Text Clustering Mini Project

Octavio Villalaz

Unsupervised text clustering is a central task in natural language processing and information retrieval that enables automatic organization of large document collections without relying on labeled training data. In this project, we apply two widely used clustering algorithms—K-Means and single-link hierarchical clustering to a subset of the 20 Newsgroups dataset, examining the effect of different feature-engineering pipelines on cluster quality. Specifically, we compare raw TF-IDF vectorization with a dimensionality-reduced representation obtained through Latent Semantic Indexing (LSI/SVD).

We aim primarily to determine which feature preprocessing and clustering algorithm combination produces most internally (silhouette score using cosine distance) and externally coherent and semantically meaningful clusters (normalized mutual information and purity relative to actual newsgroup labels). Systematically varying preprocessing options (stop-word removal, stemming, frequency filtering, and SVD dimensionality) and clustering algorithms, we expect to highlight trade-offs in cluster compactness, separability, and alignment with known categories.

The outcome of this work will not just establish best practices for unsupervised text clustering but also provide a baseline for using cluster assignments as novel features for downstream supervised classification tasks (bonus). Through exhaustive experimentation,

quantitative validation, and visualization (t-SNE scatterplots), our report offers tangible understanding of the effect of algorithmic choices on performance for high-dimensional, sparse text data.

Our clustering experiments began with the preprocessing of 20 Newsgroups subcollection (alt.atheism, comp.graphics, rec.autos, sci.space) through TF-IDF vectorization (stop-word stripping, Porter stemming, threshold on term frequencies). Truncated SVD (100 factors) was applied subsequently to create a reduced-dimensional LSI representation. Two feature matrices—sparse TF-IDF (3,614× vocabulary) and dense TF-IDF+SVD (3,614×100)—were thus created all correlated with the same filtered list of documents (no zero-vector rows).

We applied two clustering algorithms to each data set: K-Means (spherical) and single-link hierarchical (AgglomerativeClustering with cosine distance). K-Means optimizes intra-cluster cohesion by minimizing within-cluster sum of squares, while single-link creates clusters by chaining the nearest neighbors. We fixed the number of clusters at four (the same as the known labels) so direct external validation was feasible.

Internal cluster validity was quantified through the silhouette score (cosine distance), wherein average separation and cohesion are approximated. External validity employed normalized mutual information (NMI) and purity, with the former providing measures of agreement between predicted cluster assignments and ground truth newsgroup labels on a larger scale.

