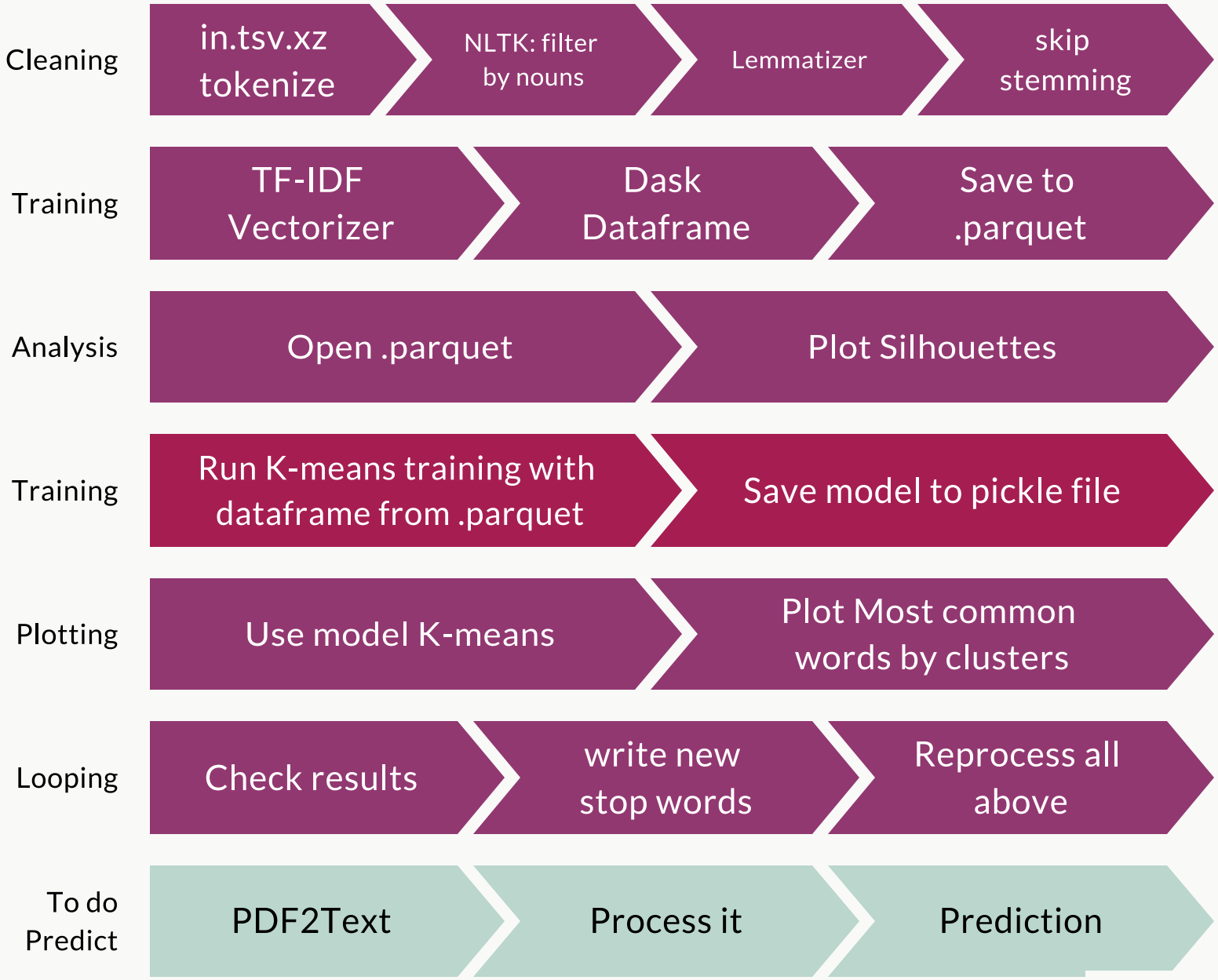
AI DOCLUSTERING

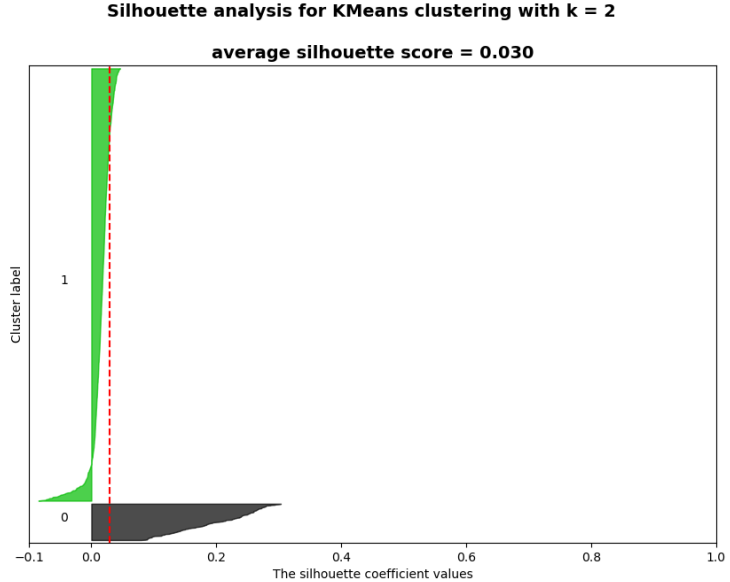
by Ke Thien

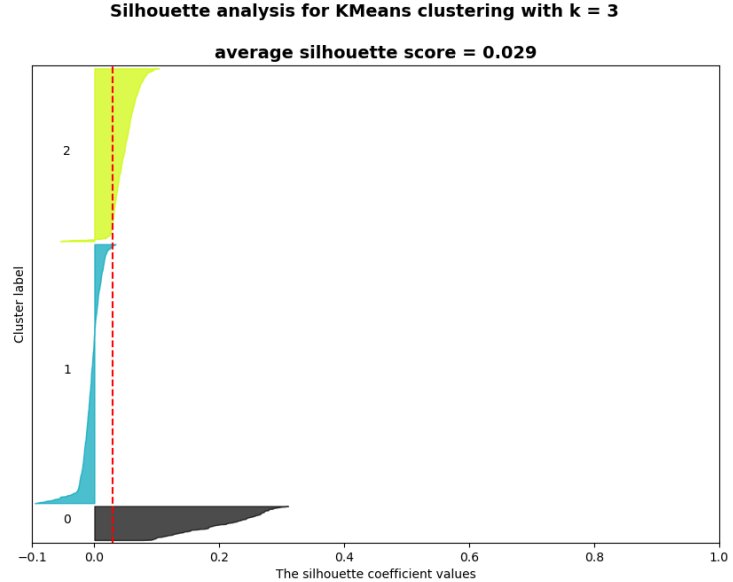
PROCESS TIMELINE

