# Lab 4: Information Retrieval

Vittorio Dinovi - vdinovi@calpoly.edu

Billy Gottenstrater - wgottens@calpoly.edu

## Notes

- Okapi distance was implemented but not considered in this report because its performance was consistently subpar compared to cosine similarity

## Vectorization

Note: all cases receive a basic filter for invalid characters

| **C50Train** | No filter | stopwords | stemming | stopwords & stemming |
|---|---|---|---|---|
| Vocab Size | 52066 | 51523 | 43065 | 42732 |

| **C50Test** | No filter | stopwords | stemming | stopwords & stemming |
|---|---|---|---|---|
| Vocab Size | 53488 | 52949 | 44224 | 43895 |

stopwords used - stopwords-long.txt
stemmer used - Porter's Stemmer

Vectors are stored in the following *sparse* format

```
word1,#docs,word2,#docs,...
author,docid,doc_word1,freq,doc_word2,freq,...
…
```

# KNN Performance

Observed overall accuracies for KNN using **cosine-sim** and **k=5**

| **Overall Accuracy** | No filter | Stopwords | Stemming | Stopwords & Stemming |
|---|---|---|---|---|
| C50train | 0.4644 | 0.6156 | 0.4436 | 0.5780 |
| C50test | 0.4676 | 0.6240 | 0.4476 | 0.5756 |

→ Best choice: **only stopwords**

| **K** | **C50Train Accuracy** | **C50Test Accuracy** |
|---|---|---|
| 1 | 0.6560 | 0.6716 |
| 3 | 0.6144 | 0.6272 |
| 5 | 0.5208 | 0.6352 |
| 7 | 0.5988 | 0.6304 |
| 9 | 0.6040 | 0.6207 |
| 11 | 0.5996 | 0.6212 |
| 13 | 0.5988 | 0.6156 |
| 15 | 0.5984 | 0.6104 |
| 17 | 0.5876 | 0.6068 |
| 19 | 0.5864 | 0.6164 |
| 21 | 0.5812 | 0.6184 |
| 23 | 0.5760 | 0.6064 |
| 25 | 0.5674 | 0.6096 |

→ Best choice: **k=1**

This seems a bit strange, though I suppose every additional k increases the likelihood of a misclassification. Though increasing k *should* provide the benefit of balancing out errors when the first k is wrong.

# Best Results

## C50Train: Cosine-Sim, k=1, stopword only

```
-> reading ground truths from  ../C50train_stop/truths.out
-> reading results from  best.out
-- Evaluation Results --
Total Correct:    1640
Total Incorrect:  860
Overall Accuracy: 0.6560
Authors:
  - GrahamEarnshaw:
      hits = 20,  strikes = 7,  misses = 30
      precision = 0.7407,  recall = 0.4000,  f-measure = 0.5195
  - KirstinRidley:
      hits = 28,  strikes = 7,  misses = 22
      precision = 0.7742,  recall = 0.4800,  f-measure = 0.5926
  - KevinDrawbaugh:
      hits = 30,  strikes = 9,  misses = 20
      precision = 0.7723,  recall = 0.5200,  f-measure = 0.6215
  - FumikoFujisaki:
      hits = 37,  strikes = 0,  misses = 13
      precision = 0.8333,  recall = 0.5750,  f-measure = 0.6805
  - MartinWolk:
      hits = 43,  strikes = 35,  misses = 7
      precision = 0.7315,  recall = 0.6320,  f-measure = 0.6781
  - SamuelPerry:
      hits = 34,  strikes = 21,  misses = 16
      precision = 0.7085,  recall = 0.6400,  f-measure = 0.6725
  - WilliamKazer:
      hits = 10,  strikes = 46,  misses = 40
      precision = 0.6177,  recall = 0.5771,  f-measure = 0.5968
  - MichaelConnor:
      hits = 41,  strikes = 12,  misses = 9
      precision = 0.6395,  recall = 0.6075,  f-measure = 0.6231
  - PatriciaCommins:
      hits = 39,  strikes = 6,  misses = 11
      precision = 0.6635,  recall = 0.6267,  f-measure = 0.6446
  - JoeOrtiz:
      hits = 36,  strikes = 13,  misses = 14
      precision = 0.6709,  recall = 0.6360,  f-measure = 0.6530
```

