# Analysis of Clustering Results

Overview

We clustered the Wikipedia people articles using our EM algorithm for Gaussian Mixture Models (GMMs) on TF-IDF vectors, with k = 5 clusters. Below we interpret the cluster assignments, cluster statistics, word clouds, and convergence behavior, using the files generated during the run.

1. Cluster Assignments & Distribution

- Cluster Sizes (from weights in em\_parameters.txt):

- Cluster 0: ~21.0% of articles

- Cluster 1: ~29.4%

- Cluster 2: ~12.2%

- Cluster 3: ~18.2%

- Cluster 4: ~19.2%

This shows that all clusters are well-populated—there are no empty or trivial clusters. The distribution is not perfectly balanced but ensures every cluster contains a significant subset of the data.

2. Cluster Interpretability

2.1 Top Words per Cluster

From cluster\_stats.txt, here are the most characteristic words for each cluster:

- Cluster 0: chicksin, teamthe, ransacked, herpetologists, actorwith

- Cluster 1: replacementhe, litigating, frags, ineke, akpalanivel

- Cluster 2: replacementhe, nahan, keates, jaxx, nonissues

- Cluster 3: schuurman, jpmorgan, maraveyas, occasionsin, scaletta

- Cluster 4: scaletta, consternation, splmanasir, jaxx, kia

Some of these words may look unusual (likely due to tokenization and dataset artifacts), but within the TF-IDF space, they are highly distinctive for their respective clusters. For example:

- Cluster 3 seems to have a financial or musical theme ("jpmorgan", "maraveyas", "mandolinist").

- Cluster 4 includes terms like "scaletta" and "consternation", hinting at specific topics or recurring entities in that subset.

2.2 ASCII Word Clouds

Visual “word clouds” for each cluster, from ascii\_wordclouds.txt, reinforce the above, with the top words represented by more stars (“\*”), indicating higher importance:

- Cluster 0:

chicksin : \*\*\*\*\*\*\*\*\*\*

teamthe : \*\*\*\*\*\*\*\*\*

ransacked : \*\*\*\*\*\*\*\*

...

- Cluster 3:

schuurman : \*\*\*\*\*\*\*\*\*\*

jpmorgan : \*\*\*\*\*\*

...

These ASCII word clouds make the dominant cluster words easy to spot at a glance, which is useful for quick, qualitative analysis.

3. Model Convergence

The log-likelihood increased consistently during training, as seen in convergence\_log.txt. Sample values:

992,749,638.52

1,020,027,470.82

1,030,369,221.39

...

1,056,595,849.64

This monotonic increase is a clear sign of successful EM convergence, with the model steadily improving its fit to the data.

4. Strengths and Limitations

Strengths:

- Stable, non-trivial clusters: Every cluster captures a significant share of the data.

- Clear convergence: The EM algorithm did not diverge or get stuck in degenerate solutions.

- Interpretability: Despite some unusual vocabulary, each cluster has top words that can serve as a starting point for understanding the group.

Limitations:

- Tokenization artifacts: Some cluster-defining words (e.g., “replacementhe”, “teamthe”) appear due to data quirks or TF-IDF feature construction.

- Rare/obscure words: In some clusters, top words are not always semantically interpretable without further preprocessing.

5. Recommendations

- Further text cleaning (e.g., merging or correcting tokens, removing rare tokens) could improve interpretability.

- Manual inspection of a few articles from each cluster would help confirm the semantic themes.

- For deeper topic discovery, consider integrating topic models (LDA/NMF) or running with more clusters.

Conclusion

The EM-based clustering produced stable, non-trivial clusters with clear defining words, verified by both quantitative metrics (log-likelihood) and qualitative outputs (word clouds). While additional preprocessing could enhance interpretability, the results robustly demonstrate unsupervised grouping of textual data.

End of Analysis