Clustering Results

Dejvis Toptani

2023-05-17

Clustering of bbc dataset

Base classifier:

Average F1-score: 0.9697

Average Accuracy: 0.9700

Average Precision: 0.9709

Average Recall: 0.9689

DBSCAN

min_samples	eps	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
5	0.729	4	0.809	BERT	0.954	0.955	0.954	0.955
5	0.825	4	0.809	BERT	0.955	0.957	0.954	0.957
5	0.837	4	0.809	BERT	0.951	0.953	0.952	0.951
5	0.696	8	0.765	fasttext	0.949	0.951	0.951	0.948
6	0.922	8	0.765	fasttext	0.952	0.953	0.955	0.951
6	0.982	8	0.765	fasttext	0.951	0.952	0.953	0.950
5	0.638	8	0.692	tfidf	0.931	0.931	0.932	0.931
5	0.432	4	0.686	tfidf	0.967	0.967	0.967	0.968
5	0.433	4	0.686	tfidf	0.919	0.921	0.926	0.917
6	0.623	16	0.743	word2vec	0.949	0.950	0.950	0.948
7	0.660	16	0.743	word2vec	0.948	0.950	0.949	0.949
8	0.652	16	0.743	word2vec	0.951	0.952	0.952	0.952

KMeans

n_clusters	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
5	8	0.866	BERT	0.953	0.955	0.953	0.955
5	8	0.866	BERT	0.956	0.957	0.956	0.957
5	8	0.866	BERT	0.954	0.956	0.954	0.955
5	16	0.848	fasttext	0.937	0.938	0.941	0.937
5	16	0.848	fasttext	0.936	0.937	0.939	0.937
5	16	0.848	fasttext	0.939	0.940	0.941	0.939
5	4	0.785	tfidf	0.923	0.923	0.924	0.923
5	4	0.785	tfidf	0.905	0.907	0.907	0.910
5	4	0.785	tfidf	0.935	0.936	0.936	0.936
5	2	0.793	word2vec	0.896	0.899	0.898	0.897
5	2	0.793	word2vec	0.909	0.910	0.912	0.910

n_clusters	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
5	2	0.793	word2vec	0.895	0.896	0.897	0.899

Spectral Clustering

n_clusters	umap_n_components	NMI	embedding
6	4	0.841	BERT
6	4	0.841	BERT
6	4	0.841	BERT
6	16	0.834	fasttext
6	16	0.834	fasttext
6	16	0.834	fasttext
6	4	0.680	tfidf
6	4	0.680	tfidf
6	4	0.680	tfidf
6	2	0.798	word2vec
6	2	0.798	word2vec
6	2	0.798	word2vec

OPTICS

min_samples	xi	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
108	0.125	16	0.855	BERT	0.955	0.957	0.956	0.956
108	0.124	16	0.855	BERT	0.952	0.953	0.952	0.953
106	0.121	16	0.855	BERT	0.955	0.956	0.955	0.957
96	0.088	8	0.834	fasttext	0.947	0.948	0.950	0.946
120	0.265	2	0.777	fasttext	0.952	0.954	0.953	0.952
110	0.269	2	0.777	fasttext	0.953	0.954	0.953	0.954
161	0.031	8	0.780	tfidf	0.917	0.919	0.918	0.918
190	0.037	8	0.778	tfidf	0.931	0.931	0.935	0.929
189	0.033	8	0.776	tfidf	0.914	0.916	0.914	0.915
174	0.027	4	0.789	word2vec	0.929	0.929	0.932	0.928
173	0.039	4	0.788	word2vec	0.934	0.936	0.938	0.933
172	0.045	4	0.788	word2vec	0.925	0.926	0.930	0.924

DBHD

min_cluster_size	rho	beta	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
170	0.157	0.317	16	0.846	BERT	0.960	0.961	0.960	0.962
169	0.062	0.316	16	0.846	BERT	0.957	0.959	0.958	0.957
172	0.150	0.301	16	0.846	BERT	0.964	0.965	0.964	0.964
135	0.773	0.251	16	0.765	fasttext	0.911	0.911	0.919	0.914
132	0.775	0.250	16	0.765	fasttext	0.916	0.916	0.924	0.918
132	0.069	0.314	16	0.765	fasttext	0.947	0.948	0.949	0.945
192	0.814	0.200	16	0.703	tfidf	0.900	0.902	0.902	0.900
195	0.833	0.233	4	0.701	tfidf	0.907	0.908	0.911	0.906
195	0.827	0.272	4	0.701	tfidf	0.948	0.949	0.949	0.949
178	0.287	0.307	4	0.735	word2vec	0.938	0.939	0.941	0.936
171	0.128	0.238	4	0.735	word2vec	0.937	0.938	0.940	0.937

min_cluster_size	rho	beta	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
171	0.141	0.297	4	0.735	word2vec	0.930	0.932	0.934	0.929

