**Project 1**

**STQD6114 – Unstructured Data Analytics**

**P152419 – Hazim Fitri Bin Ahmad Faudzi**

**Part 2 – Task 2**

Fifty news articles were analyzed using unsupervised clustering to discover underlying thematic groupings. The articles were first transformed into high-dimensional vectors using the Term Frequency–Inverse Document Frequency (TF‑IDF) representation, which weights words by their importance in each article. No predefined labels or categories were provided, so the clustering aimed to let the data reveal its own topic clusters. Three clustering algorithms with different approaches were applied: K-Means (a centroid-based algorithm), Hierarchical Agglomerative Clustering (a linkage-based method), and DBSCAN (Density-Based Spatial Clustering of Applications with Noise, a density-based algorithm). These methods were chosen to compare how each partitions the articles and to identify prominent themes or topics in the news dataset. Each clustering result is discussed below, followed by a comparison of their performance in grouping articles and handling outlier articles.

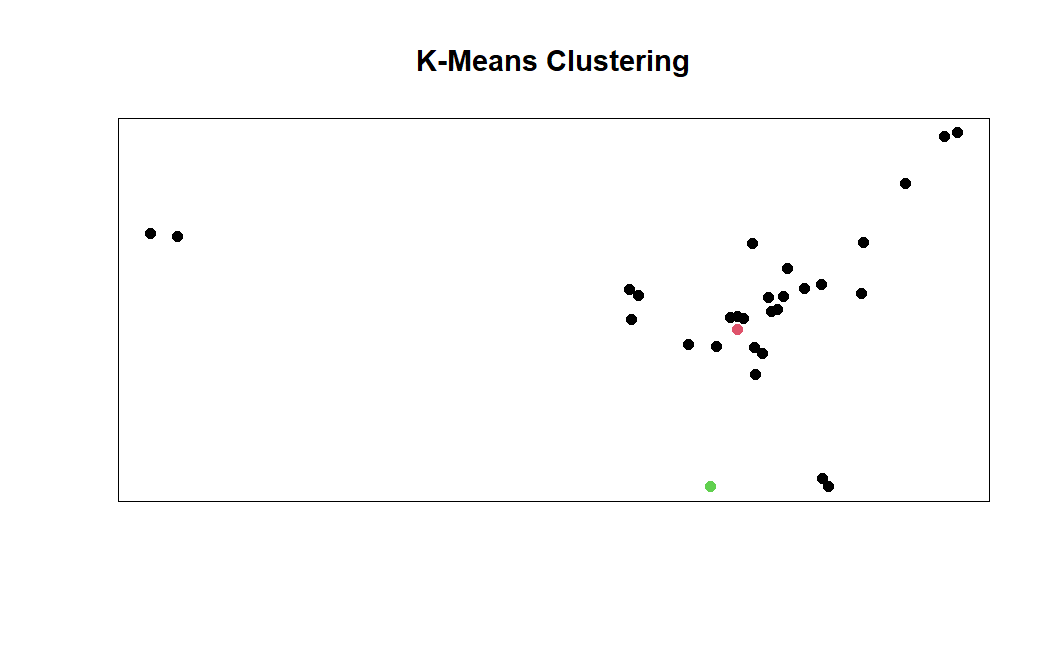


Figure 1 K-Means Clustering

Hierarchical agglomerative clustering produced a more granular grouping of the news articles. In the hierarchical clustering plot, the articles naturally formed several distinct clusters, each indicated by a different color. The grouping is clearer and more separated than in the K-Means result – articles that were all in one large K-Means cluster are split into smaller clusters here. This method does not require pre-specifying the number of clusters; instead, a dendrogram (a tree of merges) is cut at a chosen level to yield clusters. For this analysis, the cut produced multiple clusters, perhaps three or more, indicating that the algorithm found multiple meaningful groupings in the data. Each colored cluster likely corresponds to a specific theme or sub-topic among the articles. Notably, the previously outlier articles (which were far apart in the K-Means plot) appear here as their own small clusters or single-member clusters. This means the hierarchical approach isolated those unique articles into separate groups rather than forcing them into a larger cluster. While this yields very clear group separation, it can lead to over-segmentation – the formation of perhaps too many clusters. In other words, hierarchical clustering may split what could be considered one theme into several sub-clusters if the articles within that theme are somewhat dissimilar. Overall, the hierarchical clustering result suggests a detailed thematic structure: it cleanly divides the news data into distinct groups (including highlighting outlier articles as standalone clusters), potentially corresponding to nuanced topics or categories in the news.

