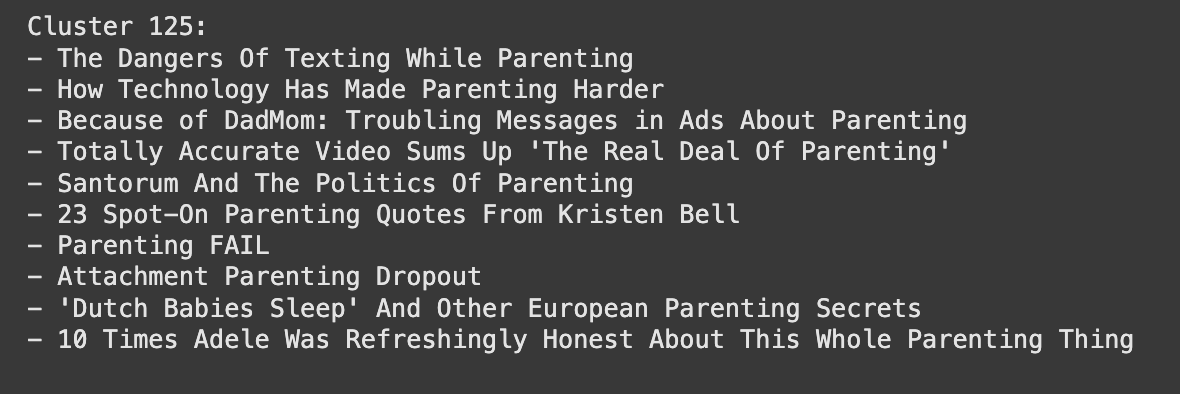
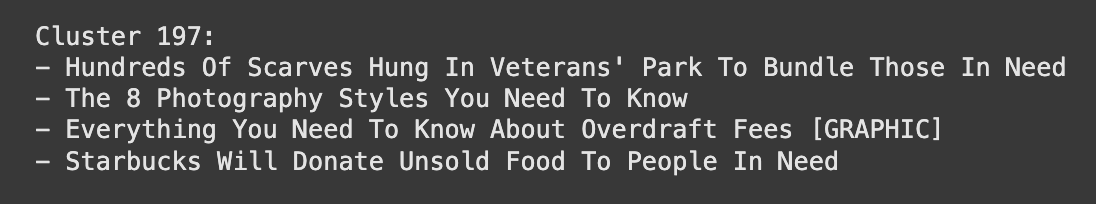
Assessment 3: Clustering News Headlines

For K-Means clustering, I got efficient cluster to be 146 by using KneedLocator to locate efficiently the cluster. Important insights which I got was that the cluster is formed according to certain specific words like PHOTOS and various words like in this cluster:

We could observe that there is a common word: Parenting (both upper and lower case) in all the sentences. This might be because TF-IDF gives more weightage to uncommon words like here the words might be parenting.

For Agglomerative Hierarchical Clustering is used for Hierarchical Clustering by taking samples for 1000 as random. It was difficult to get the effective number of clusters from seeing the dendogram. There might be some error, but I got 225 as number of clusters. The increment might be due to random samples taken; they might not effectively take cluster sentences.



In this also, we notice that there was Need word common in all sentences. The same explanation goes for this, but for others it was observed that there was no common word in the sentences so there might be some error or it might cluster based on other things.

This is the overall analysis done for both K-Means and Hierarchical Clustering .