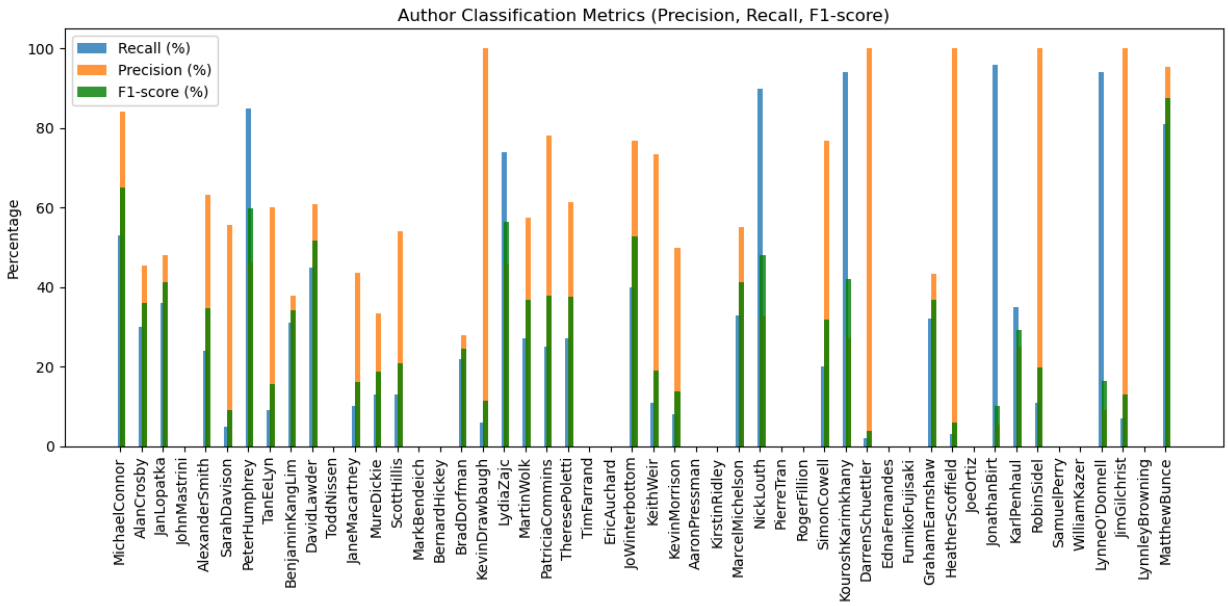
Summary

In general, knn on tf-idf with cosine similarity seems like a good method for determining authorship, resulting in near perfect f1 scores for an author in the best case and ~0.5 in the worst case, which is impressive given that there were fifty authors to choose from (i.e. baseline random is 0.02).

KMeans Clustering

No Stemming / No stopwords

* see below for tabular output



Average Cluster Purity: 0.64

Rand Score: 0.82

Top 5 Easiest Authors To Identify (F1):

MatthewBunce: 87.57%

MichaelConnor: 65.03%

PeterHumphrey: 59.86%

LydiaZajc: 56.49%

JoWinterbottom: 52.63%

Top 5 Hardest Authors To Identify (F1):