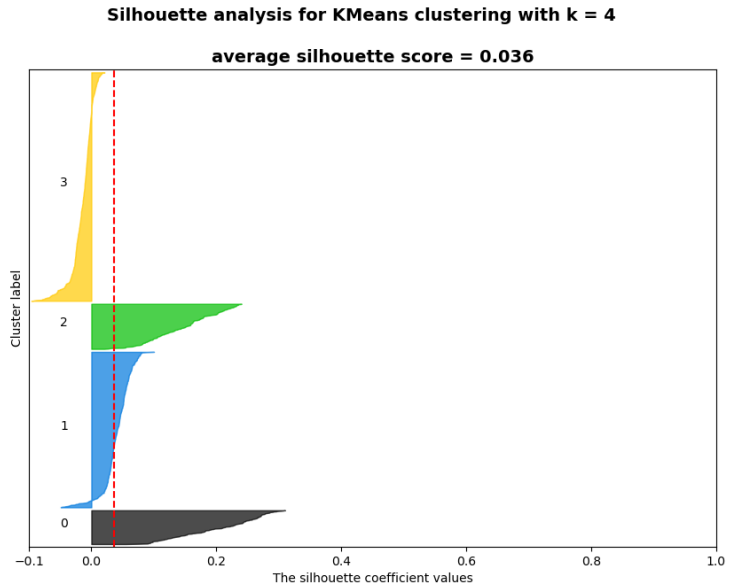


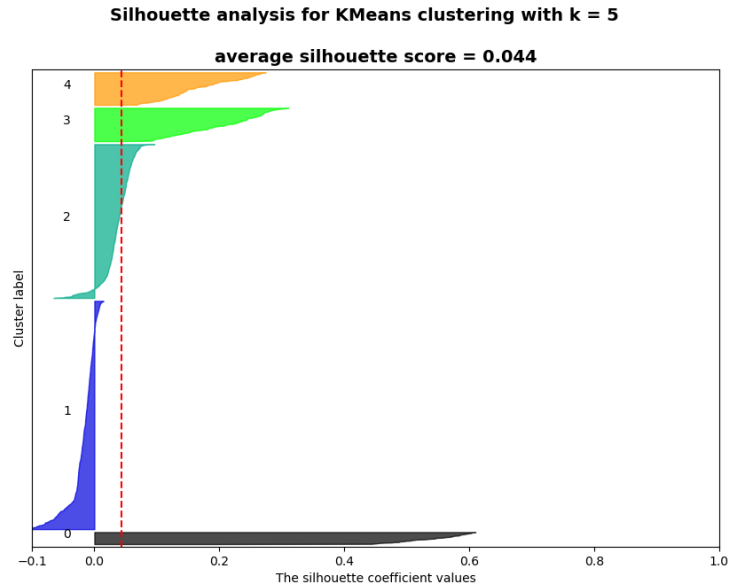
Silhouette Analysis

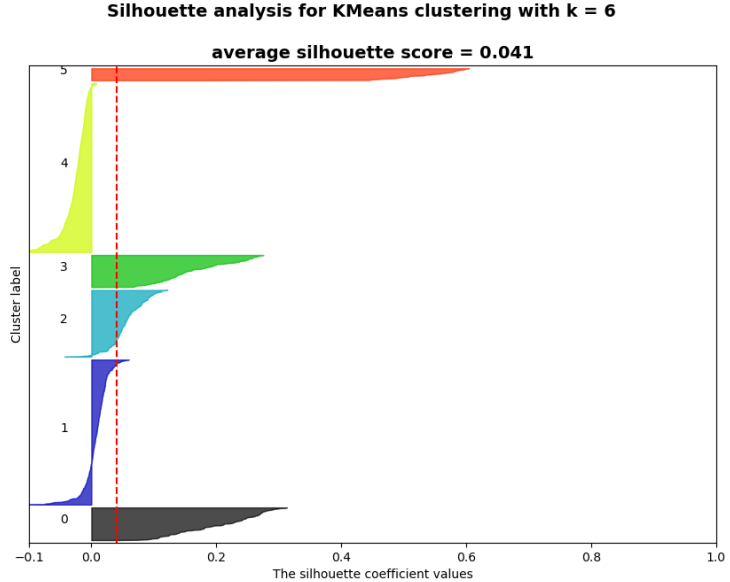
Choose the best k

