Name - Vijay Tadikamalla
Roll. no- CS17BTECH11040

Assignment 2 - Clustering problems in NLP

Foundation of Machine Learning

Clustering Tasks in NLP

We compare the performance of **Kernelized k -means** and **Mixture Model** algorithms on the NLP problem, **Topic Clustering** on 20 Newsgroups Dataset**.**

TF-IDF

Measures	Kernelized K-means	Mixture model
Adjusted Random index	0.00063	-3.47E-06
NMI	0.010	0.0043
AMI	0.0041	0.0019
Homogeneity	0.0067	0.0022
Completeness	0.018	0.26
V-measure	0.0096	0.0044
FMI	0.17	0.22

Sub-Linear TF-Scaling

Measures	Kernelized K-means	Mixture model
Adjusted Random index	0.00073	0.039
NMI	0.012	0.13
AMI	0.0053	0.12
Homogeneity	0.0078	0.093
Completeness	0.020	0.19
V-measure	0.011	0.12
FMI	0.17	0.16

Maximum TF-Normalization

Measures	Kernelized K-means	Mixture model
Adjusted Random index	0.018	0.093
NMI	0.016	0.28
AMI	0.017	0.22
Homogeneity	0.015	0.21
Completeness	0.019	0.22
V-measure	0.016	0.20
FMI	0.10	0.19

Maximum TF-Normalization with Sublinear TF-Scaling

Measures	Kernelized K-means	Mixture model
Adjusted Random index	0.017	0.098
NMI	0.015	0.23
AMI	0.016	0.25
Homogeneity	0.014	0.22
Completeness	0.019	0.21
V-measure	0.019	0.26
FMI	0.08	0.18