FumikoFujisaki: 0.00%

JoeOrtiz: 0.00%

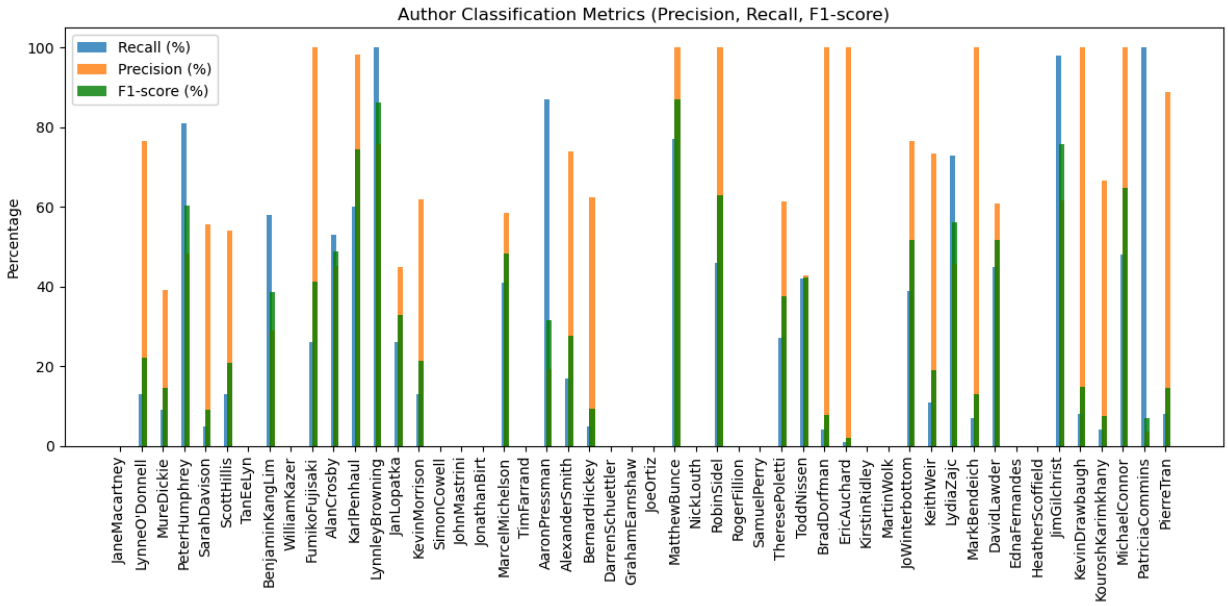
SamuelPerry: 0.00%

WilliamKazer: 0.00%

LynnleyBrowning: 0.00%

Stemming / No Stopwords

* see below for tabular output



Average Cluster Purity: 0.75

Rand Score: 0.68

Top 5 Easiest Authors To Identify (Highest F1-Score):

MatthewBunce: 87.01%

LynnleyBrowning: 86.21%

JimGilchrist: 75.68%

KarlPenhaul: 74.53%

MichaelConnor: 64.86%

Top 5 Hardest Authors To Identify (Lowest F1-Score):