- LynnleyBrowning:
  hits = 50,  strikes = 44,  misses = 0
  precision = 0.6479,  recall = 0.6691,  f-measure = 0.6583
- JaneMacartney:
  hits = 9,  strikes = 4,  misses = 41
  precision = 0.6489,  recall = 0.6283,  f-measure = 0.6384
- MatthewBunce:
  hits = 29,  strikes = 0,  misses = 21
  precision = 0.6656,  recall = 0.6246,  f-measure = 0.6444
- TanEeLyn:
  hits = 29,  strikes = 121,  misses = 21
  precision = 0.5724,  recall = 0.6214,  f-measure = 0.5959
- LydiaZajc:
  hits = 41,  strikes = 1,  misses = 9
  precision = 0.5935,  recall = 0.6347,  f-measure = 0.6134
- LynneO'Donnell:
  hits = 45,  strikes = 55,  misses = 5
  precision = 0.5776,  recall = 0.6512,  f-measure = 0.6122
- AlanCrosby:
  hits = 34,  strikes = 4,  misses = 16
  precision = 0.5904,  recall = 0.6529,  f-measure = 0.6201
- JonathanBirt:
  hits = 38,  strikes = 31,  misses = 12
  precision = 0.5877,  recall = 0.6589,  f-measure = 0.6213
- BenjaminKangLim:
  hits = 17,  strikes = 16,  misses = 33
  precision = 0.5854,  recall = 0.6421,  f-measure = 0.6124
- TheresePoletti:
  hits = 27,  strikes = 7,  misses = 23
  precision = 0.5920,  recall = 0.6370,  f-measure = 0.6137
- PeterHumphrey:
  hits = 0,  strikes = 6,  misses = 50
  precision = 0.5887,  recall = 0.6067,  f-measure = 0.5976
- MureDickie:
  hits = 8,  strikes = 43,  misses = 42
  precision = 0.5693,  recall = 0.5864,  f-measure = 0.5777
- TimFarrand:
  hits = 35,  strikes = 5,  misses = 15
  precision = 0.5797,  recall = 0.5913,  f-measure = 0.5854
- SarahDavison:
  hits = 14,  strikes = 2,  misses = 36
  precision = 0.5837,  recall = 0.5783,  f-measure = 0.5810
- HeatherScoffield:
  hits = 40,  strikes = 23,  misses = 10
  precision = 0.5863,  recall = 0.5872,  f-measure = 0.5867
- KeithWeir:
  hits = 35,  strikes = 13,  misses = 15

```
    precision = 0.5915,  recall = 0.5915,  f-measure = 0.5915
- JimGilchrist:
    hits = 18,  strikes = 3,  misses = 32
    precision = 0.5958,  recall = 0.5830,  f-measure = 0.5893
- MarkBendeich:
    hits = 30,  strikes = 4,  misses = 20
    precision = 0.6030,  recall = 0.5836,  f-measure = 0.5931
- SimonCowell:
    hits = 34,  strikes = 6,  misses = 16
    precision = 0.6100,  recall = 0.5869,  f-measure = 0.5982
- AaronPressman:
    hits = 43,  strikes = 24,  misses = 7
    precision = 0.6115,  recall = 0.5960,  f-measure = 0.6036
- NickLouth:
    hits = 43,  strikes = 6,  misses = 7
    precision = 0.6201,  recall = 0.6045,  f-measure = 0.6122
- RogerFillion:
    hits = 42,  strikes = 2,  misses = 8
    precision = 0.6296,  recall = 0.6119,  f-measure = 0.6206
- DarrenSchuettler:
    hits = 46,  strikes = 25,  misses = 4
    precision = 0.6304,  recall = 0.6212,  f-measure = 0.6258
- BradDorfman:
    hits = 26,  strikes = 9,  misses = 24
    precision = 0.6328,  recall = 0.6182,  f-measure = 0.6254
- ToddNissen:
    hits = 36,  strikes = 8,  misses = 14
    precision = 0.6375,  recall = 0.6211,  f-measure = 0.6292
- AlexanderSmith:
    hits = 31,  strikes = 11,  misses = 19
    precision = 0.6400,  recall = 0.6211,  f-measure = 0.6304
- JanLopatka:
    hits = 36,  strikes = 18,  misses = 14
    precision = 0.6408,  recall = 0.6238,  f-measure = 0.6322
- ScottHillis:
    hits = 26,  strikes = 69,  misses = 24
    precision = 0.6224,  recall = 0.6211,  f-measure = 0.6217
- PierreTran:
    hits = 32,  strikes = 9,  misses = 18
    precision = 0.6257,  recall = 0.6215,  f-measure = 0.6236
- MarcelMichelson:
    hits = 43,  strikes = 11,  misses = 7
    precision = 0.6303,  recall = 0.6275,  f-measure = 0.6289
- KarlPenhaul:
    hits = 34,  strikes = 4,  misses = 16
    precision = 0.6353,  recall = 0.6288,  f-measure = 0.6320
- EricAuchard:
```

```
    hits = 41,  strikes = 29,  misses = 9
    precision = 0.6336,  recall = 0.6333,  f-measure = 0.6335
  - KouroshKarimkhany:
    hits = 41,  strikes = 19,  misses = 9
    precision = 0.6350,  recall = 0.6377,  f-measure = 0.6363
  - EdnaFernandes:
    hits = 43,  strikes = 17,  misses = 7
    precision = 0.6372,  recall = 0.6427,  f-measure = 0.6400
  - RobinSidel:
    hits = 47,  strikes = 3,  misses = 3
    precision = 0.6439,  recall = 0.6493,  f-measure = 0.6466
  - JoWinterbottom:
    hits = 23,  strikes = 5,  misses = 27
    precision = 0.6461,  recall = 0.6452,  f-measure = 0.6456
  - JohnMastrini:
    hits = 42,  strikes = 18,  misses = 8
    precision = 0.6474,  recall = 0.6494,  f-measure = 0.6484
  - DavidLawder:
    hits = 49,  strikes = 13,  misses = 1
    precision = 0.6511,  recall = 0.6562,  f-measure = 0.6537
  - BernardHickey:
    hits = 32,  strikes = 5,  misses = 18
    precision = 0.6543,  recall = 0.6559,  f-measure = 0.6551
  - KevinMorrison:
    hits = 33,  strikes = 11,  misses = 17
    precision = 0.6560,  recall = 0.6560,  f-measure = 0.6560
-> writing confusion matrix to  best_conf.cs
```

