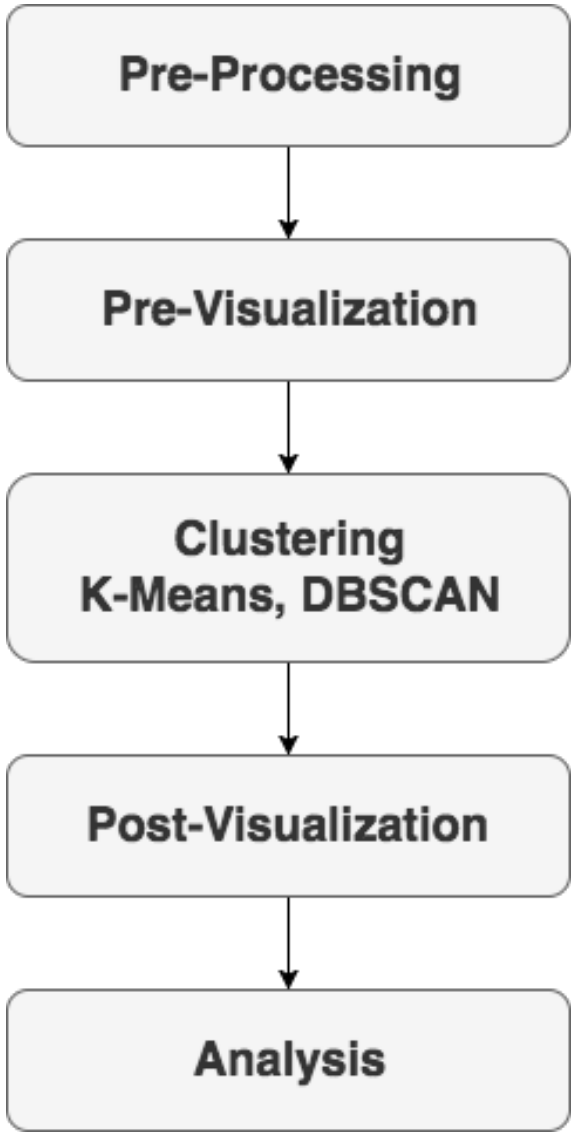
Natural Language Processing Project: Text Clustering

Sudat Tuladhar

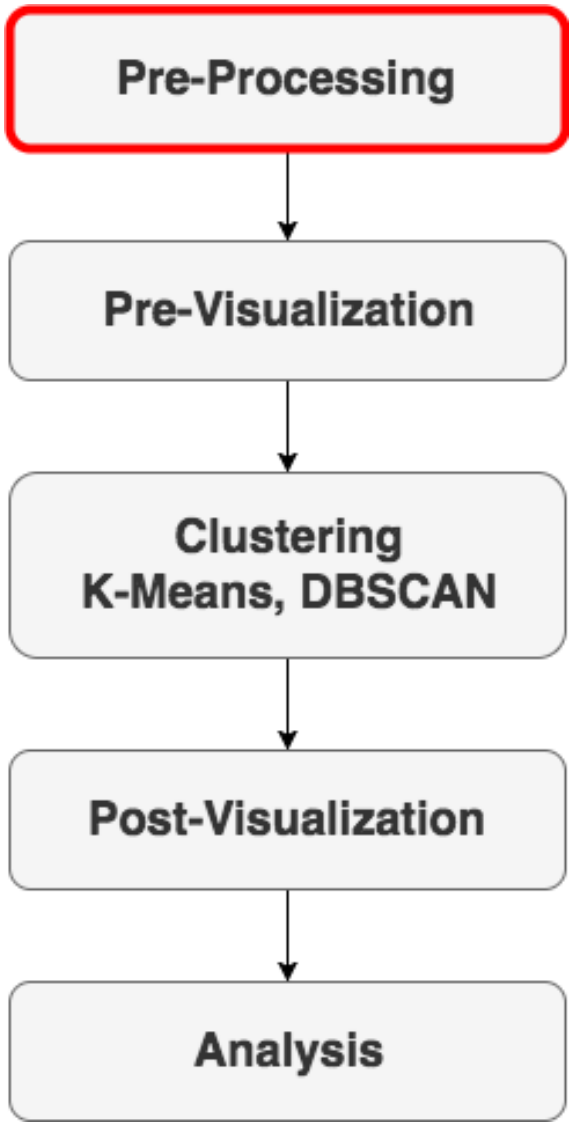
Department of Electrical Engineering
The University of Mississippi