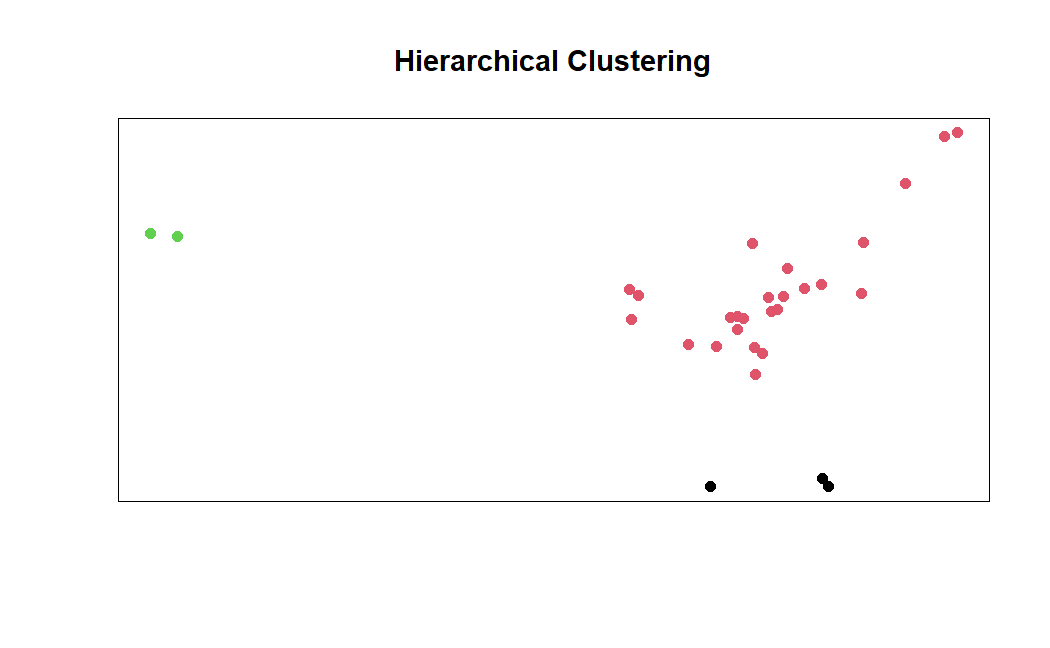


Figure 1 K-Means Clustering

Hierarchical agglomerative clustering produced a more granular grouping of the news articles. In the hierarchical clustering plot, the articles naturally formed several distinct clusters, each indicated by a different color. The grouping is clearer and more separated than in the K-Means result – articles that were all in one large K-Means cluster are split into smaller clusters here. This method does not require pre-specifying the number of clusters; instead, a dendrogram (a tree of merges) is cut at a chosen level to yield clusters. For this analysis, the cut produced multiple clusters, perhaps three or more, indicating that the algorithm found multiple meaningful groupings in the data. Each colored cluster likely corresponds to a specific theme or sub-topic among the articles. Notably, the previously outlier articles (which were far apart in the K-Means plot) appear here as their own small clusters or single-member clusters. This means the hierarchical approach isolated those unique articles into separate groups rather than forcing them into a larger cluster. While this yields very clear group separation, it can lead to over-segmentation – the formation of perhaps too many clusters. In other words, hierarchical clustering may split what could be considered one theme into several sub-clusters if the articles within that theme are somewhat dissimilar. Overall, the hierarchical clustering result suggests a detailed thematic structure: it cleanly divides the news data into distinct groups (including highlighting outlier articles as standalone clusters), potentially corresponding to nuanced topics or categories in the news.