## C50Test: Cosine-Sim, k=1, stopword only

```
-> reading ground truths from  ../C50test_stop/truths.out
-> reading results from  best.out
-- Evaluation Results --
Total Correct:    1679
Total Incorrect:  821
Overall Accuracy: 0.6716
Authors:
  - LydiaZajc:
    hits = 9,  strikes = 2,  misses = 41
    precision = 0.8182,  recall = 0.1800,  f-measure = 0.2951
  - KirstinRidley:
    hits = 32,  strikes = 8,  misses = 18
    precision = 0.8039,  recall = 0.4100,  f-measure = 0.5430
  - AlexanderSmith:
    hits = 20,  strikes = 6,  misses = 30
```

```
    precision = 0.7922,  recall = 0.4067,  f-measure = 0.5374
- SarahDavison:
    hits = 6,  strikes = 6,  misses = 44
    precision = 0.7528,  recall = 0.3350,  f-measure = 0.4637
- JimGilchrist:
    hits = 30,  strikes = 2,  misses = 20
    precision = 0.8017,  recall = 0.3880,  f-measure = 0.5229
- SamuelPerry:
    hits = 36,  strikes = 53,  misses = 14
    precision = 0.6333,  recall = 0.4433,  f-measure = 0.5216
- MartinWolk:
    hits = 29,  strikes = 16,  misses = 21
    precision = 0.6353,  recall = 0.4629,  f-measure = 0.5355
- TheresePoletti:
    hits = 42,  strikes = 8,  misses = 8
    precision = 0.6689,  recall = 0.5100,  f-measure = 0.5787
- KarlPenhaul:
    hits = 50,  strikes = 47,  misses = 0
    precision = 0.6318,  recall = 0.5644,  f-measure = 0.5962
- DavidLawder:
    hits = 33,  strikes = 37,  misses = 17
    precision = 0.6081,  recall = 0.5740,  f-measure = 0.5905
- RogerFillion:
    hits = 40,  strikes = 4,  misses = 10
    precision = 0.6337,  recall = 0.5945,  f-measure = 0.6135
- MichaelConnor:
    hits = 47,  strikes = 5,  misses = 3
    precision = 0.6585,  recall = 0.6233,  f-measure = 0.6404
- WilliamKazer:
    hits = 16,  strikes = 50,  misses = 34
    precision = 0.6151,  recall = 0.6000,  f-measure = 0.6075
- KevinMorrison:
    hits = 40,  strikes = 23,  misses = 10
    precision = 0.6169,  recall = 0.6143,  f-measure = 0.6156
- AlanCrosby:
    hits = 33,  strikes = 6,  misses = 17
    precision = 0.6291,  recall = 0.6173,  f-measure = 0.6231
- ToddNissen:
    hits = 27,  strikes = 3,  misses = 23
    precision = 0.6397,  recall = 0.6125,  f-measure = 0.6258
- EdnaFernandes:
    hits = 41,  strikes = 31,  misses = 9
    precision = 0.6337,  recall = 0.6247,  f-measure = 0.6291
- JoWinterbottom:
    hits = 43,  strikes = 9,  misses = 7
    precision = 0.6449,  recall = 0.6378,  f-measure = 0.6413
- PatriciaCommins:
```