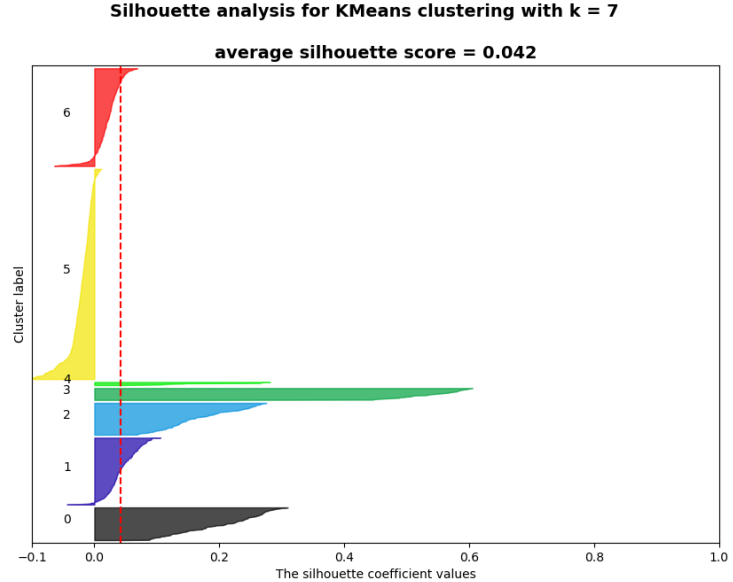


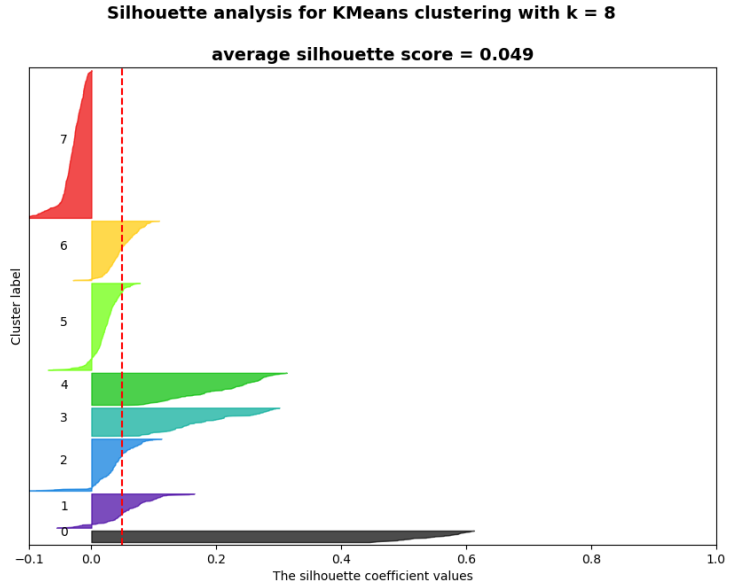


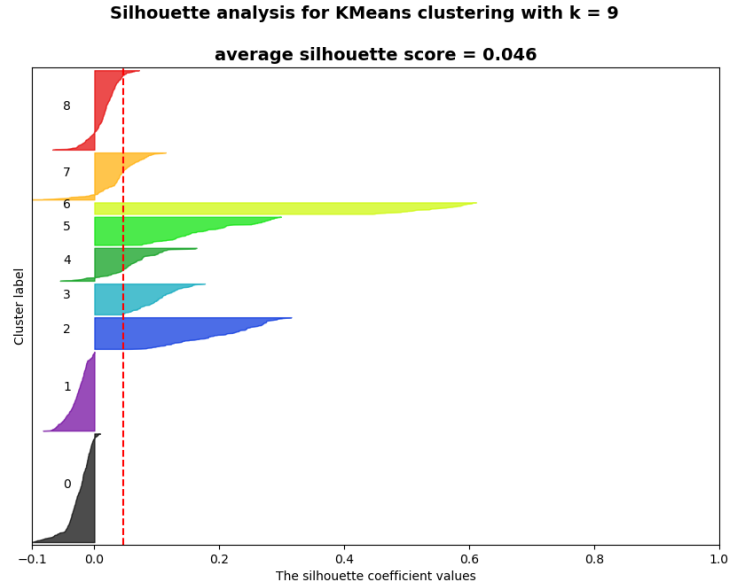






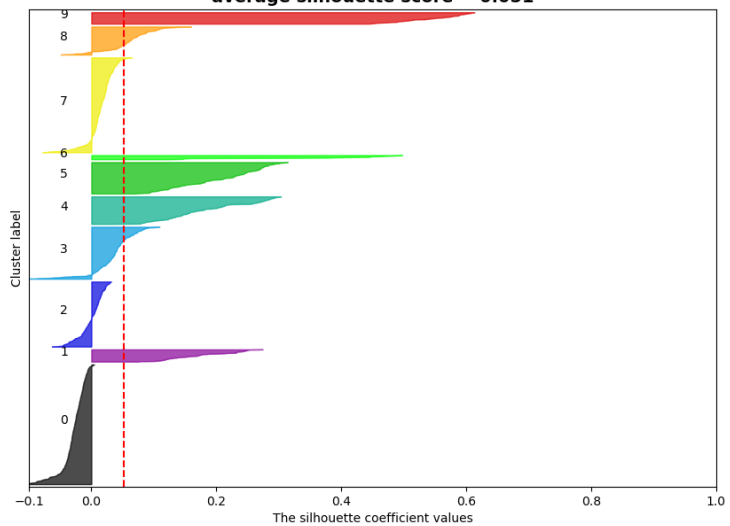