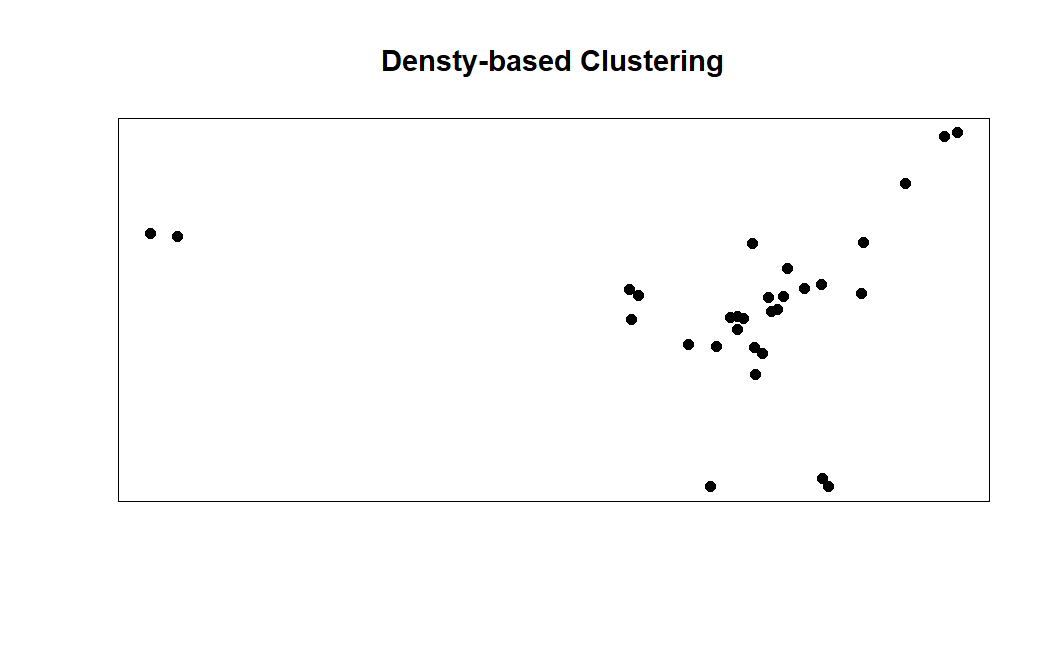


Figure 1 K-Means Clustering

DBSCAN clustering resulted in a markedly different outcome, characterized by one dominant cluster and several points classified as noise. In the DBSCAN plot, nearly all the article points appear as part of a single large cluster (plotted with one color), while a few points do not belong to any cluster. These unassigned points are treated as noise by the algorithm, meaning they did not have a sufficient number of neighboring points within the specified distance (ε) to form their own clusters. In practical terms, those noise points correspond to articles that are highly dissimilar to all others (isolated in the TF‑IDF feature space), reinforcing their status as outliers. The fact that DBSCAN found only one substantial cluster implies that most articles are densely grouped together, likely sharing a broad common theme. Unlike K-Means and hierarchical clustering, DBSCAN did not split this large group into smaller clusters; it was less sensitive to any sub-structure within that big cluster. This is partly due to how DBSCAN works: it merges points into the same cluster if they are within the density radius, so if the majority of articles form one contiguous dense region, DBSCAN will regard them as one cluster. It will only form a new cluster if a set of points is sufficiently dense and separate from the rest. Here, no strong secondary clusters were identified – any potential smaller theme groups were either absorbed into the main cluster or not dense enough and thus labeled as noise. Consequently, the DBSCAN result suggests one prevailing theme encompassing most articles, with a handful of articles so unique that they stand alone as noise. This outcome highlights DBSCAN’s strength in flagging outliers explicitly, but also its tendency to under-segment when the data does not contain well-separated dense regions.

Overall, the clustering analysis highlights how each method uncovers themes at different levels of granularity. K-Means provided a broad-strokes segmentation of the news articles, identifying a few major themes but glossing over niche topics. Hierarchical clustering offered a finer segmentation, distinguishing even closely related articles into separate thematic groups (useful for detailed topic discovery, albeit with the risk of over-segmentation). DBSCAN confirmed the dominant theme by grouping the bulk of articles together and was effective in flagging unique articles as outliers, though it did not separate the dataset into multiple topic clusters beyond the main one. In a thesis context, these findings suggest that the choice of clustering algorithm can significantly influence the interpretation of underlying themes. For a dataset of this size and type, K-Means and hierarchical clustering are effective for discovering distinct topic clusters (with hierarchical giving more nuanced splits), while DBSCAN excels in identifying whether a single theme prevails and which items do not belong to any theme cluster. Each method’s handling of sparse or outlier points also provides insight: outliers may either subtly influence broad clusters (K-Means), form their own small clusters (hierarchical), or be explicitly recognized as noise (DBSCAN). By comparing these approaches, we gain a comprehensive understanding of the thematic structure in the news articles and the robustness of each clustering method in dealing with heterogeneous textual data.