hits = 33, strikes = 3, misses = 17
precision = 0.6555, recall = 0.6389, f-measure = 0.6471
- LynnleyBrowning:
hits = 48, strikes = 9, misses = 2
precision = 0.6663, recall = 0.6550, f-measure = 0.6606
- MatthewBunce:
hits = 40, strikes = 0, misses = 10
precision = 0.6794, recall = 0.6619, f-measure = 0.6705
- AaronPressman:
hits = 41, strikes = 13, misses = 9
precision = 0.6834, recall = 0.6691, f-measure = 0.6762
- JonathanBirt:
hits = 33, strikes = 1, misses = 17
precision = 0.6922, recall = 0.6687, f-measure = 0.6802
- TimFarrand:
hits = 37, strikes = 8, misses = 13
precision = 0.6972, recall = 0.6717, f-measure = 0.6842
- MarcelMichelson:
hits = 39, strikes = 21, misses = 11
precision = 0.6949, recall = 0.6760, f-measure = 0.6853
- DarrenSchuettler:
hits = 47, strikes = 45, misses = 3
precision = 0.6820, recall = 0.6862, f-measure = 0.6840
- JanLopatka:
hits = 42, strikes = 12, misses = 8
precision = 0.6858, recall = 0.6919, f-measure = 0.6888
- GrahamEarnshaw:
hits = 33, strikes = 8, misses = 17
precision = 0.6892, recall = 0.6907, f-measure = 0.6900
- NickLouth:
hits = 21, strikes = 4, misses = 29
precision = 0.6919, recall = 0.6814, f-measure = 0.6866
- JoeOrtiz:
hits = 42, strikes = 17, misses = 8
precision = 0.6927, recall = 0.6867, f-measure = 0.6897
- ScottHillis:
hits = 30, strikes = 14, misses = 20
precision = 0.6924, recall = 0.6839, f-measure = 0.6881
- TanEeLyn:
hits = 41, strikes = 46, misses = 9
precision = 0.6805, recall = 0.6881, f-measure = 0.6843
- EricAuchard:
hits = 32, strikes = 41, misses = 18
precision = 0.6700, recall = 0.6867, f-measure = 0.6782
- JaneMacartney:
hits = 7, strikes = 18, misses = 43
precision = 0.6643, recall = 0.6706, f-measure = 0.6674