November 28, 2018

Process Description

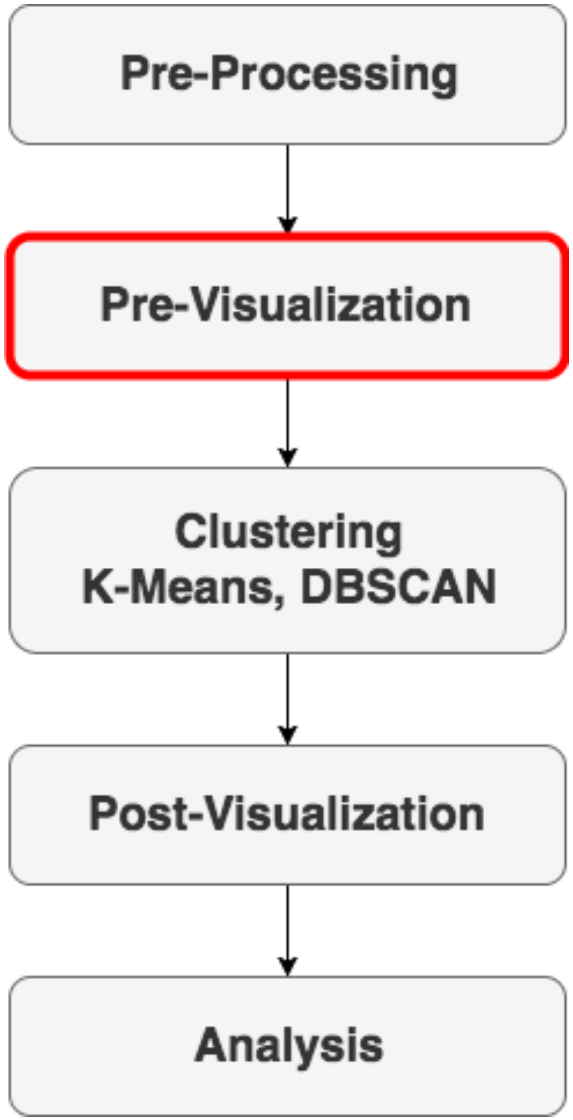


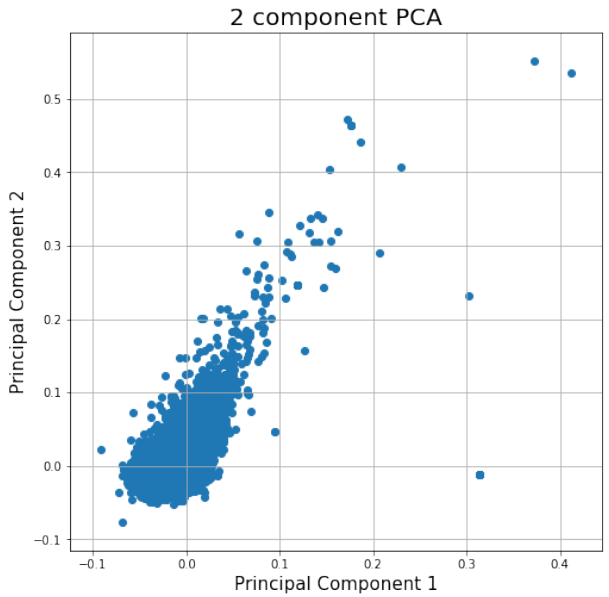
Pre-Processing



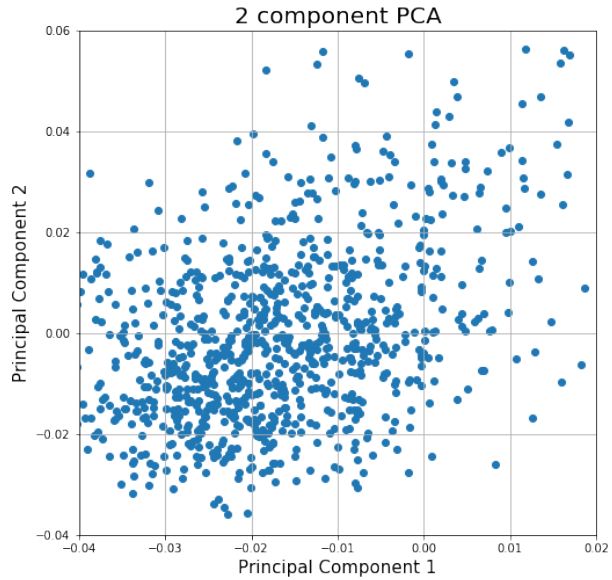
- AwardNo and abstract extraction
- Null Document Filtering
- Word2Vec library
- Unknown words Filtering
- Doc2Vec as mean of all Word2Vec in a document

