Clustering Results

Dejvis Toptani

2023-05-17

Clustering of 20news dataset

DBSCAN

min__samples	eps	umap__n__components	NMI	embedding
6	0.195	8	0.327	BERT
6	0.195	8	0.327	BERT
6	0.196	8	0.326	BERT
7	0.217	2	0.542	fasttext
7	0.215	2	0.542	fasttext
7	0.206	2	0.542	fasttext
5	0.469	8	0.558	tfidf
5	0.470	8	0.558	tfidf
5	0.470	8	0.558	tfidf
10	0.460	16	0.467	word2vec
10	0.460	16	0.467	word2vec
10	0.459	16	0.467	word2vec

KMeans

n__clusters	umap__n__components	NMI	embedding
6	4	0.237	BERT
6	4	0.237	BERT
6	4	0.237	BERT
6	2	0.559	fasttext
6	2	0.559	fasttext
6	2	0.559	fasttext
6	8	0.686	tfidf
6	8	0.686	tfidf
6	8	0.686	tfidf
6	2	0.514	word2vec
6	2	0.513	word2vec
6	2	0.512	word2vec

Spectral Clustering

n__clusters	umap__n__components	NMI	embedding
7	2	0.176	BERT
7	2	0.176	BERT

n_clusters	umap_n_components	NMI	embedding
7	2	0.176	BERT
6	16	0.554	fasttext
6	16	0.554	fasttext
6	16	0.554	fasttext
5	8	0.693	tfidf
5	8	0.693	tfidf
5	8	0.693	tfidf
7	8	0.503	word2vec
7	8	0.503	word2vec
7	8	0.503	word2vec

OPTICS

min_samples	xi	umap_n_components	NMI	embedding
46	0.012	16	0.336	BERT
46	0.012	16	0.336	BERT
46	0.011	16	0.336	BERT
172	0.013	16	0.562	fasttext
141	0.013	8	0.554	fasttext
137	0.010	16	0.551	fasttext
193	0.011	8	0.688	tfidf
191	0.010	8	0.686	tfidf
161	0.041	16	0.670	tfidf
160	0.011	16	0.515	word2vec
160	0.010	16	0.515	word2vec
148	0.011	16	0.509	word2vec

DBHD

min_cluster_size	rho	beta	umap_n_components	NMI	embedding
97	1.832	0.143	8	0.342	BERT
97	1.874	0.129	8	0.337	BERT
97	1.770	0.165	8	0.336	BERT
171	1.713	0.279	2	0.552	fasttext
175	1.836	0.356	2	0.548	fasttext
176	1.831	0.396	2	0.547	fasttext
172	1.749	0.425	2	0.609	tfidf
177	1.726	0.429	2	0.606	tfidf
178	1.756	0.419	2	0.605	tfidf
182	1.858	0.177	2	0.538	word2vec
182	1.910	0.205	2	0.538	word2vec
176	1.947	0.113	2	0.537	word2vec

Meanshift

bandwidth	umap_n_components	NMI	embedding
0.010	16	0.341	BERT
0.010	16	0.341	BERT

bandwidth	umap_n_components	NMI	embedding
0.010	16	0.341	BERT
0.880	16	0.543	fasttext
0.891	16	0.543	fasttext
0.889	16	0.543	fasttext
0.909	16	0.605	tfidf
0.909	16	0.605	tfidf
0.908	16	0.604	tfidf
0.908	2	0.529	word2vec
0.904	2	0.528	word2vec
0.906	2	0.528	word2vec

SNN-DPC

k	nc	umap_n_components	NMI	embedding
45	6	8	0.261	BERT
45	6	8	0.261	BERT
45	6	8	0.261	BERT
30	6	16	0.553	fasttext
30	6	16	0.553	fasttext
30	6	16	0.553	fasttext
48	6	4	0.644	tfidf
48	6	4	0.644	tfidf
48	6	4	0.644	tfidf
39	6	16	0.498	word2vec
39	6	16	0.498	word2vec
39	6	16	0.498	word2vec

HDBSCAN

min_cluster_size	min_samples	umap_n_components	NMI	embedding
29	5	2	0.300	BERT
27	5	2	0.299	BERT
27	15	2	0.298	BERT
90	114	4	0.258	fasttext
129	116	4	0.258	fasttext
111	116	4	0.258	fasttext
32	171	8	0.558	tfidf
22	171	8	0.558	tfidf
47	171	8	0.558	tfidf
101	5	2	0.491	word2vec
146	5	2	0.491	word2vec
160	5	2	0.491	word2vec

SpectralACL

n_clusters	epsilon	umap_n_components	NMI	embedding
4	0.073	16	0	BERT
4	0.866	16	0	BERT

n_clusters	epsilon	umap_n_components	NMI	embedding
4	0.998	16	0	BERT
4	0.665	8	0	fasttext
4	0.505	4	0	fasttext
4	0.344	16	0	fasttext
4	0.217	16	0	tfidf
4	0.839	4	0	tfidf
4	0.581	16	0	tfidf
4	0.954	8	0	word2vec
4	0.840	16	0	word2vec
4	0.362	8	0	word2vec

DBADV

perplexity	MinPts	probability	umap_n_components	NMI	embedding
1	1	0.997	4	0.341	BERT
1	1	0.997	4	0.341	BERT
1	1	0.997	4	0.341	BERT
4	7	0.997	2	0.454	fasttext
4	7	0.997	2	0.454	fasttext
4	7	0.997	2	0.454	fasttext
5	4	0.997	4	0.495	tfidf
5	4	0.997	4	0.495	tfidf
5	4	0.997	4	0.495	tfidf
13	30	0.997	2	0.493	word2vec
13	30	0.997	2	0.493	word2vec
12	28	0.997	2	0.472	word2vec

DPC

density_threshold	distance_threshold	umap_n_components	NMI	embedding
0.352	0.234	8	0.349	BERT
0.370	0.190	16	0.349	BERT
0.277	0.197	4	0.348	BERT
0.478	0.605	2	0.559	fasttext
0.583	0.722	2	0.559	fasttext
0.609	0.721	2	0.559	fasttext
0.645	0.898	4	0.620	tfidf
0.609	0.991	4	0.620	tfidf
0.628	0.904	4	0.620	tfidf
0.470	0.996	2	0.530	word2vec
0.452	0.996	2	0.530	word2vec
0.449	0.998	2	0.530	word2vec