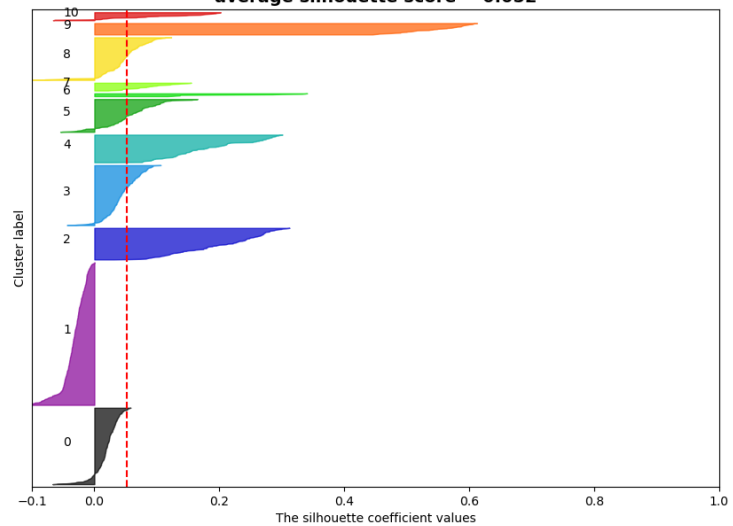
Silhouette analysis for KMeans clustering with k = 10

average silhouette score = 0.051



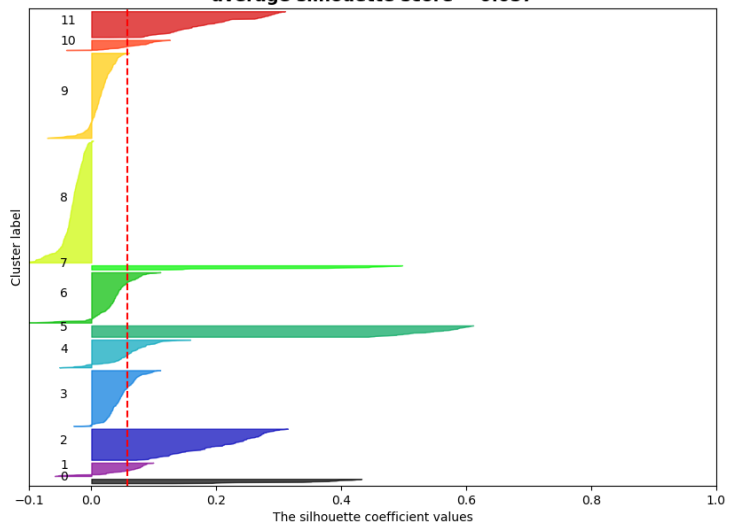
