7.1.1 Benefits of the Consensus Matrix

The consensus matrix from Figure 7.2 is very interesting because the ensemble that was created using the ensemble of 4, 3, and 2 clusterings in the ensemble:

\[ A_{ij} = \text{number of times word } i \text{ appears in document } j \]

The consensus matrix as the adjacency matrix for the graph (Chapter 6 Section 6.2) to count the number of clusters in this dataset.

These two definitions are of course equivalent.

As an example, the consensus matrix and traditional similarity matrices. One problem with these traditional methods is that they do not account for the number of clusters in the data.

Noise removal: 4 algorithms

1. Used consensus matrix and drop tolerance
2. Used DBSCAN on distance matrix
3. Used DBSCAN on consensus matrix
4. If tweet was marked as noise in at least 2 out of 3 previous algorithms, it was considered a noise point

Conclusion

- Used cluster analysis to find topics in tweets
- Ran k-means on consensus matrix and procured similar results with NMF
- NMF is faster and provides more easily interpreted results
- NMF can pick out representative tweet
- Future work: further explore Gephi visualization and use other algorithms

Acknowledgements

We would like to thank Dr. Carl Meyer and Dr. Shaina Race for guiding us through this research and the NSF (DMS-1063010) and NSA (H98230-12-1-0299) for funding the project.