Data Mining Assignment 4 – Clustering

Problem Setting

For this assignment, you will work with a dataset of food product reviews from Amazon, available in the file **“reviews.csv”**. The goal is to discover groupings of reviews by clustering their textual content. Clustering is a key task in unsupervised learning and can be used to group similar customer feedback, detect themes, or identify outliers in large collections of text.

Your task is to apply and evaluate different clustering techniques and analyze what kind of reviews are grouped together. Before applying clustering algorithms, you need to preprocess the review texts. This includes cleaning and transforming the raw text into a numerical format suitable for clustering.

Task 1: Data Exploration and Preprocessing

Start by inspecting the dataset, are there any groups you expect to find?

Continue with preprocessing the data: Since the reviews are in free-text form, you must preprocess them to make them usable by clustering algorithms. One of the simplest approaches is the "bag-of-words" model: a matrix where each row is a review and each column is a word. The cells contain how often each word occurs in a review. For example:

1. This chocolate was absolutely delicious
2. Terrible taste and horrible packaging

The bag-of-words matrix might look like:

chocolate	**delicious**	**taste**	**horrible**	**packaging**	**absolutely**
1	1	0	0	0	1
0	0	1	1	1	0

The minimum requirement for this task is to preprocess the data into a bag-of-words representation. You will receive additional credit for additional preprocessing steps, such as:

- Removing stopwords (e.g., "the", "and", "is")
- Stemming or lemmatization of words (e.g., "tasting" → "taste")
- Removing rare or overly frequent terms
- Using more advanced text representations such as word or sentence embeddings (e.g., Word2Vec, BERT)
- Any other techniques you find appropriate based on patterns you observe

In your report, clearly describe and motivate the preprocessing steps you took.

Task 2: Clustering and Evaluation

Once your data is prepared, apply clustering algorithms to your dataset. You are encouraged to experiment with different combinations of:

- Clustering algorithms
- Distance functions
- Number of clusters

For this assignment, do **not exceed 10 clusters**. After clustering the reviews, analyze:

- What combination of clustering method, distance function, and number of clusters yielded the most interesting or interpretable results?

- Based on what metrics did you evaluate the results (e.g., silhouette score)?

- What patterns or topics do the different clusters seem to represent?

Summarize your findings in your report. Alongside the report, also submit the provided **“clusters.csv”** file. Fill this file in by assigning a cluster ID to each review based on your best clustering model. Make sure to preserve the original order of the reviews as they appear in **“reviews.csv”**.

Submission Details

- Implement your code in any language (Python recommended). Use libraries like:
  1. **pandas** for data manipulation
  2. **sklearn** for preprocessing and evaluation
  3. **nltk** for text processing
  4. **matplotlib** or **seaborn** for visualizations
- Include your findings in a concise report of maximum 4 pages.
- Submit both the code, your results and the report(pdf) as a **.zip** file via Blackboard by **23/05/2025**.
- The ZIP file should include your full name like, **Lastname_Firstname-studentNumber.zip.**

Evaluation Criteria

Your submission will be evaluated based on the following grading rubric.

	Less than 7	**7 to 10**	**10 to 13**	**14 to 17**	**18 to 20**
Writing Style	The report is very confusing; the writing style is below average.	At places the report is not very clear; there is a lack of clear structure.	The text is overall clear although some parts could be improved.	Most of the text is easy to follow, findings are explained in a clear way.	The text is very clearly and concisely written, illustrative examples and figures are not overused, but added where needed.
Task 1	Task was not completed.	Task was only completed in a minimal way. Incorrect or insufficient argumentation is given for the preprocessing choices that were made. The result analysis is not clear.	Bag of words approach was implemented correctly, but not much further preprocessing steps were done.	Some additional preprocessing steps were executed on the textual data, motivation for the choices are provided.	The student executed new and creative preprocessing steps, that were not suggested in the assignment, but are well motivated.
Task 2	Task was not completed.	Task was only completed in a minimal way. There are some mistakes in the evaluation of the cluster validities.	Basic clustering approaches were tried out, some information about the evaluation of each clustering were given.	Interesting clustering approaches were tried out and well motivated. Results were evaluated rigorously.	Interesting clustering approaches were tried out and well motivated. Results were evaluated rigorously, focussing both on quantitative and qualitative evaluation methods.
Code	Code raises many errors.	Code raises some errors or is very unclear.	Code runs but lacks clear structure and readability, only little documentation is given.	Code is readable and sufficiently documented.	Code is very readable and well documented. It is structured in a way that only by (un)commenting single lines, the code for the different tasks can be run.