Silhouette analysis for KMeans clustering with k = 11

average silhouette score = 0.052



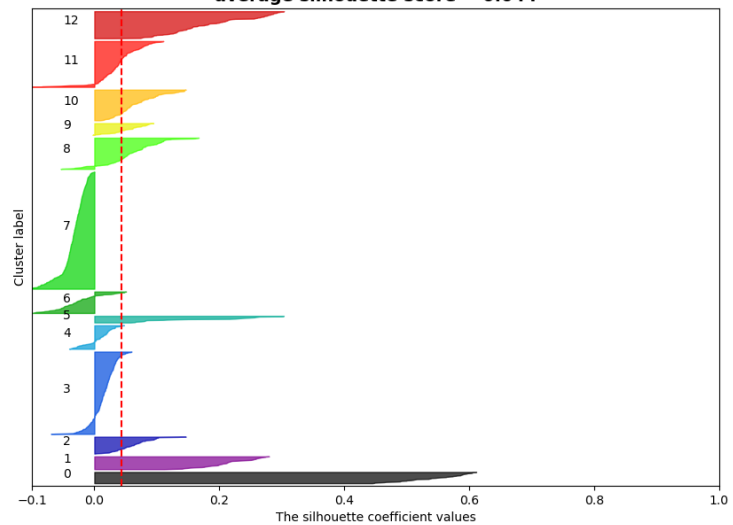
Silhouette analysis for KMeans clustering with k = 12