- LynneO'Donnell:
  hits = 38,  strikes = 15,  misses = 12
  precision = 0.6659,  recall = 0.6731,  f-measure = 0.6695
- RobinSidel:
  hits = 45,  strikes = 1,  misses = 5
  precision = 0.6738,  recall = 0.6794,  f-measure = 0.6766
- BenjaminKangLim:
  hits = 25,  strikes = 32,  misses = 25
  precision = 0.6667,  recall = 0.6746,  f-measure = 0.6706
- SimonCowell:
  hits = 39,  strikes = 3,  misses = 11
  precision = 0.6724,  recall = 0.6774,  f-measure = 0.6749
- MureDickie:
  hits = 22,  strikes = 20,  misses = 28
  precision = 0.6692,  recall = 0.6713,  f-measure = 0.6703
- MarkBendeich:
  hits = 36,  strikes = 9,  misses = 14
  precision = 0.6722,  recall = 0.6725,  f-measure = 0.6723
- BernardHickey:
  hits = 31,  strikes = 4,  misses = 19
  precision = 0.6758,  recall = 0.6712,  f-measure = 0.6735
- KouroshKarimkhany:
  hits = 44,  strikes = 56,  misses = 6
  precision = 0.6648,  recall = 0.6762,  f-measure = 0.6704
- PeterHumphrey:
  hits = 23,  strikes = 43,  misses = 27
  precision = 0.6553,  recall = 0.6712,  f-measure = 0.6631
- PierreTran:
  hits = 24,  strikes = 5,  misses = 26
  precision = 0.6576,  recall = 0.6668,  f-measure = 0.6622
- HeatherScoffield:
  hits = 27,  strikes = 5,  misses = 23
  precision = 0.6602,  recall = 0.6640,  f-measure = 0.6621
- KevinDrawbaugh:
  hits = 27,  strikes = 1,  misses = 23
  precision = 0.6639,  recall = 0.6613,  f-measure = 0.6626
- KeithWeir:
  hits = 38,  strikes = 11,  misses = 12
  precision = 0.6662,  recall = 0.6634,  f-measure = 0.6648
- FumikoFujisaki:
  hits = 46,  strikes = 21,  misses = 4
  precision = 0.6668,  recall = 0.6687,  f-measure = 0.6678
- BradDorfman:
  hits = 36,  strikes = 6,  misses = 14
  precision = 0.6701,  recall = 0.6698,  f-measure = 0.6699
- JohnMastrini:
  hits = 38,  strikes = 13,  misses = 12

```
     precision = 0.6716,  recall = 0.6716,  f-measure = 0.6716
-> writing confusion matrix to  best_conf.csv
```

# Clustering Performance

Clustering did not produce great results. When using single link, the results were the worst, involving very few merges between large clusters, and mostly merges between a large cluster and one point. To make this better, we implemented a priority queue  We also specified a parameter, *min_size*.

**C50test**

| **Clusters found** | Single Link | Average Link | Complete Link |
|---|---|---|---|
| min_size=2 | 50 | 50 | 50 |
| min_size=5 | 15 | 50 | 50 |
| min_size=10 | 7 | 50 | 50 |
| min_size=20 | 6 | 50 | 50 |
| min_size=35 | 4 | 33 | 47 |
| min_size=50 | 2 | 25 | 31 |

**C50train**

For the training set, we only performed single link.

| **Clusters found** | Single Link |
|---|---|
| min_size=2 | 45 |
| min_size=5 | 17 |
| min_size=10 | 11 |
| min_size=20 | 7 |
| min_size=35 | 6 |
| min_size=50 | 5 |

The best results are from C50test with complete link. So, we will use that for our analysis.