SamuelPerry: 0.00%

KirstinRidley: 0.00%

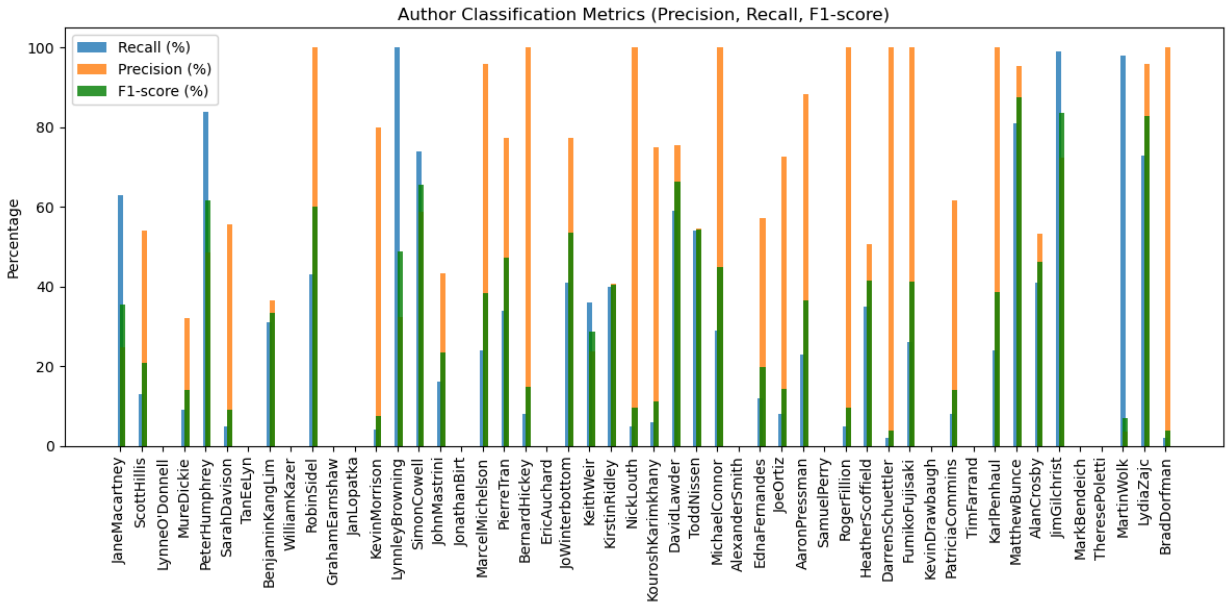
MartinWolk: 0.00%

EdnaFernandes: 0.00%

HeatherScoffield: 0.00%

No Stemming / Stopwords

* see below for tabular output



Average Cluster Purity: 0.73

Rand Score: 0.69

Top 5 Easiest Authors To Identify (Highest F1-Score):

MatthewBunce: 87.57%

JimGilchrist: 83.54%

LydiaZajc: 82.95%

DavidLawder: 66.29%

SimonCowell: 65.49%

Top 5 Hardest Authors To Identify (Lowest F1-Score):