Pre-Visualization





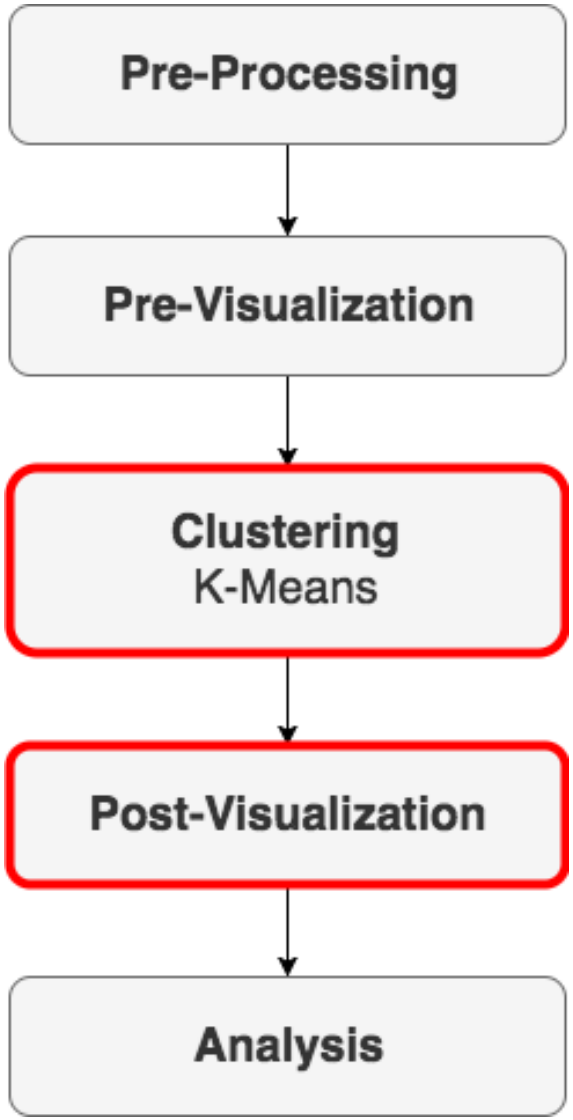
(a) All Data, n = 51715



(b) Samples, n = 1000

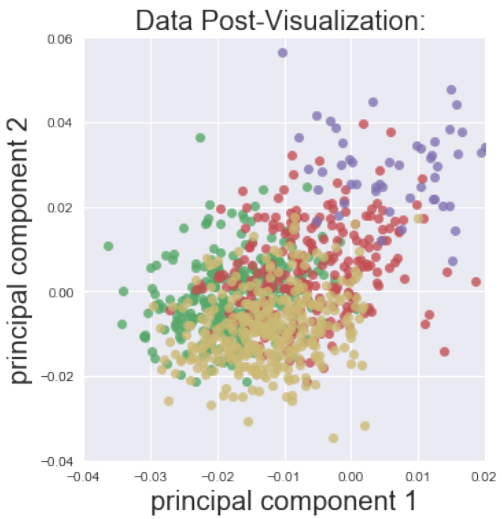
- Word2Vec model: 300 dimension
- PCA: 2-component for 2D visualization
- Variation: [0.644 0.061]

Clustering: K-Means

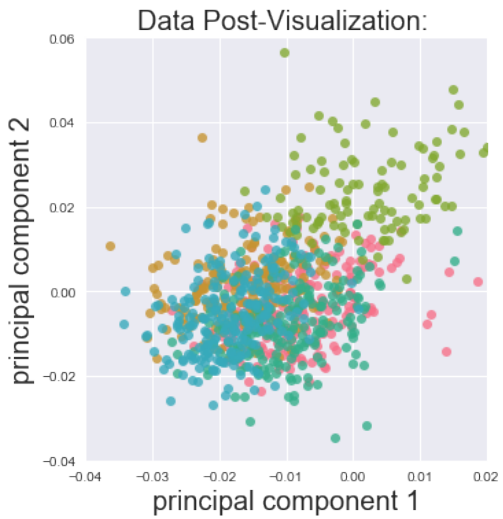


- K-Means for K = 5, 7, 10, 12, 15
- Pandas data-frames as a data structure
- Random data point as initial centers of the cluster
- Updating Labels: Cosine Similarity
- Center Updates
- SSE

