CATEGORIZING YOUTUBE VIDEOS

NITESH GUPTA

TANMAY GORE

WHY CLUSTER YOUTUBE VIDEOS?

- BENEFITS:
  - MONETARY BENEFITS
  - VIDEO RECOMMENDATION

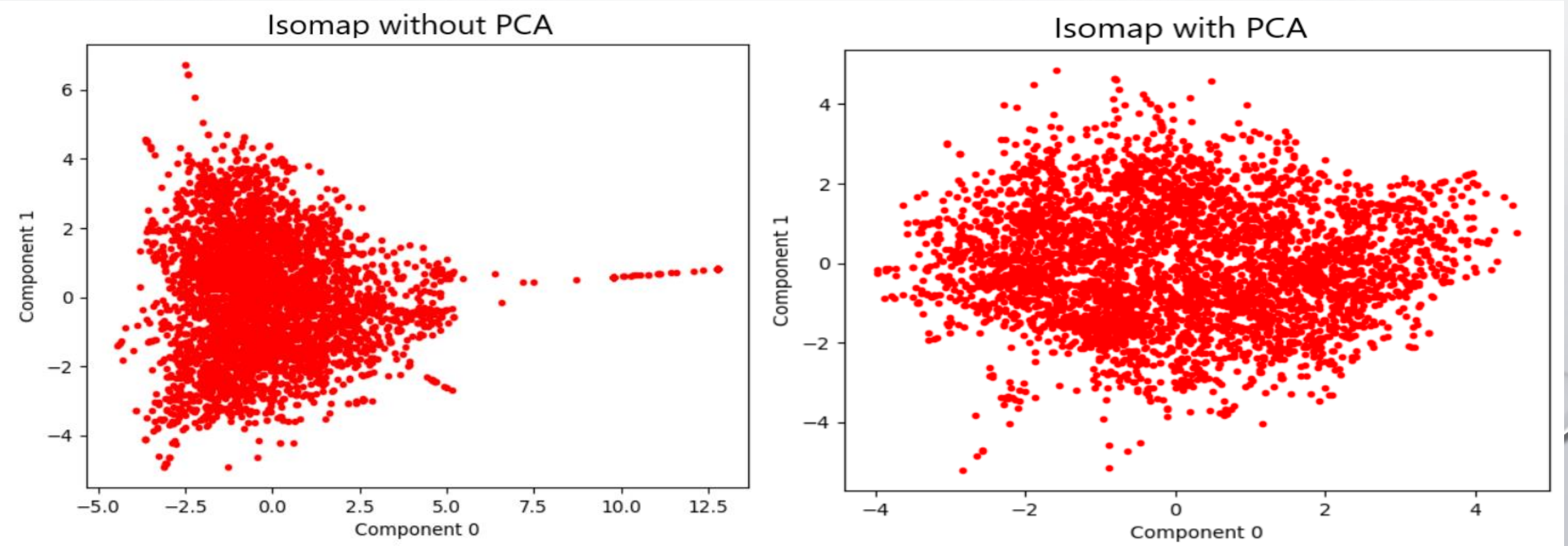
- WHY CLUSTERING:
  - VIDEO CATEGORIES MIGHT NOT BE KNOWN BEFOREHAND.
  - CAN CONSIDER METADATA FOR CLASSIFICATION

TFIDF

- DATA SET CONTAINS FOLLOWING FEATURES:
  - TITLE
  - TAGS
  - DESCRIPTION
- CONVERTS TEXT TO NUMBERS
- PRIORITIZES MEANINGFUL INFORMATION OVER STOP WORDS
- RESULT – 50,000 DIMENSIONS

PCA

- DIMENSIONALITY REDUCTION
- REDUCES 50, 000 DIMENSIONS TO 3000 WITH ONLY 5% LOSS OF INFORMATION.



KMEANS

- CENTROID BASED CLUSTERING
- USES SILHOUETTE COEFFICIENT TO IDENTIFY CORRELATION BETWEEN CLUSTERS
  - +1 : BEST CLUSTER (CLUSTERS ARE NON OVERLAPPING)
  - 0 : CLUSTER ARE OVERLAPPING
  - -1 : WORST CLUSTERS (DATA POINTS ARE ASSIGNED TO WRONG CLUSTERS)