SamuelPerry: 0.00%

KevinDrawbaugh: 0.00%

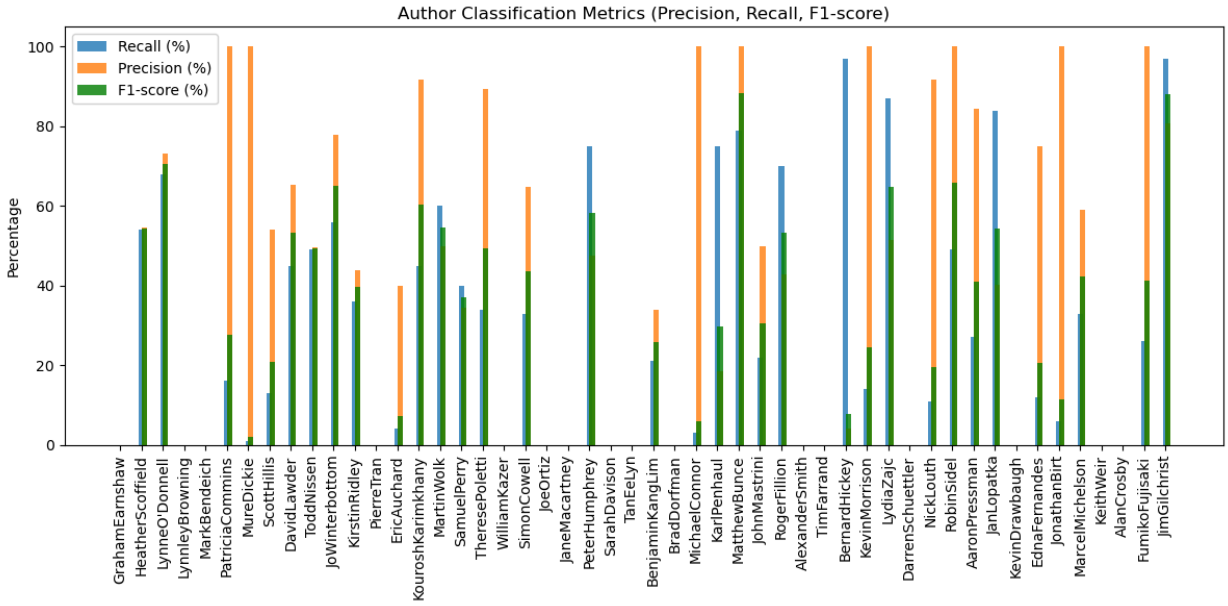
TimFarrand: 0.00%

MarkBendeich: 0.00%

TheresePoletti: 0.00%

Stemming / Stopwords

* see below for tabular output



Average Cluster Purity: 0.72

Rand Score: 0.75

Like KNN, KMeans performed the best (i.e. best in both metrics above) when preprocessing documents with both stemming and stopword removal. There is still a lot of crossover between the results of the other 3 results above. Authors like Matthew Bunce, Jim Gilchrist, and Jo Winterbottom are consistently in the top easiest to identify authors, suggesting they have very distinct writing styles to the other authors. Authors like Kevin Drawbaugh, Tim Farrand, and Samuel Taylor are consistently top hardest to identify authors, suggesting a lack of a unique style compared to the other authors or a lack of sufficient data, making them harder to cluster. It's interesting as well is how authors flip from being easy to identify into hard to identify when different preprocessing techniques are applied. For example, authors like Lynnley Browning go from hard to identify (0% for no stemming and stopwords) to easy to identify (86% for stemming and no stopwords). It seems like techniques like stemming present the documents in a different way that makes authors more distinguishable from each other.

Top 5 Easiest Authors To Identify (Highest F1-Score):

MatthewBunce: 88.27%

JimGilchrist: 88.18%

Top 5 Hardest Authors To Identify (Lowest F1-Score):

TimFarrand: 0.00%

DarrenSchuettler: 0.00%

LynneO'Donnell: 70.47%
RobinSidel: 65.77%
JoWinterbottom: 65.12%
KevinDrawbaugh: 0.00%
KeithWeir: 0.00%
AlanCrosby: 0.00%

Conclusion

Crossover best authors between two methods:

Top 5 authors (F1 Score):
MatthewBunce: KNN - 97%, KMeans - 88%
LynnleyBrowning: KNN - 92%, KMeans - 86%
Karl Penhaul: KNN - 95%, KMeans - 75%

Bottom 5 authors (F1 Score):
WilliamKazer: KNN - 51%, KMeans - 0%

Overall, we find that using supervised learning (KNN classification) for author attribution on the Reuters 50-50 dataset produced the best results. From the crossover between results above, you can see the KNN model with cosine similarity at n=1 consistently produces better F1 Scores over KMeans models. For the easiest to identify authors, the KMeans models are never able to break 90%, while the KNN model is consistently over 90%. For the hardest to identify authors, KMeans models are almost always 0%, while the KNN model still maintains a respectable 40-60%. We can draw conclusions from this that the classification approach, which leveraged known labels, made more accurate predictions even for less unique authors, whereas the clustering approach struggled more with assigning authors with similar styles to the correct groups without prior knowledge.

Addendum: Tabular Outputs

KNN (best run, n=1/cosine/stem+stop)