average silhouette score = 0.057



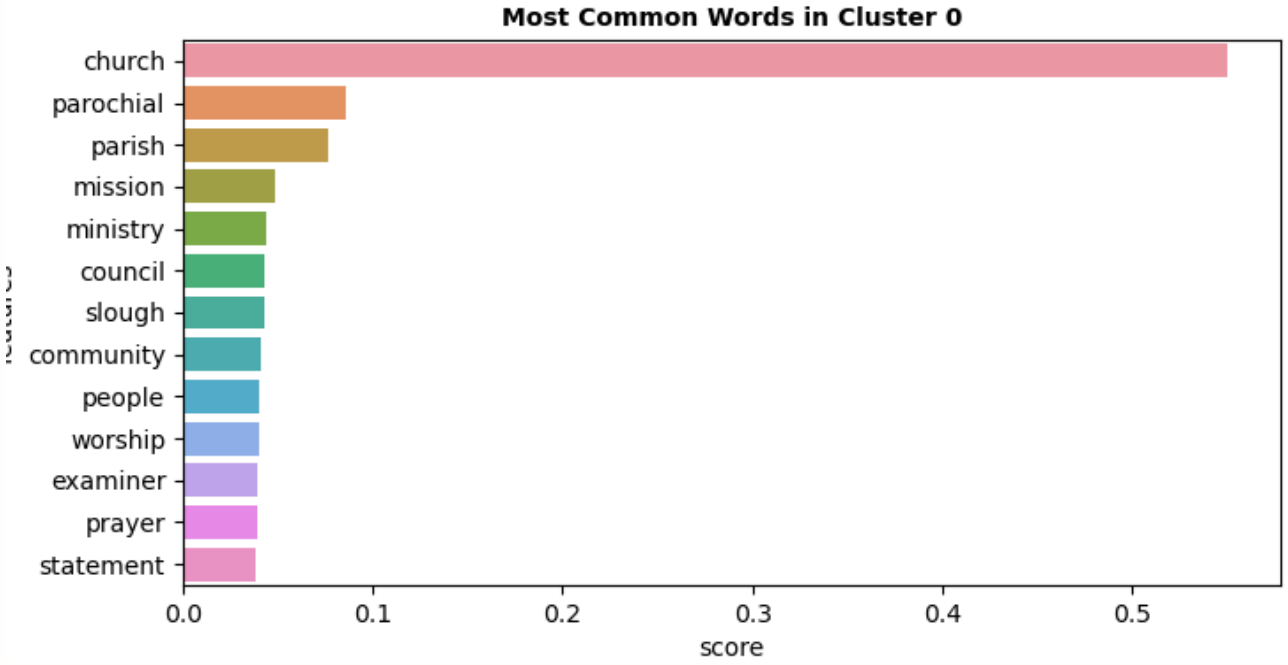
Silhouette analysis for KMeans clustering with k = 13

