Table with 2199 data, less ambiguous:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Micro Accuracy | Micro Recall | Macro Accuracy | Macro Recall | Pairwise Accuracy | Pairwise Recall |
| **n\_cluster = 150** |  |  |  |  |  |  |
| Cosine+average | 0.9936 | 0.9972 | 0.98 | 0.98 | 0.9973 | 0.9936 |
| Euclidean+average | 0.9545 | 0.9981 | 0.9867 | 0.98 | 0.9981 | 0.9408 |
| Euclidean+ward | 0.9727 | 0.9263 | 0.9733 | 0.9067 | 0.9067 | 0.9727 |
| Cosine+complete | 0.9713 | 0.9627 | 0.9733 | 0.9133 | 0.9713 | 0.962 |
| Euclidean+complete | 0.9850 | 0.9804 | 0.9667 | 0.9467 | 0.9704 | 0.9850 |
| **n\_cluster = 140** |  |  |  |  |  |  |
| Cosine+average | 0.9718 | 0.9995 | 0.9533 | 0.9933 | 0.9995 | 0.9572 |
| Euclidean+average | 0.8758 | 0.9981 | 0.98 | 0.98 | 0.9981 | 0.8622 |
| Euclidean+ward | 0.9581 | 0.9326 | 0.96 | 0.92 | 0.9135 | 0.9577 |
| Cosine+complete | 0.9413 | 0.9627 | 0.9733 | 0.9133 | 0.9527 | 0.9331 |
| Euclidean+complete | 0.9618 | 0.9813 | 0.9467 | 0.9533 | 0.9713 | 0.9608 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Type | Micro Accuracy | Micro Recall | Macro Accuracy | Macro Recall | Pairwise Accuracy | | Pairwise Recall |
| **n\_cluster = 150** |  |  |  |  | |  |  |
| String Similarity | 0.9472 | 0.9450 | 0.88 | 0.8733 | | 0.9272 | 0.9068 |
| String Identity | 0.8945 | 0.9108 | 0.9933 | 0.7866 | | 0.8785 | 0.8872 |
| Attribute Overlap | 0.1569 | 0.3192 | 0.3133 | 0.0333 | | 0.2456 | 0.1091 |
| Entity Overlap |  |  |  |  | |  |  |
| Work Overlap |  |  |  |  | |  |  |
| Tf-Idf |  |  |  |  | |  |  |
| Combined Features |  |  |  |  | |  |  |
| **n\_cluster = 140** |  |  |  |  | |  |  |
| String Similarity | 0.9236 | 0.9449 | 0.85 | 0.8642 | | 0.9272 | 0.8840 |
| String Identity | 0.8726 | 0.9117 | 0.9928 | 0.7785 | | 0.8794 | 0.8653 |
| Attribute Overlap | 0.1505 | 0.3383 | 0.3357 | -0.035 | | 0.2701 | 0.1027 |
| Entity Overlap |  |  |  |  | |  |  |
| Work Overlap |  |  |  |  | |  |  |
| Tf-Idf |  |  |  |  | |  |  |
| Combined Features |  |  |  |  | |  |  |

Table with larger amount of data, more ambiguous, multiple word, with noise:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Micro Accuracy | Micro Recall | Macro Accuracy | Macro Recall | Pairwise Accuracy | Pairwise Recall |
| **n\_cluster = 150** |  |  |  |  |  |  |
| Cosine+average | 0.9817 | 0.9855 | 0.96 | 0.9533 | 0.9692 | 0.9699 |
| Euclidean+average | 0.9459 | 0.9855 | 0.9866 | 0.9466 | 0.9692 | 0.9395 |
| Euclidean+ward | 0.9846 | 0.9463 | 0.98 | 0.8866 | 0.9008 | 0.984 |
| Cosine+complete | 0.9434 | 0.9540 | 0.9866 | 0.86 | 0.9090 | 0.9277 |
| Euclidean+complete | 0.9732 | 0.9595 | 0.9533 | 0.88 | 0.9151 | 0.9572 |
| **n\_cluster = 140** |  |  |  |  |  |  |
| Cosine+average | 0.9611 | 0.9858 | 0.9533 | 0.96 | 0.9694 | 0.9421 |
| Euclidean+average | 0.8975 | 0.9855 | 0.9866 | 0.9466 | 0.9692 | 0.8911 |
| Euclidean+ward | 0.9738 | 0.9512 | 0.98 | 0.9 | 0.9074 | 0.9735 |
| Cosine+complete | 0.9199 | 0.9540 | 0.9866 | 0.86 | 0.9090 | 0.9041 |
| Euclidean+complete | 0.9476 | 0.9595 | 0.9533 | 0.88 | 0.9151 | 0.9313 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Micro Accuracy | Micro Recall | Macro Accuracy | Macro Recall | Pairwise Accuracy | Pairwise Recall |
| **n\_cluster = 150** |  |  |  |  |  |  |
| String Similarity | 0.9413 | 0.9508 | 0.8666 | 0.7266 | 0.9132 | 0.8881 |
| String Identity | 0.8785 | 0.9270 | 0.9933 | 0.68 | 0.8660 | 0.8719 |
| Attribute Overlap | 0.2194 | 0.5063 | 0.4266 | 0.0066 | 0.3872 | 0.0565 |
| Entity Overlap |  |  |  |  |  |  |
| Work Overlap |  |  |  |  |  |  |
| Tf-Idf |  |  |  |  |  |  |
| Combined Features |  |  |  |  |  |  |
| **n\_cluster = 140** |  |  |  |  |  |  |
| String Similarity | 0.9190 | 0.9512 | 0.8857 | 0.7071 | 0.9137 | 0.8609 |
| String Identity | 0.8625 | 0.9281 | 0.9928 | 0.6785 | 0.8713 | 0.8558 |
| Attribute Overlap | 0.2187 | 0.5076 | 0.45 | -0.0642 | 0.3960 | 0.0544 |
| Entity Overlap |  |  |  |  |  |  |
| Work Overlap |  |  |  |  |  |  |
| Tf-Idf |  |  |  |  |  |  |
| Combined Features |  |  |  |  |  |  |

Plot of HAC with different number of cluster: — for relatively noisy dataset

1. Micro Analysis (precision & recall):



1. Macro Analysis (precision & recall):



1. Pairwise Analysis (precision & recall):



Plot using paper-defined similarity function with different number of cluster: — for relatively noisy dataset

1. Micro Analysis (precision & recall):
2. Macro Analysis (precision & recall):
3. Pairwise Analysis (precision & recall):