Author	Correct	False Positive	False Negative	Precision	Recall	F1
AaronPressman	85	19	15	0.8173076923	0.85	0.8333333333
AlanCrosby	90	18	10	0.8333333333	0.9	0.8653846154
AlexanderSmith	66	21	34	0.7586206897	0.66	0.7058823529
BenjaminKangLim	62	55	38	0.5299145299	0.62	0.5714285714
BernardHickey	71	21	29	0.7717391304	0.71	0.7395833333
BradDorfman	75	19	25	0.7978723404	0.75	0.7731958763
DarrenSchuettler	86	15	14	0.8514851485	0.86	0.855721393
DavidLawder	84	22	16	0.7924528302	0.84	0.8155339806
EdnaFernandes	89	13	11	0.8725490196	0.89	0.8811881188
EricAuchard	72	30	28	0.7058823529	0.72	0.7128712871
FumikoFujisaki	96	7	4	0.932038835	0.96	0.9458128079
GrahamEarnshaw	78	23	22	0.7722772277	0.78	0.776119403
HeatherScoffield	87	15	13	0.8529411765	0.87	0.8613861386
JanLopatka	83	18	17	0.8217821782	0.83	0.8258706468
JaneMacartney	63	48	37	0.5675675676	0.63	0.5971563981
JimGilchrist	90	7	10	0.9278350515	0.9	0.9137055838
JoWinterbottom	91	17	9	0.8425925926	0.91	0.875
JoeOrtiz	82	25	18	0.7663551402	0.82	0.7922705314
JohnMastrini	77	22	23	0.7777777778	0.77	0.7738693467
JonathanBirt	85	12	15	0.8762886598	0.85	0.8629441624
KarlPenhaul	95	5	5	0.95	0.95	0.95
KeithWeir	86	17	14	0.8349514563	0.86	0.8472906404
KevinDrawbaugh	77	9	23	0.8953488372	0.77	0.8279569892
KevinMorrison	76	32	24	0.7037037037	0.76	0.7307692308
KirstinRidley	76	23	24	0.7676767677	0.76	0.7638190955
KouroshKarimkhany	85	33	15	0.7203389831	0.85	0.7798165138
LydiaZajc	83	6	17	0.9325842697	0.83	0.8783068783
LynnleyBrowning	96	12	4	0.8888888889	0.96	0.9230769231
MarcelMichelson	83	12	17	0.8736842105	0.83	0.8512820513

MarkBendeich	83	21	17	0.7980769231	0.83	0.8137254902
MartinWolk	77	14	23	0.8461538462	0.77	0.8062827225
MatthewBunce	99	6	1	0.9428571429	0.99	0.9658536585
MichaelConnor	89	8	11	0.9175257732	0.89	0.9035532995
MureDickie	42	48	58	0.4666666667	0.42	0.4421052632
NickLouth	83	13	17	0.8645833333	0.83	0.8469387755
PatriciaCommins	82	11	18	0.8817204301	0.82	0.8497409326
PeterHumphrey	64	36	36	0.64	0.64	0.64
PierreTran	83	17	17	0.83	0.83	0.83
RobinSidel	88	6	12	0.9361702128	0.88	0.9072164948
RogerFillion	94	11	6	0.8952380952	0.94	0.9170731707
SamuelPerry	64	34	36	0.6530612245	0.64	0.6464646465
SarahDavison	67	19	33	0.7790697674	0.67	0.7204301075
ScottHillis	45	56	55	0.4455445545	0.45	0.447761194
SimonCowell	80	11	20	0.8791208791	0.8	0.8376963351
TanEeLyn	61	37	39	0.6224489796	0.61	0.6161616162
TheresePoletti	82	33	18	0.7130434783	0.82	0.7627906977
TimFarrand	85	18	15	0.8252427184	0.85	0.8374384236
ToddNissen	86	14	14	0.86	0.86	0.86
WilliamKazer	46	35	54	0.5679012346	0.46	0.5082872928

KMeans Clustering

No Stemming / No stopwords

	Total Clusters	Plurality Clusters	Recall	Precision	F1 Score
MichaelConnor	8	3	53	84.1269841269841	65.0306748466258
AlanCrosby	5	1	30	45.4545454545455	36.144578313253
JanLopatka	9	4	36	48	41.1428571428571
JohnMastrini	7	0	0	0	0
AlexanderSmith	10	2	24	63.1578947368421	34.7826086956522
SarahDavison	12	1	5	55.5555555555556	9.1743119266055
PeterHumphrey	8	3	85	46.1956521739131	59.8591549295775