# Best Results

## C50test: Complete Link, min_size=20, priority queue used

```
-> loading dendrogram from  ../cluster_results/C50test_cluster_complete/dendrogram.json
-> reading ground truths from  ../cluster_results/C50test_cluster_complete/truths.out
-- Evaluation Results --
Total correct:    171
Total incorrect:  2329
Overall accuracy: 0.0684
Authors:
  - MureDickie
      hits = 3, strikes = 44, misses = 47
      precision =   0.0600, recall =   0.0638, f-measure =   0.0619
  - EricAuchard
      hits = 9, strikes = 100, misses = 41
      precision =   0.1800, recall =   0.0826, f-measure =   0.1132
  - MichaelConnor
      hits = 7, strikes = 97, misses = 43
      precision =   0.1400, recall =   0.0673, f-measure =   0.0909
  - SamuelPerry
      hits = 4, strikes = 66, misses = 46
      precision =   0.0800, recall =   0.0571, f-measure =   0.0667
  - NickLouth
      hits = 0, strikes = 0, misses = 50
      precision =   0.0000, recall =   0.0000, f-measure =   0.0000
  - KouroshKarimkhany
      hits = 6, strikes = 90, misses = 44
      precision =   0.1200, recall =   0.0625, f-measure =   0.0822
  - LynneO'Donnell
      hits = 3, strikes = 46, misses = 47
      precision =   0.0600, recall =   0.0612, f-measure =   0.0606
  - BradDorfman
      hits = 0, strikes = 0, misses = 50
      precision =   0.0000, recall =   0.0000, f-measure =   0.0000
  - LynnleyBrowning
      hits = 5, strikes = 72, misses = 45
      precision =   0.1000, recall =   0.0649, f-measure =   0.0787
  - BernardHickey
      hits = 4, strikes = 39, misses = 46
      precision =   0.0800, recall =   0.0930, f-measure =   0.0860
  - MarcelMichelson
      hits = 0, strikes = 0, misses = 50
      precision =   0.0000, recall =   0.0000, f-measure =   0.0000
  - AlexanderSmith
```

hits = 11, strikes = 162, misses = 39
precision = 0.2200, recall = 0.0636, f-measure = 0.0987
- WilliamKazer
hits = 8, strikes = 104, misses = 42
precision = 0.1600, recall = 0.0714, f-measure = 0.0988
- JimGilchrist
hits = 0, strikes = 0, misses = 50
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- RogerFillion
hits = 4, strikes = 45, misses = 46
precision = 0.0800, recall = 0.0816, f-measure = 0.0808
- MarkBendeich
hits = 0, strikes = 0, misses = 50
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- JaneMacartney
hits = 7, strikes = 98, misses = 43
precision = 0.1400, recall = 0.0667, f-measure = 0.0903
- GrahamEarnshaw
hits = 3, strikes = 34, misses = 47
precision = 0.0600, recall = 0.0811, f-measure = 0.0690
- JanLopatka
hits = 0, strikes = 0, misses = 50
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- KeithWeir
hits = 6, strikes = 80, misses = 44
precision = 0.1200, recall = 0.0698, f-measure = 0.0882
- PeterHumphrey
hits = 0, strikes = 0, misses = 50
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- KevinMorrison
hits = 3, strikes = 23, misses = 47
precision = 0.0600, recall = 0.1154, f-measure = 0.0789
- DavidLawder
hits = 0, strikes = 0, misses = 50
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- AaronPressman
hits = 7, strikes = 78, misses = 43
precision = 0.1400, recall = 0.0824, f-measure = 0.1037
- DarrenSchuettler
hits = 3, strikes = 47, misses = 47
precision = 0.0600, recall = 0.0600, f-measure = 0.0600
- ToddNissen
hits = 4, strikes = 66, misses = 46
precision = 0.0800, recall = 0.0571, f-measure = 0.0667
- KirstinRidley
hits = 4, strikes = 60, misses = 46
precision = 0.0800, recall = 0.0625, f-measure = 0.0702