PERFORMANCE MEASURE

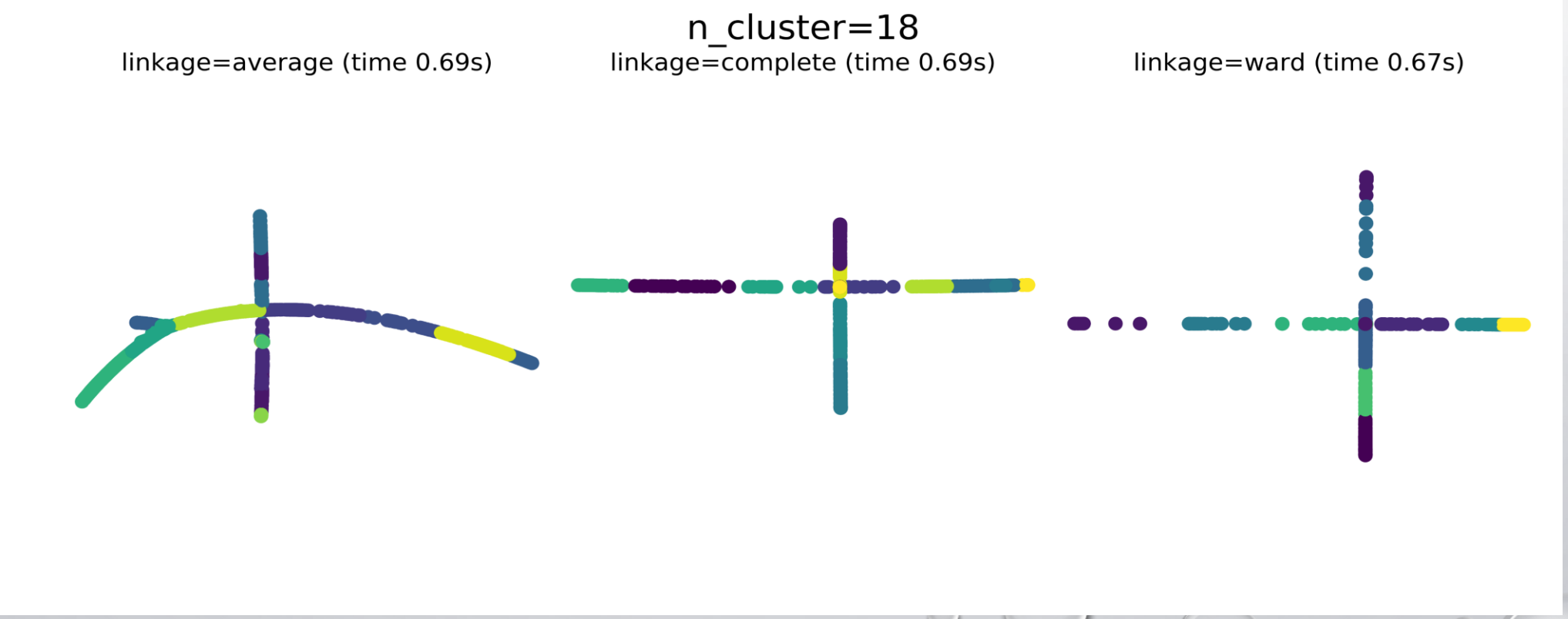
CLUSTERS	SIHOUETTE VALUE	TRAINING TIME (Sec)	CLUSTERS	SILHOUETTE VALUE	TRAINING TIME (Sec)
2	-0.006	26.17	14	0.017	51.75
3	-0.003	28.08	15	0.015	56.36
4	-0.002	28.56	16	0.016	55.47
5	0.002	31.67	17	0.019	65.68
6	0.002	35.15	18	0.020	65.99
7	0.004	36.53	19	0.021	66.00
8	0.006	37.03	20	0.022	66.40
9	0.009	42.22	21	0.024	71.24
10	0.009	45.15	22	0.024	72.98
11	0.01	42.73	23	0.029	73.82
12	0.007	46.68	24	0.027	72.95
13	0.016	50.25	25	0.031	74.10

AGGLOMERATIVE HIERARCHICAL CLUSTERING (AHC)

- WHY HIERARCHICAL CLUSTERING
  - NUMBER OF CLUSTERS NOT KNOWN BEFOREHAND IN SEVERAL DATA ANALYSIS PROBLEM
- TWO TYPES
  - TOP DOWN : START WITH A ONE LARGE CLUSTER
  - BOTTOM UP : INITALIZE EACH DATA POINTS AS ITS OWN CLUSTER AND RECURSIVELY MERGE THE CLUSTERS
- CLUSTER DISTANCE
  - SINGLE LINKAGE
  - COMPLETE LINKAGE
  - AVERAGE LINKAGE
  - WARD LINKAGE : MINIMIZES THE TOTAL WITHIN-CLUSTER VARIANCE

RESULTS

- AHC


n_cluster=18
linkage=average (time 0.69s)
linkage=complete (time 0.69s)
linkage=ward (time 0.67s)

CONCLUSION

- AGGLOMERATIVE HIERARCHICAL CLUSTERING
  - CLUSTERS: 18, LINKAGE : WARD, SILHOUETTE COEFFICIENT: 0.023, TRAINING TIME: 0.67 S
- KMEANS CLUSTERING
  - CLUSTERS: 18, N_ITER : 20, SILHOUETTE COEFFICIENT: 0.020, TRAINING TIME: 65.19 S

REFERENCES

- LECTURES SLIDES BY DR. CHINMAY HEGDE FOR EE 525X SPRING 2018, ISU
- HTTPS://EN.WIKIPEDIA.ORG/WIKI/ISOMAP
- HTTPS://STATS.STACKEXCHANGE.COM/QUESTIONS/124534/HOW-TO-UNDERSTAND-NONLINEAR-AS-IN-NONLINEAR-DIMENSIONALITY-REDUCTION
- HTTPS://EN.WIKIPEDIA.ORG/WIKI/HIERARCHICAL_CLUSTERING