TanEeLyn	10	1	9	60	15.6521739130435
BenjaminKangLim	12	3	31	37.8048780487805	34.0659340659341
DavidLawder	5	1	45	60.8108108108108	51.7241379310345
ToddNissen	4	0	0	0	0
JaneMacartney	11	1	10	43.4782608695652	16.260162601626
MureDickie	12	1	13	33.3333333333333	18.705035971223
ScottHillis	11	1	13	54.1666666666667	20.9677419354839
MarkBendeich	4	0	0	0	0
BernardHickey	4	0	0	0	0
BradDorfman	5	2	22	27.8481012658228	24.5810055865922
KevinDrawbaugh	6	1	6	100	11.3207547169811
LydiaZajc	5	1	74	45.679012345679	56.4885496183206
MartinWolk	6	1	27	57.4468085106383	36.734693877551
PatriciaCommins	5	2	25	78.125	37.8787878787879
TheresePoletti	5	1	27	61.3636363636364	37.5
TimFarrand	5	0	0	0	0
EricAuchard	5	0	0	0	0
JoWinterbottom	6	2	40	76.9230769230769	52.6315789473684
KeithWeir	4	1	11	73.3333333333333	19.1304347826087
KevinMorrison	7	1	8	50	13.7931034482759
AaronPressman	5	0	0	0	0
KirstinRidley	7	0	0	0	0
MarcelMichelson	5	1	33	55	41.25
NickLouth	3	1	90	32.7272727272727	48
PierreTran	4	0	0	0	0

RogerFillion	4	0	0	0	0
SimonCowell	7	1	20	76.9230769230769	31.7460317460317
KouroshKarimkhany	4	2	94	27.0893371757925	42.0581655480984
DarrenSchuettler	4	1	2	100	3.92156862745098
EdnaFernandes	4	0	0	0	0
FumikoFujisaki	2	0	0	0	0
GrahamEarnshaw	4	1	32	43.2432432432432	36.7816091954023
HeatherScoffield	5	1	3	100	5.8252427184466
JoeOrtiz	6	0	0	0	0
JonathanBirt	3	1	96	5.40540540540541	10.2345415778252
KarlPenhaul	4	1	35	25	29.1666666666667
RobinSidel	3	1	11	100	19.8198198198198
SamuelPerry	4	0	0	0	0
WilliamKazer	10	0	0	0	0
LynneO'Donnell	4	1	94	8.96091515729266	16.3620539599652
JimGilchrist	4	2	7	100	13.0841121495327
LynnleyBrowning	2	0	0	0	0
MatthewBunce	3	2	81	95.2941176470588	87.5675675675676

Stemming / No Stopwords

	Total Clusters	Plurality Clusters	Recall	Precision	F1 Score
JaneMacartney	7	0	0	0	0
LynneO'Donnell	7	1	13	76.4705882352941	22.2222222222222
MureDickie	7	1	9	39.1304347826087	14.6341463414634
PeterHumphrey	4	1	81	48.2142857142857	60.4477611940298

SarahDavison	7	1	5	55.5555555555556	9.1743119266055
ScottHillis	8	1	13	54.1666666666667	20.9677419354839
TanEeLyn	6	0	0	0	0
BenjaminKangLim	7	1	58	29	38.6666666666667
WilliamKazer	8	0	0	0	0
FumikoFujisaki	3	1	26	100	41.2698412698413
AlanCrosby	7	3	53	45.2991452991453	48.8479262672811
KarlPenhaul	2	1	60	98.3606557377049	74.5341614906832
LynnleyBrowning	2	2	100	75.7575757575758	86.2068965517241
JanLopatka	6	2	26	44.8275862068966	32.9113924050633
KevinMorrison	6	2	13	61.9047619047619	21.4876033057851
SimonCowell	4	0	0	0	0
JohnMastrini	6	0	0	0	0
JonathanBirt	3	0	0	0	0
MarcelMichelson	7	3	41	58.5714285714286	48.2352941176471
TimFarrand	2	0	0	0	0
AaronPressman	4	2	87	19.3333333333333	31.6363636363636
AlexanderSmith	9	3	17	73.9130434782609	27.6422764227642
BernardHickey	4	1	5	62.5	9.25925925925926
DarrenSchuettler	4	0	0	0	0
GrahamEarnshaw	5	0	0	0	0
JoeOrtiz	6	0	0	0	0
MatthewBunce	3	1	77	100	87.0056497175141
NickLouth	2	0	0	0	0
RobinSidel	5	2	46	100	63.013698630137