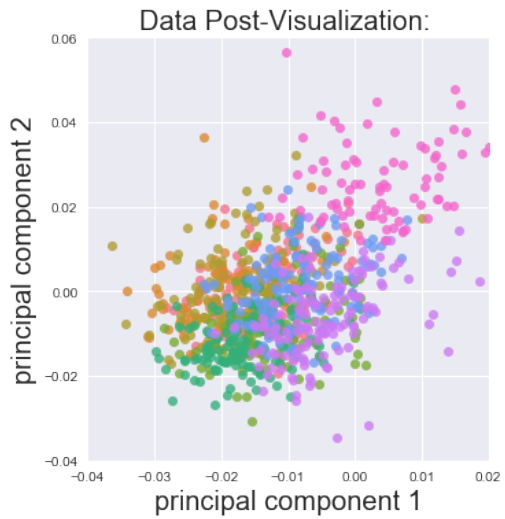
K-Means: Post-Visualization, Data and Count



(a) Data, K = 5



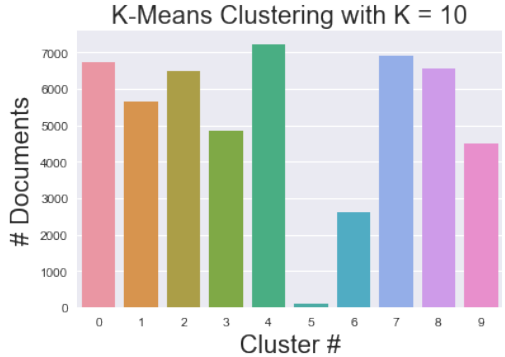
(b) Data, K = 7



(c) Data, K = 10

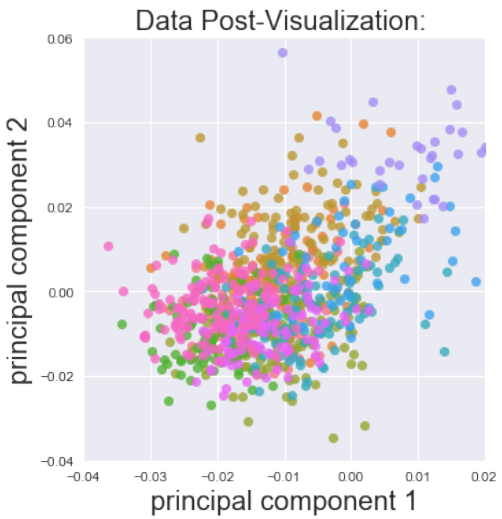
(a) Count, K = 5

(b) Count, K = 7

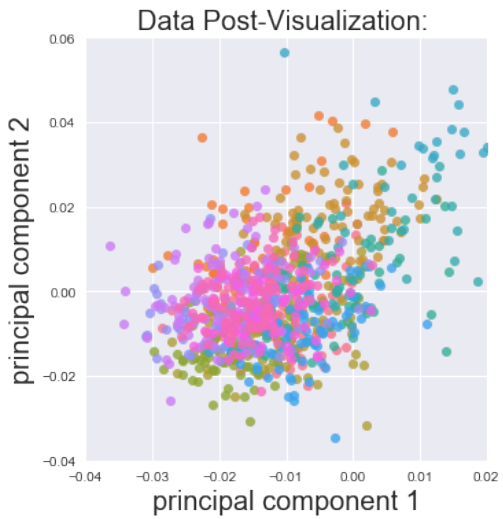


(c) Count, K = 10

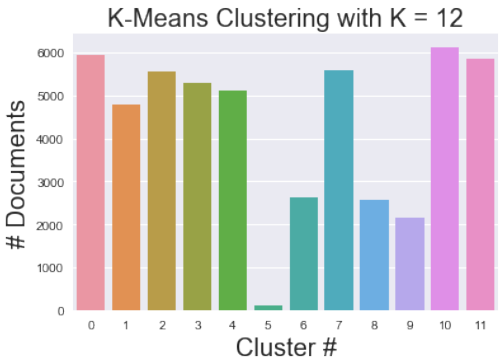
K-Means: Post-Visualization



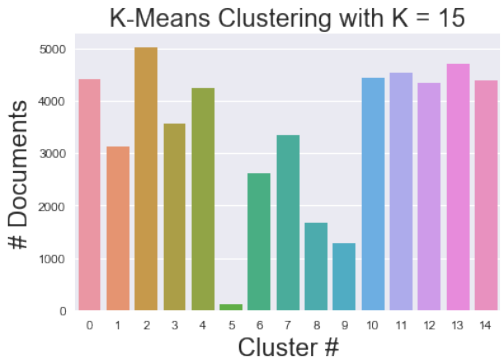
(a) Data, K = 12



(b) Data , K = 15