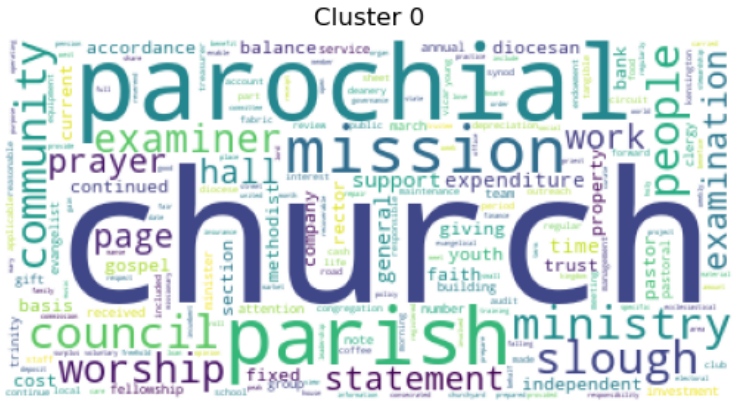
average silhouette score = 0.044

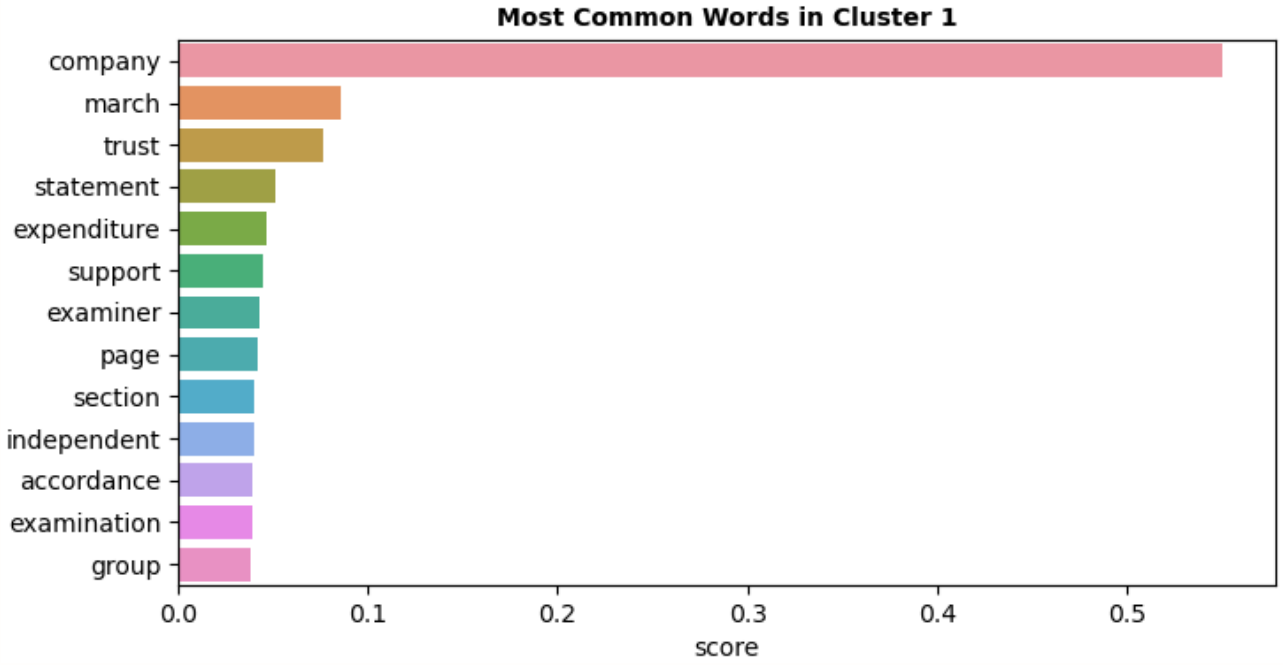


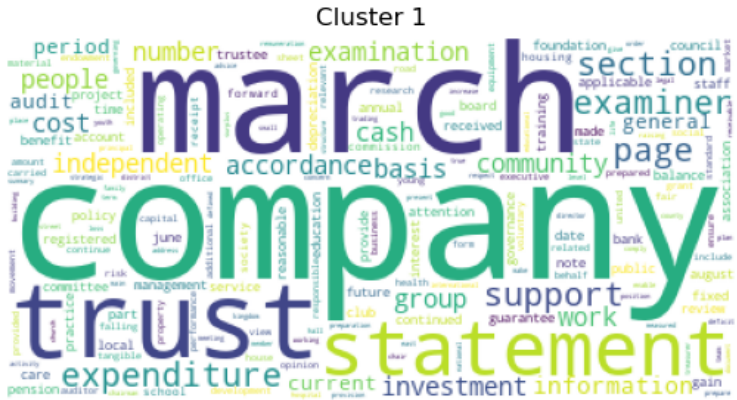
Plot of most common words

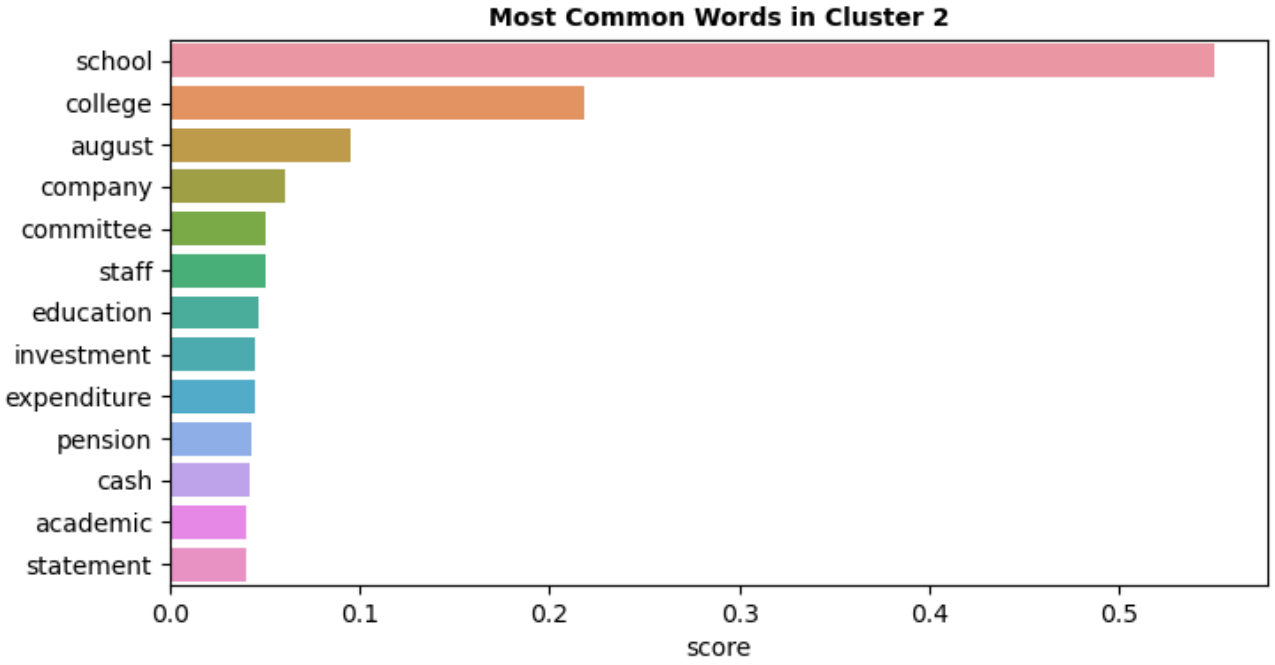
by cluster with k=3

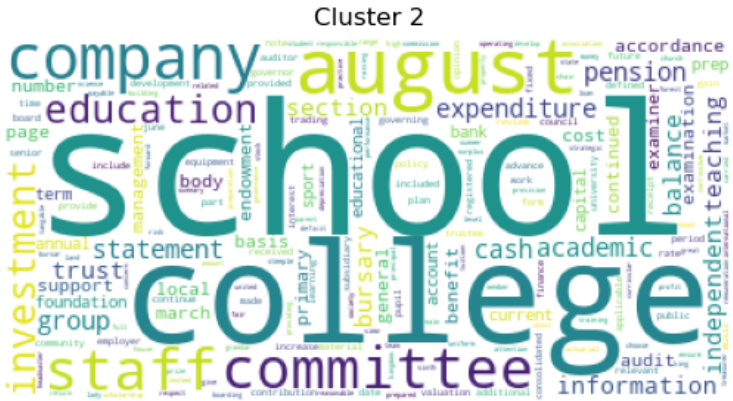












MY DIFFICULTIES

- ## Using Dask

  It add more layers of complexities, lot of frustrations. But I keep trying and it kinda works

- ## Package dependencies

  When switching computers, differences of versions, missing packages sucks.
  Next time: use docker from the start

- ## Losing time when trying to save it

  By trying to use package that looks good but don't works properly.

  e.g.: Yellobrick that uninstall latest numpy to reinstall old version of numpy

- ## Transfer of knowledge

  From simple examples to my own code

THANK YOU