RogerFillion	2	0	0	0	0
SamuelPerry	3	0	0	0	0
TheresePoletti	3	1	27	61.3636363636364	37.5
ToddNissen	4	1	42	42.8571428571429	42.4242424242424
BradDorfman	6	2	4	100	7.69230769230769
EricAuchard	4	1	1	100	1.98019801980198
KirstinRidley	4	0	0	0	0
MartinWolk	3	0	0	0	0
JoWinterbottom	4	2	39	76.4705882352941	51.6556291390728
KeithWeir	2	1	11	73.3333333333333	19.1304347826087
LydiaZajc	3	2	73	45.625	56.1538461538462
MarkBendeich	4	2	7	100	13.0841121495327
DavidLawder	3	1	45	60.8108108108108	51.7241379310345
EdnaFernandes	4	0	0	0	0
HeatherScoffield	2	0	0	0	0
JimGilchrist	2	1	98	61.6352201257862	75.6756756756757
KevinDrawbaugh	4	1	8	100	14.8148148148148
KouroshKarimkhany	3	2	4	66.6666666666667	7.54716981132076
MichaelConnor	5	2	48	100	64.8648648648649
PatriciaCommins	1	1	100	3.58937544867193	6.93000693000693
PierreTran	5	1	8	88.8888888888889	14.6788990825688

No Stemming / Stopwords

	Total Clusters	Plurality Clusters	Recall	Precision	F1 Score
JaneMacartney	9	3	63	24.8031496062992	35.5932203389831

ScottHillis	10	1	13	54.1666666666667	20.9677419354839
LynneO'Donnell	7	0	0	0	0
MureDickie	9	1	9	32.1428571428571	14.0625
PeterHumphrey	7	1	84	48.5549132947977	61.5384615384615
SarahDavison	8	1	5	55.5555555555556	9.1743119266055
TanEeLyn	8	0	0	0	0
BenjaminKangLim	9	2	31	36.4705882352941	33.5135135135135
WilliamKazer	8	0	0	0	0
RobinSidel	8	4	43	100	60.1398601398602
GrahamEarnshaw	6	0	0	0	0
JanLopatka	4	0	0	0	0
KevinMorrison	6	1	4	80	7.61904761904762
LynnleyBrowning	1	1	100	32.258064516129	48.7804878048781
SimonCowell	5	1	74	58.7301587301587	65.4867256637168
JohnMastrini	5	1	16	43.2432432432432	23.3576642335766
JonathanBirt	5	0	0	0	0
MarcelMichelson	8	1	24	96	38.4
PierreTran	6	2	34	77.2727272727273	47.2222222222222
BernardHickey	4	1	8	100	14.8148148148148
EricAuchard	4	0	0	0	0
JoWinterbottom	5	1	41	77.3584905660377	53.5947712418301
KeithWeir	3	1	36	23.6842105263158	28.5714285714286
KirstinRidley	6	1	40	40.8163265306122	40.4040404040404
NickLouth	4	1	5	100	9.52380952380952
KouroshKarimkhany	3	1	6	75	11.1111111111111

DavidLawder	5	3	59	75.6410256410256	66.2921348314607
ToddNissen	4	2	54	54.5454545454545	54.2713567839196
MichaelConnor	5	1	29	100	44.9612403100775
AlexanderSmith	6	0	0	0	0
EdnaFernandes	6	1	12	57.1428571428571	19.8347107438017
JoeOrtiz	6	1	8	72.7272727272727	14.4144144144144
AaronPressman	4	1	23	88.4615384615385	36.5079365079365
SamuelPerry	2	0	0	0	0
RogerFillion	5	1	5	100	9.52380952380952
HeatherScoffield	6	1	35	50.7246376811594	41.4201183431953
DarrenSchuettler	4	1	2	100	3.92156862745098
FumikoFujisaki	3	1	26	100	41.2698412698413
KevinDrawbaugh	5	0	0	0	0
PatriciaCommins	3	1	8	61.5384615384615	14.1592920353982
TimFarrand	3	0	0	0	0
KarlPenhaul	3	1	24	100	38.7096774193548
MatthewBunce	3	2	81	95.2941176470588	87.5675675675676
AlanCrosby	4	2	41	53.2467532467532	46.3276836158192
JimGilchrist	2	1	99	72.2627737226277	83.5443037974684
MarkBendeich	3	0	0	0	0
TheresePoletti	2	0	0	0	0
MartinWolk	2	1	98	3.58580314672521	6.91846099541123
LydiaZajc	3	2	73	96.0526315789474	82.9545454545455
BradDorfman	4	1	2	100	3.92156862745098