Meanshift

bandwidth	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
0.848	16	0.734	BERT	0.935	0.937	0.940	0.934
0.860	16	0.734	BERT	0.952	0.954	0.955	0.952
0.860	16	0.734	BERT	0.832	0.837	0.874	0.825
0.902	4	0.739	fasttext	0.942	0.943	0.944	0.942
0.899	4	0.739	fasttext	0.907	0.905	0.916	0.904
0.895	4	0.739	fasttext	0.936	0.936	0.939	0.935
0.908	4	0.653	tfidf	0.932	0.933	0.934	0.932
0.903	4	0.652	tfidf	0.893	0.895	0.896	0.892
0.907	4	0.652	tfidf	0.936	0.937	0.937	0.936
0.910	8	0.716	word2vec	0.873	0.868	0.889	0.868
0.910	8	0.716	word2vec	0.906	0.907	0.916	0.904
0.874	4	0.715	word2vec	0.893	0.888	0.907	0.889

SNN-DPC

k	nc	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
31	6	4	0.858	BERT	0.950	0.952	0.951	0.951
30	6	4	0.858	BERT	0.952	0.953	0.951	0.954
29	6	4	0.858	BERT	0.948	0.949	0.950	0.947
36	6	16	0.836	fasttext	0.943	0.944	0.945	0.943
36	6	16	0.836	fasttext	0.942	0.943	0.944	0.942
36	6	16	0.836	fasttext	0.942	0.943	0.943	0.942
36	6	2	0.764	tfidf	0.893	0.895	0.899	0.892
36	6	2	0.764	tfidf	0.876	0.880	0.880	0.876
36	6	2	0.764	tfidf	0.927	0.928	0.927	0.928
35	6	8	0.771	word2vec	0.927	0.928	0.932	0.926
35	6	8	0.771	word2vec	0.915	0.916	0.918	0.914
35	6	8	0.771	word2vec	0.919	0.920	0.924	0.919

HDBSCAN

min_cluster_size	min_samples	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
143	77	4	0.876	BERT	0.953	0.955	0.953	0.954
137	77	4	0.876	BERT	0.954	0.956	0.953	0.956
161	77	4	0.876	BERT	0.958	0.959	0.958	0.960
113	100	8	0.777	fasttext	0.948	0.949	0.952	0.947
48	52	16	0.777	fasttext	0.948	0.950	0.951	0.947
69	51	16	0.777	fasttext	0.946	0.947	0.949	0.945
145	5	4	0.733	tfidf	0.923	0.925	0.925	0.924
172	5	4	0.733	tfidf	0.970	0.970	0.970	0.970
174	5	4	0.733	tfidf	0.927	0.928	0.928	0.928
172	89	16	0.798	word2vec	0.953	0.955	0.954	0.953
170	89	16	0.798	word2vec	0.970	0.970	0.971	0.970

min_cluster_size	min_samples	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
172	89	16	0.798	word2vec	0.929	0.931	0.932	0.929

SpectralACL

n_clusters	epsilon	umap_n_components	NMI	embedding
6	0.352	16	0	BERT
6	0.590	2	0	BERT
6	0.064	8	0	BERT
4	0.073	2	0	fasttext
4	0.280	8	0	fasttext
4	0.147	16	0	fasttext
6	0.704	4	0	tfidf
6	0.259	4	0	tfidf
6	0.847	4	0	tfidf
6	0.280	8	0	word2vec
6	0.658	4	0	word2vec
6	0.020	8	0	word2vec

DBADV

perplexity	MinPts	probability	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
23	18	0.997	4	0.814	BERT	0.955	0.956	0.955	0.956
23	18	0.997	4	0.814	BERT	0.953	0.954	0.952	0.955
22	18	0.997	4	0.814	BERT	0.953	0.955	0.954	0.955
23	11	0.997	16	0.765	fasttext	0.952	0.954	0.956	0.951
29	23	0.997	16	0.765	fasttext	0.953	0.954	0.955	0.951
29	16	0.997	8	0.765	fasttext	0.950	0.952	0.953	0.949
19	30	0.997	4	0.674	tfidf	0.928	0.928	0.934	0.926
19	30	0.997	4	0.674	tfidf	0.965	0.964	0.966	0.964
19	30	0.997	4	0.674	tfidf	0.935	0.936	0.940	0.933
9	6	0.997	16	0.708	word2vec	0.948	0.949	0.948	0.949
9	6	0.997	16	0.708	word2vec	0.931	0.931	0.934	0.931
9	6	0.997	16	0.708	word2vec	0.912	0.912	0.915	0.911

DPC

density_threshold	distance_threshold	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
0.135	0.972	4	0.736	BERT	0.926	0.926	0.929	0.930
0.126	0.973	4	0.736	BERT	0.919	0.919	0.922	0.923
0.113	1.000	4	0.736	BERT	0.950	0.951	0.949	0.952
0.227	1.000	4	0.727	fasttext	0.941	0.943	0.942	0.942
0.197	0.971	4	0.727	fasttext	0.931	0.931	0.934	0.931
0.201	0.976	4	0.727	fasttext	0.871	0.868	0.895	0.875
0.586	0.929	16	0.643	tfidf	0.886	0.885	0.889	0.889
0.588	0.914	16	0.643	tfidf	0.920	0.922	0.921	0.921
0.487	0.916	16	0.643	tfidf	0.893	0.896	0.898	0.893
0.294	0.999	4	0.732	word2vec	0.882	0.896	0.877	0.895
0.237	1.000	4	0.732	word2vec	0.921	0.922	0.926	0.921

density_threshold	distance_threshold	umap_n_components	NMI	embedding	f1_score	accuracy	precision	recall
0.242	0.998	4	0.732	word2vec	0.881	0.883	0.883	0.888