- TimFarrand
  hits = 11, strikes = 156, misses = 39
  precision = 0.2200, recall = 0.0659, f-measure = 0.1014
- JoeOrtiz
  hits = 4, strikes = 41, misses = 46
  precision = 0.0800, recall = 0.0889, f-measure = 0.0842
- KevinDrawbaugh
  hits = 9, strikes = 133, misses = 41
  precision = 0.1800, recall = 0.0634, f-measure = 0.0938
- TheresePoletti
  hits = 0, strikes = 0, misses = 50
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- FumikoFujisaki
  hits = 4, strikes = 68, misses = 46
  precision = 0.0800, recall = 0.0556, f-measure = 0.0656
- JoWinterbottom
  hits = 2, strikes = 21, misses = 48
  precision = 0.0400, recall = 0.0870, f-measure = 0.0548
- JohnMastrini
  hits = 7, strikes = 84, misses = 43
  precision = 0.1400, recall = 0.0769, f-measure = 0.0993
- EdnaFernandes
  hits = 3, strikes = 50, misses = 47
  precision = 0.0600, recall = 0.0566, f-measure = 0.0583
- TanEeLyn
  hits = 4, strikes = 46, misses = 46
  precision = 0.0800, recall = 0.0800, f-measure = 0.0800
- BenjaminKangLim
  hits = 0, strikes = 0, misses = 50
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- PierreTran
  hits = 0, strikes = 0, misses = 50
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- SarahDavison
  hits = 0, strikes = 0, misses = 50
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- LydiaZajc
  hits = 7, strikes = 110, misses = 43
  precision = 0.1400, recall = 0.0598, f-measure = 0.0838
- PatriciaCommins
  hits = 7, strikes = 68, misses = 43
  precision = 0.1400, recall = 0.0933, f-measure = 0.1120
- SimonCowell
  hits = 0, strikes = 0, misses = 50
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- AlanCrosby
  hits = 3, strikes = 61, misses = 47

```
    precision =   0.0600, recall =   0.0469, f-measure =   0.0526
 - JonathanBirt
    hits = 3, strikes = 53, misses = 47
    precision =   0.0600, recall =   0.0536, f-measure =   0.0566
 - KarlPenhaul
    hits = 0, strikes = 0, misses = 50
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
 - ScottHillis
    hits = 3, strikes = 37, misses = 47
    precision =   0.0600, recall =   0.0750, f-measure =   0.0667
 - RobinSidel
    hits = 3, strikes = 50, misses = 47
    precision =   0.0600, recall =   0.0566, f-measure =   0.0583
 - HeatherScoffield
    hits = 0, strikes = 0, misses = 50
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
 - MatthewBunce
    hits = 0, strikes = 0, misses = 50
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
 - MartinWolk
    hits = 0, strikes = 0, misses = 50
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
```

## C50test: Complete Link, min_size=20, no priority queue

```
-> loading dendrogram from  ../cluster_results/C50test_cluster_complete/dendrogram.json
-> reading ground truths from  ../cluster_results/C50test_cluster_complete/truths.out
-- Evaluation Results --
Total correct:    50
Total incorrect:  102
Overall accuracy:   0.3289
Authors:
 - SimonCowell
    hits = 1, strikes = 1, misses = 3
    precision =   0.2500, recall =   0.5000, f-measure =   0.3333
 - KarlPenhaul
    hits = 0, strikes = 0, misses = 3
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
 - DavidLawder
    hits = 2, strikes = 2, misses = 4
    precision =   0.3333, recall =   0.5000, f-measure =   0.4000
 - SarahDavison
    hits = 0, strikes = 0, misses = 0
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
 - LynnleyBrowning
```

hits = 5, strikes = 13, misses = 1
precision = 0.8333, recall = 0.2778, f-measure = 0.4167
- NickLouth
hits = 1, strikes = 0, misses = 1
precision = 0.5000, recall = 1.0000, f-measure = 0.6667
- BernardHickey
hits = 0, strikes = 0, misses = 0
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- FumikoFujisaki
hits = 1, strikes = 1, misses = 1
precision = 0.5000, recall = 0.5000, f-measure = 0.5000
- AaronPressman
hits = 2, strikes = 6, misses = 2
precision = 0.5000, recall = 0.2500, f-measure = 0.3333
- EdnaFernandes
hits = 1, strikes = 2, misses = 4
precision = 0.2000, recall = 0.3333, f-measure = 0.2500
- KevinMorrison
hits = 1, strikes = 0, misses = 0
precision = 1.0000, recall = 1.0000, f-measure = 1.0000
- EricAuchard
hits = 4, strikes = 6, misses = 1
precision = 0.8000, recall = 0.4000, f-measure = 0.5333
- ToddNissen
hits = 0, strikes = 0, misses = 2
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- GrahamEarnshaw
hits = 0, strikes = 0, misses = 0
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- AlexanderSmith
hits = 0, strikes = 0, misses = 1
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- RogerFillion
hits = 0, strikes = 0, misses = 2
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- MatthewBunce
hits = 0, strikes = 0, misses = 3
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- KevinDrawbaugh
hits = 2, strikes = 5, misses = 2
precision = 0.5000, recall = 0.2857, f-measure = 0.3636
- ScottHillis
hits = 0, strikes = 0, misses = 2
precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- TheresePoletti
hits = 2, strikes = 3, misses = 2
precision = 0.5000, recall = 0.4000, f-measure = 0.4444