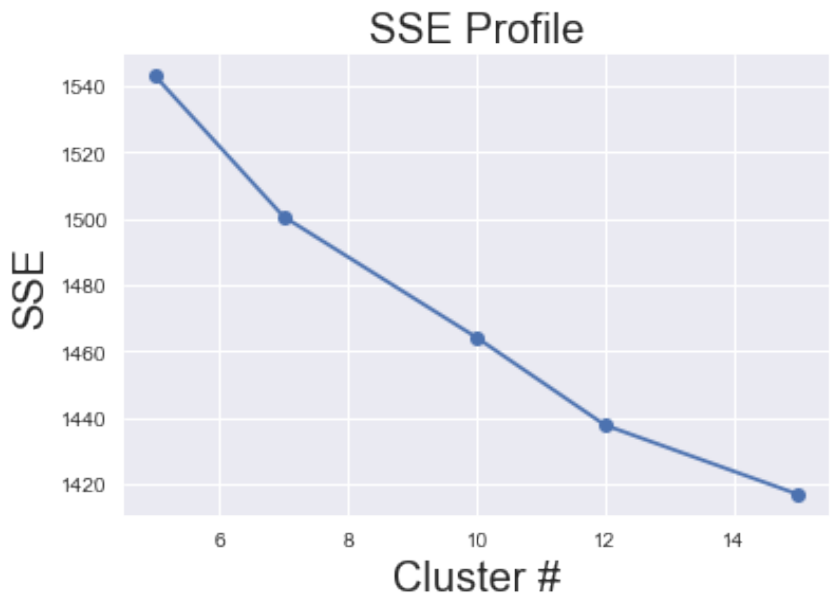


(a) Count, K = 12

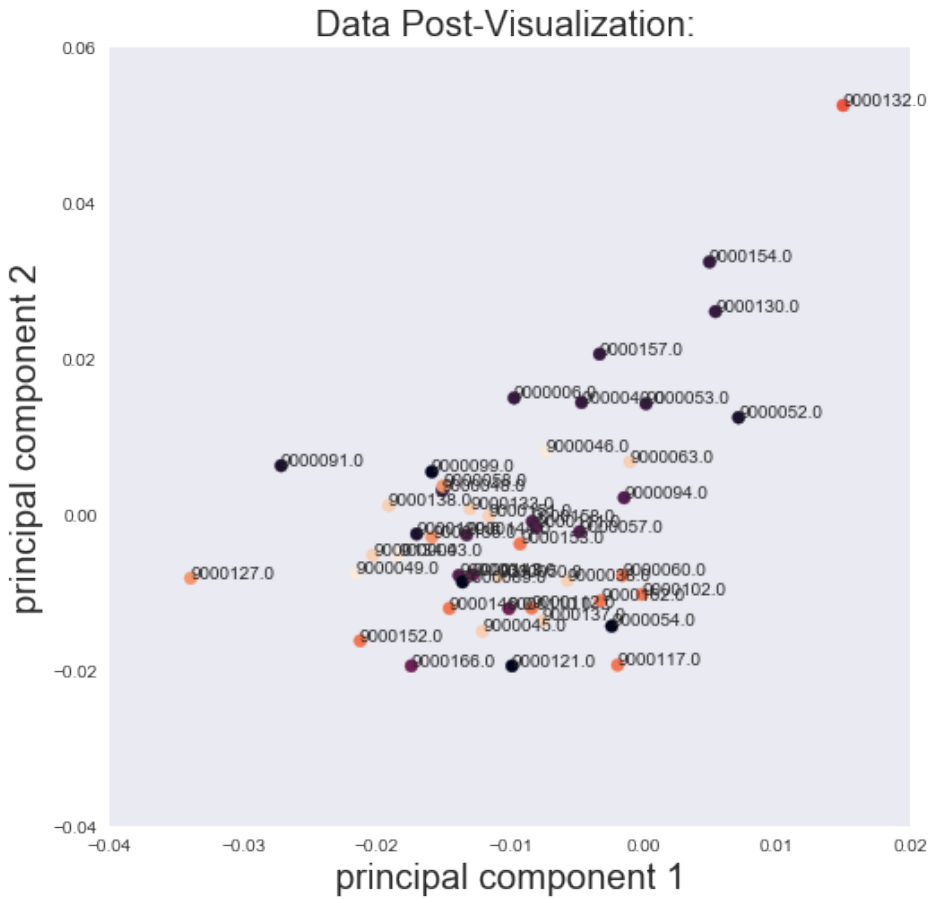


(b) Count, K = 15

K-Means: SSE



(a) SSE Profile



(b) Annotated Data , K = 15

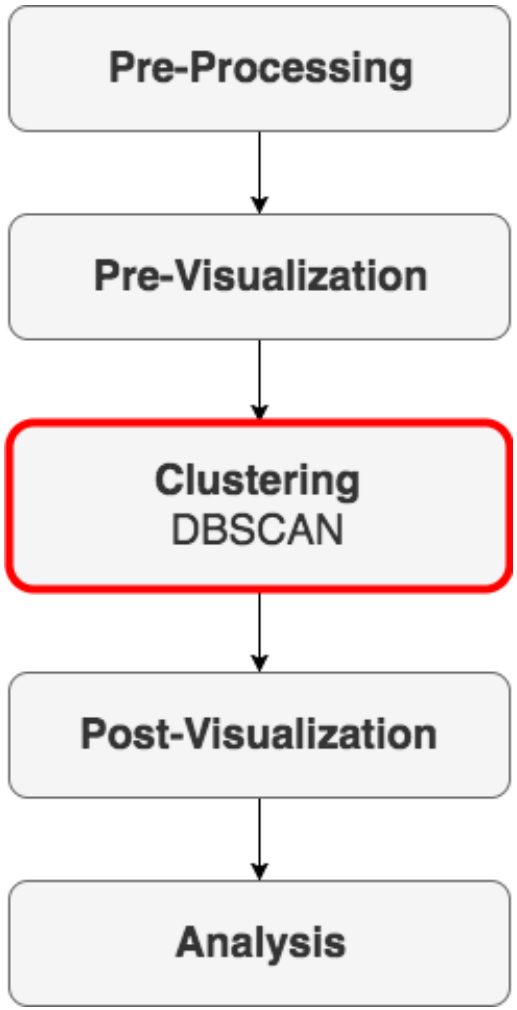
K-Means: Meta-Data of Documents in a Cluster

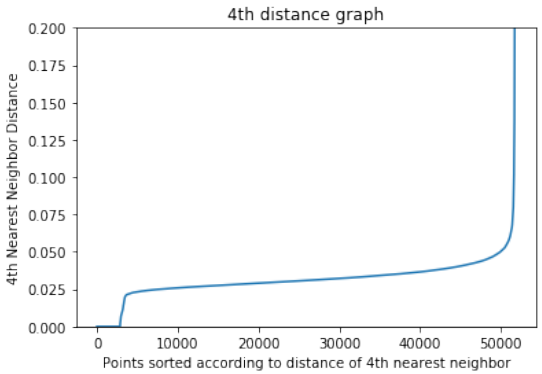
Award Number: **9000157**
Prgm Manager: Richard B. Lambert, Jr.
OCE DIVISION OF OCEAN SCIENCES, GEO DIRECTORATE FOR GEOSCIENCES
NSF Program : 1610 **PHYSICAL OCEANOGRAPHY**