Stemming / Stopwords

	Total Clusters	Plurality Clusters	Recall	Precision	F1 Score
GrahamEarnshaw	5	0	0	0	0
HeatherScoffield	6	2	54	54.5454545454545	54.2713567839196
LynneO'Donnell	6	2	68	73.1182795698925	70.4663212435233
LynnleyBrowning	2	0	0	0	0
MarkBendeich	3	0	0	0	0
PatriciaCommins	4	1	16	100	27.5862068965517
MureDickie	9	1	1	100	1.980198019801980198
ScottHillis	7	1	13	54.1666666666667	20.9677419354839
DavidLawder	4	1	45	65.2173913043478	53.2544378698225
ToddNissen	4	2	49	49.4949494949495	49.2462311557789
JoWinterbottom	5	2	56	77.7777777777778	65.1162790697674
KirstinRidley	7	1	36	43.9024390243903	39.5604395604396
PierreTran	6	0	0	0	0
EricAuchard	10	1	4	40	7.27272727272727
KouroshKarimkhany	5	1	45	91.8367346938776	60.4026845637584
MartinWolk	5	2	60	50	54.5454545454545
SamuelPerry	5	1	40	34.4827586206897	37.037037037037
TheresePoletti	6	1	34	89.4736842105263	49.2753623188406
WilliamKazer	7	0	0	0	0
SimonCowell	3	2	33	64.7058823529412	43.7086092715232
JoeOrtiz	5	0	0	0	0
JaneMacartney	6	0	0	0	0
PeterHumphrey	5	1	75	47.4683544303798	58.1395348837209
SarahDavison	4	0	0	0	0

TanEeLyn	5	0	0	0	0
BenjaminKangLim	7	2	21	33.8709677419355	25.9259259259259
BradDorfman	4	0	0	0	0
MichaelConnor	4	1	3	100	5.8252427184466
KarlPenhaul	4	2	75	18.5185185185185	29.7029702970297
MatthewBunce	3	1	79	100	88.268156424581
JohnMastrini	6	2	22	50	30.5555555555556
RogerFillion	5	1	70	42.9447852760736	53.2319391634981
AlexanderSmith	5	0	0	0	0
TimFarrand	3	0	0	0	0
BernardHickey	3	1	97	3.97867104183757	7.64381402679275
KevinMorrison	5	2	14	100	24.5614035087719
LydiaZajc	4	2	87	51.4792899408284	64.6840148698885
DarrenSchuettler	3	0	0	0	0
NickLouth	4	1	11	91.6666666666667	19.6428571428571
RobinSidel	5	3	49	100	65.7718120805369
AaronPressman	4	2	27	84.375	40.9090909090909
JanLopatka	4	2	84	40.1913875598086	54.368932038835
KevinDrawbaugh	4	0	0	0	0
EdnaFernandes	5	2	12	75	20.6896551724138
JonathanBirt	4	1	6	100	11.3207547169811
MarcelMichelson	5	1	33	58.9285714285714	42.3076923076923
KeithWeir	2	0	0	0	0
AlanCrosby	3	0	0	0	0
FumikoFujisaki	2	1	26	100	41.2698412698413

JimGilchrist	2	1	97	80.8333333333333	88.1818181818182