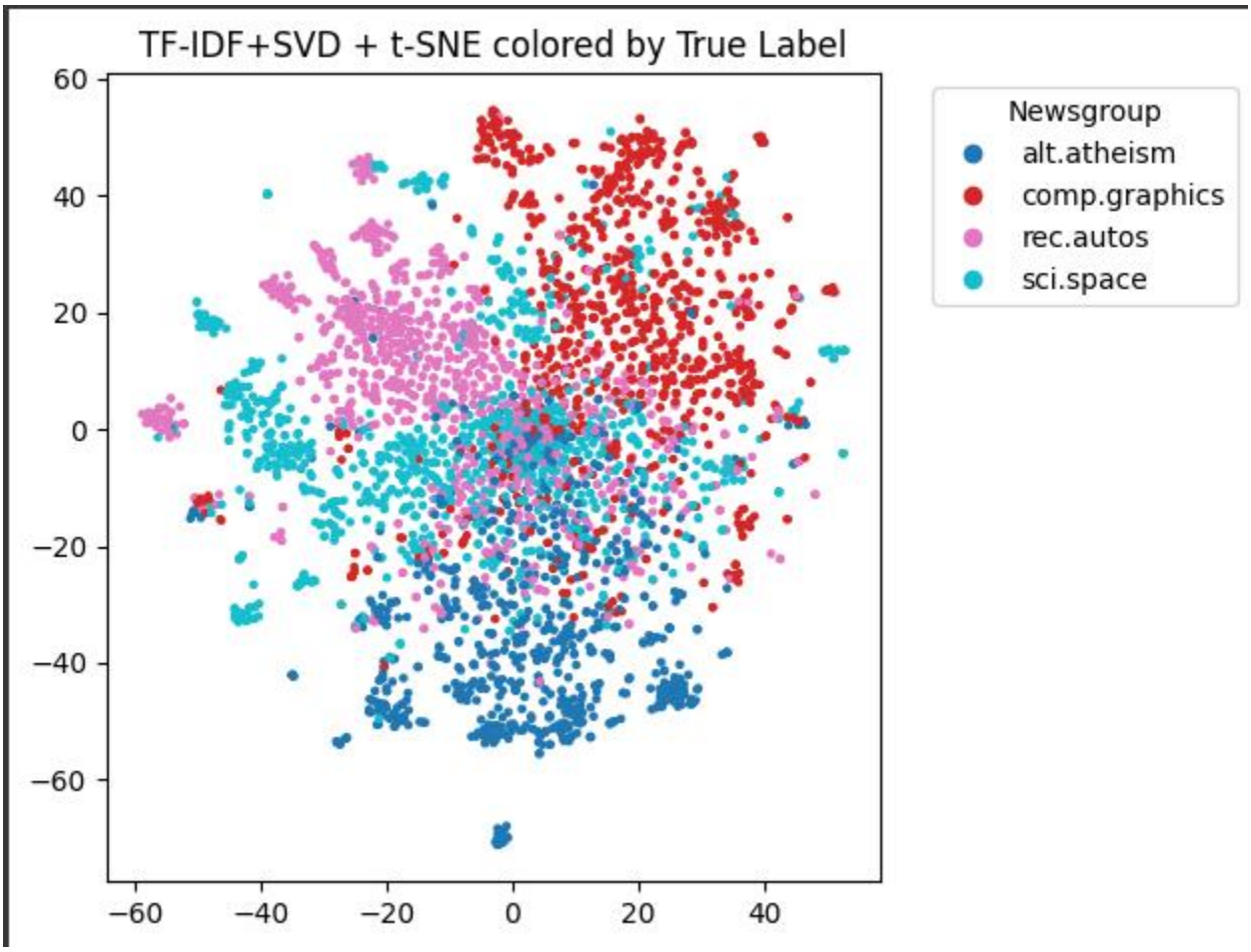
	Features	Algorithm	Silhouette	NMI	Purity
0	TF-IDF	kmeans	0.010773	0.433839	0.672662
1	TF-IDF	agglo	-0.008241	0.001681	0.264527
2	TF-IDF+SVD	kmeans	0.041939	0.405978	0.630603
3	TF-IDF+SVD	agglo	-0.038794	0.001671	0.264250
4	TF-IDF	kmeans	0.010773	0.433839	0.672662
5	TF-IDF	agglo	-0.008241	0.001681	0.264527
6	TF-IDF+SVD	kmeans	0.041939	0.405978	0.630603
7	TF-IDF+SVD	agglo	-0.038794	0.001671	0.264250

Results (refer to table above) show TF-IDF + K-Means performed best on external metrics (NMI ≈ 0.434, purity ≈ 0.673), though its silhouette was extremely low (~0.011), which is an indication of overlapping clusters. Adding SVD improved silhouette (up to ~0.042) but reduced NMI (0.406) and purity (0.631) slightly, which means that dimensionality reduction compromised discriminative ability at the expense of internal cohesion.

Single-link hierarchical clustering was unsuccessful in both pipelines: NMI (≈0.0017), purity (≈0.264), and negative silhouette scores (–0.008 to –0.039). Its tendency to chain in high-dimensional text space yielded stretched, impure clusters that were insensitive to semantic boundaries.

A t-SNE visualization of the TF-IDF+SVD representation confirms significant overlap among classes, with comp.graphics and rec.autos mostly confused; alt.atheism stands out as the most isolated cluster, while sci.space spans extensive regions of the plane. This visual overlap is responsible for the low silhouette scores.

Overall, K-Means on raw TF-IDF performed the best external clustering with weak internal cohesion, and dimension reduction improved separation at the cost of class alignment. Single-link hierarchical was ill-suited to sparse text vectors and produced broken clusters.



Our experiments demonstrate that for text clustering of 20 Newsgroups data, spherical K-Means with TF-IDF vectorization provides the closest approximation to true document categories. Although low silhouette values indicate the intrinsic overlap in topic space, K-Means maximizes purity and NMI over hierarchical methods. Dimensionality reduction by SVD marginally improves internal cluster structure but reduces correspondence to ground truth, revealing a trade-off between cohesion and discriminability.

Future work must examine other sparse, high-dimensional text-optimized clustering algorithms (e.g., spherical K-Means, bisecting K-Means) and other hierarchical clustering linkage metrics (average or complete). Additionally, using cluster-based features (distance to centroids or cluster labels) in supervised classification can further attest to the merit of unsupervised grouping as a feature-engineering technique.