Fld Applictn: 0204000 **Oceanography**

Award Number: **9000154**
Prgm Manager: Russell C. Kelz
OCE DIVISION OF OCEAN SCIENCES, GEO DIRECTORATE FOR GEOSCIENCES
NSF Program : 1610 **PHYSICAL OCEANOGRAPHY**
Fld Applictn: 0204000 **Oceanography**

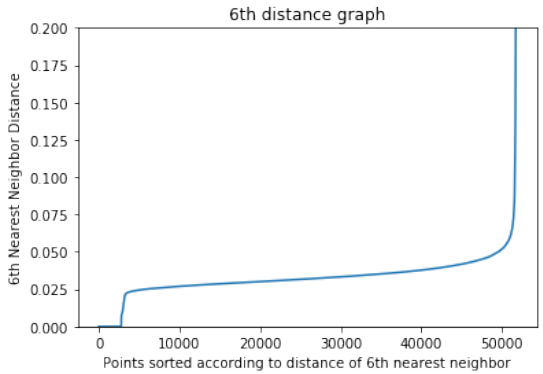
Award Number: **9000130**
Prgm Manager: Emma R. Dieter
OCE DIVISION OF OCEAN SCIENCES, GEO DIRECTORATE FOR GEOSCIENCES
NSF Program : 5411 **SHIP OPERATIONS**
Fld Applictn: 0204000 **Oceanography**

