Data Mining Assignment 3 – Clustering

Problem Setting: For this assignment you will work with a dataset consisting of news articles. The goal for you is to divide these articles into meaningful clusters, experimenting with different data preprocessing techniques, clustering algorithms and distance functions.

Task 1: Start with inspecting the dataset. Upon clustering the dataset, what information do you expect the different clusters to represent?

Continue with preprocessing the data: Since the news articles consist of texts, you must preprocess them to make them understandable to a computer, basically transforming the text into numbers. One of the simplest methods to do so is the "bag of words" approach: essentially you create a matrix, where each column represents a word in the corpus. Each row in the matrix will then describe an article by its word counts.

For example:

1. Roses are red, dark red
2. Violets are blue

#	roses	are	red	dark	violets	blue
1	1	1	2	1	0	0
2	0	1	0	0	1	1

For a more elaborate tutorial how to do this in python we refer to: https://medium.com/analytics-vidhya/introduction-to-text-classification-in-python-659eccf6b2e

The minimum requirement for this part of the assignment is to preprocess the data according to this bag-of-word approach. You can get extra points if you do additional preprocessing steps to create a more suitable representation of the data. This could for example include:

- Stemming or lemmatization of tokens (see tutorial)
- Removing non-informative functionwords like 'the' or 'and' from your dataset
- Extracting other relevant information from each document (e.g., does it contain any names, any numeric values, etc.)
- *Feel free to come up with your own creative ideas, also based on some patterns you might find in the data!*

Describe and motivate your preprocessing choices in your report.

Task 2: Apply now some clustering algorithms on your dataset, experimenting with different clustering algorithms, distance functions and number of clusters. For this assignment we ask you to not generate more than 10 clusters. Summarize your findings in the report, reflecting at least on the following points:

- What combination of cluster algorithm, distance function and number of clusters yields an interesting result? Based on what metrics did you evaluate this? What could the different clusters represent?

- What are some other combinations that you tried out? Report some aggregated results of at least 2 of them. Why do you deem these results as less useful?

With your report also hand in a file called 'clusters.xlsx'. The file should consists of the original news articles data, but with an extra column, denoting the cluster number that each article got assigned to.

Some Notes on your Code: You may implement your code in any programming language of your choice, but we do recommend Python as this is the most common language used for Data Mining tasks. You might find these libraries useful:

`pandas` for data storing/preprocessing
`sklearn` to build and evaluate clustering models
`matplotlib` or `seaborn` for generating figures
`nltk` for processing textual data

Handing in instructions: Put all your findings in a report of about 4 pages (there is no hard page limit, but be concise). The deadline for handing in your report, the code and the clusters.xlsx file is 24-05-2024

Grading Rubric:

Below you can find the rubric we will use to grade your assignment. Please consult this rubric to understand what is expected of you.

	Less than 7	**7 to 10**	**10 to 13**	**14 to 17**	**18 to 20**
Layout	Layout of the report is sloppy, report is hard to read, there are many typos		Layout is okay, but at some parts there are typos or section headings, figure titles etc. are missing	The layout of the report is good, there are barely any typos	The layout is excellent
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.	Bag of words approach was implemented correctly, but not	Some additional preprocessing steps were	The student executed new and creative

		Incorrect or insufficient argumentation is given for the preprocessing choices that were made. The result analysis is not clear.	much further preprocessing steps were done	executed on the textual data, motivation for the choices are provided	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 different tasks can be run