- BenjaminKangLim
  hits = 1, strikes = 0, misses = 1
  precision = 0.5000, recall = 1.0000, f-measure = 0.6667
- TimFarrand
  hits = 0, strikes = 0, misses = 0
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- BradDorfman
  hits = 0, strikes = 0, misses = 1
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- SamuelPerry
  hits = 2, strikes = 2, misses = 1
  precision = 0.6667, recall = 0.5000, f-measure = 0.5714
- LydiaZajc
  hits = 0, strikes = 0, misses = 2
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- PierreTran
  hits = 0, strikes = 0, misses = 1
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- MarkBendeich
  hits = 0, strikes = 0, misses = 4
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- JonathanBirt
  hits = 0, strikes = 0, misses = 3
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- KirstinRidley
  hits = 0, strikes = 0, misses = 2
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- JaneMacartney
  hits = 0, strikes = 0, misses = 2
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- JohnMastrini
  hits = 1, strikes = 1, misses = 1
  precision = 0.5000, recall = 0.5000, f-measure = 0.5000
- DarrenSchuettler
  hits = 1, strikes = 1, misses = 1
  precision = 0.5000, recall = 0.5000, f-measure = 0.5000
- JimGilchrist
  hits = 0, strikes = 0, misses = 1
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- LynneO'Donnell
  hits = 2, strikes = 7, misses = 2
  precision = 0.5000, recall = 0.2222, f-measure = 0.3077
- MarcelMichelson
  hits = 0, strikes = 0, misses = 5
  precision = 0.0000, recall = 0.0000, f-measure = 0.0000
- JoWinterbottom
  hits = 0, strikes = 0, misses = 3

```
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- RobinSidel
    hits = 5, strikes = 17, misses = 0
    precision =   1.0000, recall =   0.2273, f-measure =   0.3704
- KouroshKarimkhany
    hits = 3, strikes = 9, misses = 4
    precision =   0.4286, recall =   0.2500, f-measure =   0.3158
- JanLopatka
    hits = 1, strikes = 1, misses = 2
    precision =   0.3333, recall =   0.5000, f-measure =   0.4000
- AlanCrosby
    hits = 0, strikes = 0, misses = 1
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- MartinWolk
    hits = 3, strikes = 7, misses = 1
    precision =   0.7500, recall =   0.3000, f-measure =   0.4286
- JoeOrtiz
    hits = 4, strikes = 6, misses = 2
    precision =   0.6667, recall =   0.4000, f-measure =   0.5000
- TanEeLyn
    hits = 0, strikes = 0, misses = 2
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- PatriciaCommins
    hits = 3, strikes = 5, misses = 4
    precision =   0.4286, recall =   0.3750, f-measure =   0.4000
- PeterHumphrey
    hits = 0, strikes = 0, misses = 9
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- MichaelConnor
    hits = 2, strikes = 7, misses = 1
    precision =   0.6667, recall =   0.2222, f-measure =   0.3333
- HeatherScoffield
    hits = 0, strikes = 0, misses = 5
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- KeithWeir
    hits = 0, strikes = 0, misses = 1
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- WilliamKazer
    hits = 0, strikes = 0, misses = 4
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
- MureDickie
    hits = 0, strikes = 0, misses = 2
    precision =   0.0000, recall =   0.0000, f-measure =   0.0000
```

# Conclusion

KNN worked relatively well, with accuracy always exceeding 50%, and often exceeding 60%, when stop-words were used. Hierarchical clustering was not anywhere near as efficient. When single-link was used as a distance metric, almost all merges were the large cluster merging with one data point. average-link improved results slightly, and complete-link improved results slightly again. Even with these improved results, things were still very poor for hierarchical clustering. We attempted to improve this by implementing a priority queue for splitting up the largest clusters first. This decreased the accuracy a lot, but left out a lot more points. Either way, KNN far out-performs hierarchical clustering.