Clustering: DBSCAN





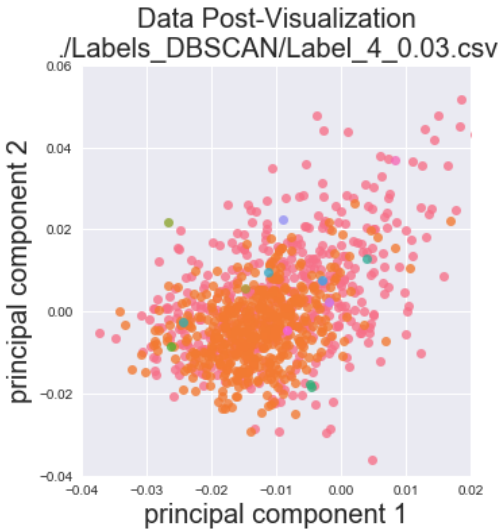
(a) 4-distance-graph



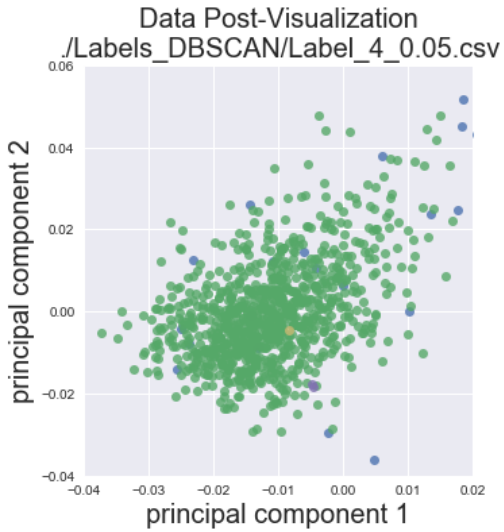
(b) 6-distance-graph

- Parameter Search: K-distance graph
- Optimum parameter at the knee
- MinPts $= 4$, eps $= 0.05$
- Simple implementation: does not use accelerating index structure for neighborhood query
- Time Complexity: $O(n^2)$ instead of $O(nlog(n))$

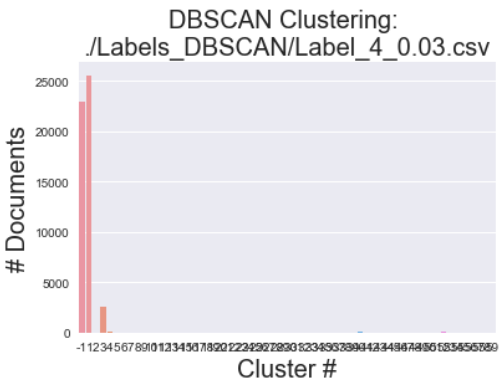
Post-Visualization: DBSCAN



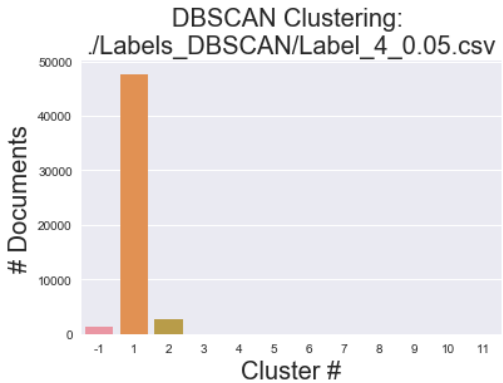
(a) PRM: (4, 0.03)



(b) PRM: (4, 0.05)

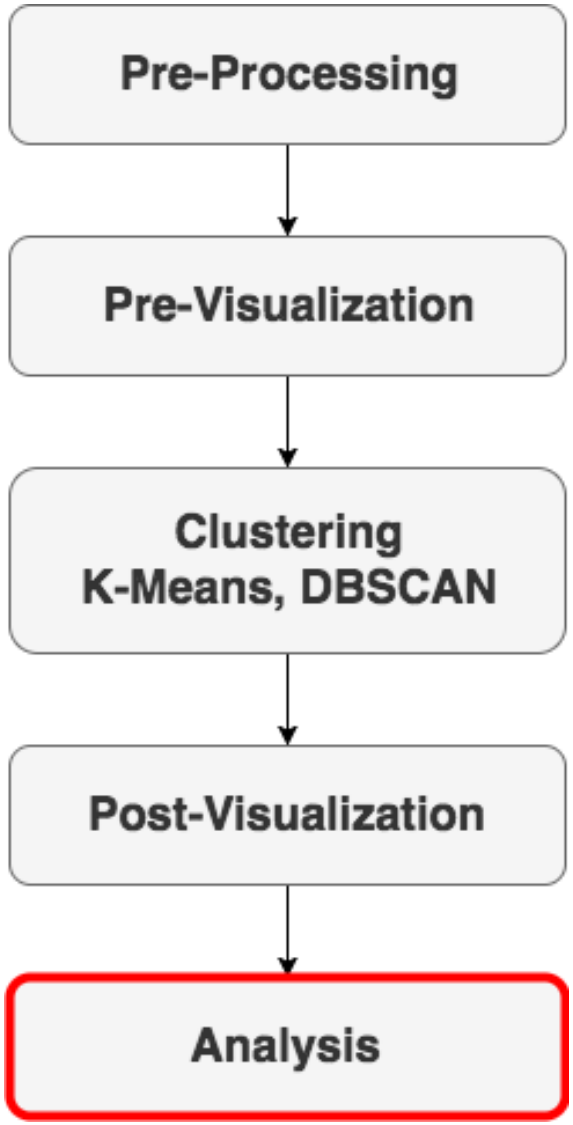


(a) PRM: (4, 0.03)



