Hausarbeit im Modul „Data Science und Machine Learning“ WS 23/24: Assignment 3

The Social Network Data Set

The dataset contains a random sample of 15,000 high school students with profiles on a popular social network in 2006. The data was collected uniformly between 2006 and 2009. The data were automatically crawled from the social network and further processed via text mining to identify student interests. The 37 most dominant words in the entire dataset were placed (e.g., football, shopping). For each student, the final dataset shows how many times each word appears in that student's profile. Other attributes include graduation year (gradyear), gender (gender), age at the time of the survey (age), and the number of contacts on the social network (NumberOffriends). The goal of this task is to identify students with similar interests for marketing purposes and analyze the respective clusters.

Tasks

1. Cluster the existing data based on appropriate procedures. Justify how you would cluster the data set based on the results.
2. Describe the clusters found based on the characteristics of the attributes of the customers and the cluster size.
3. What recommendation could you give the marketing department based on your analysis?