(b) PRM: (4, 0.05)

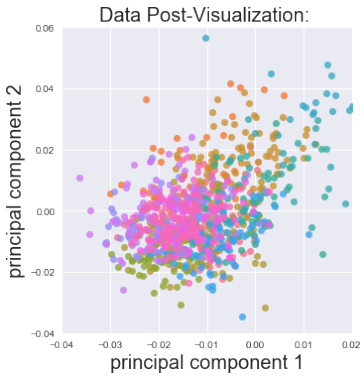
Analysis



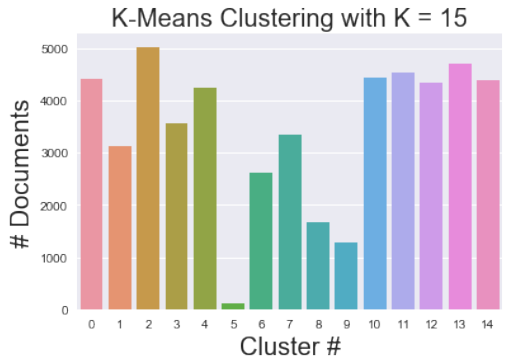
Comparison of K-Means and DBSCAN Clustering based on

- Distribution of Documents
- SSE Error

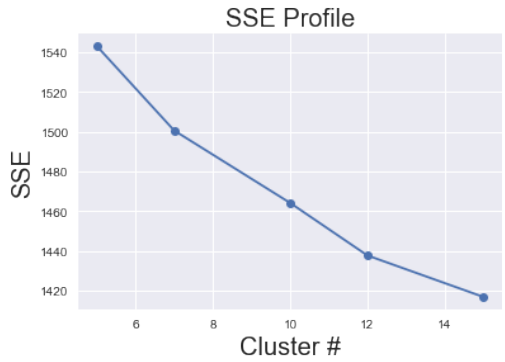
Analysis: Comparison



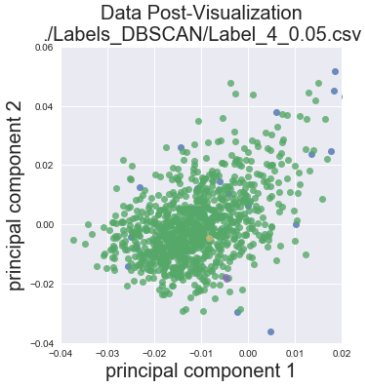
(a) K = 15



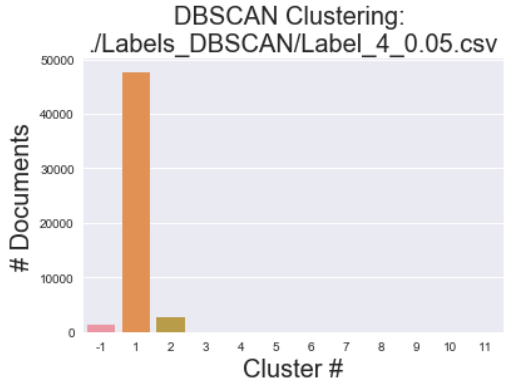
(b) K = 15



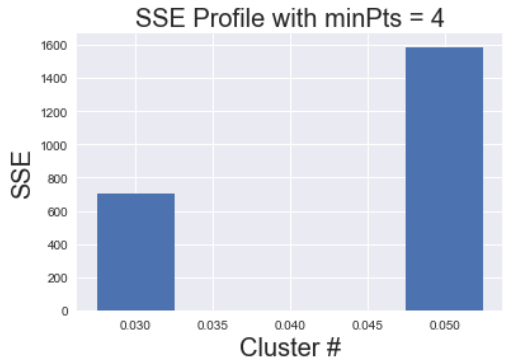
(c) SSE



(a) PRM: (4, 0.05)



(b) PRM: (4, 0.05)



(c) SSE

Conclusion

- K-Means seem to work better than DBSCAN for current dataset
- DBSCAN Failure: High dimension data
- One can try with lower dimension data using PCA