Unsupervised Learning / Data Analytics 1

WT 2024/2025

Case Study

General information.
These exercises are meant to be solved within a group of two to a maximum of four students. Assign yourself to a group by clicking on the *Group Selection*-button in Learnweb and selecting a free or not-full group. The case study exercise awards 10 additional (bonus) points in total for the *Unsupervised Learning / Data Analytics 1* exams in WT 2024/25 and ST 2025. Although these points are supplementary, meaning that you can achieve a 1.0 / A grade (i.e., full points) without completing these exercises, we strongly encourage you to undertake the following tasks for hands-on experience in various topics of the lecture and to earn extra points for the exam.
For all students who participate in the module *Unsupervised Learning*, this case study is mandatory to complete the full module. Nevertheless, the 10 bonus points also count for your exam.

Exercise. Analyze the Provided Datasets; submit by Jan. 26 [10 Points]

We provide you with a large data set of social media messages. The file *dataset.json* contains social media data crawled from a well-known platform. Some of the metadata has been removed to keep it manageable.
We extracted the following features:

Feature	Description
timestamp	A numerical feature, according to which the data can be ordered (w.r.t. time).
text	A text posted.
text_id	unique ID of the text.
user	nickname/username of the user, who posted the text.
user_id	unique ID of the user.

You can import this dataset with `Python` and `pandas` by using the following code snippet:

```
from pandas import json_normalize
import json

[...]

# load JSON-file
with open('dataset.json', 'r') as file:
    data = json.load(file)

# normalize data (return data frame)
    df = json_normalize(data)
```

In addition, we provide you (as a service ;-)) with a data set that specifies some user relations (as an edge list) in a `.csv` format. That is, for many users of the first data set (and even some more users), you have the *friendship / following* relations. Relations may be considered undirected (that is, as both-sided following). We provide here a plot which gives some rough impression of these data (possibly, this is not the best way to analyze the graph).

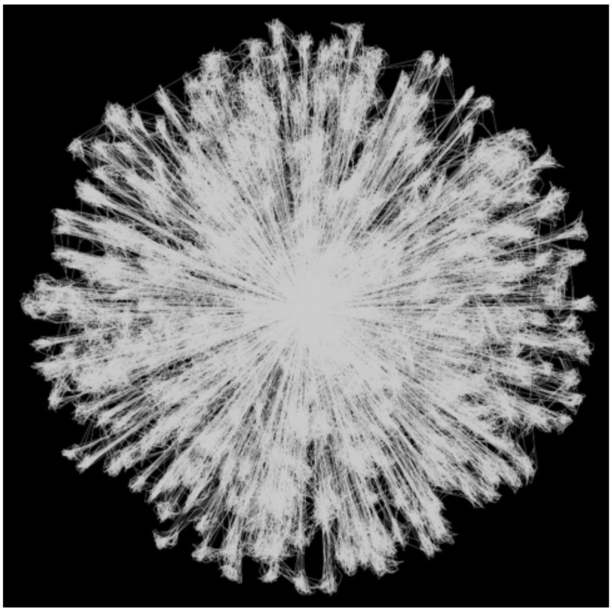

Figure 1: Looks like a snow flake but is the network of the users from our provided dataset.

Your task is: Find out as much as possible about the data and the users. Use methods you know from the lecture and tutorials. Do not hesitate to do own research and go beyond the lecture's content. Please use Python as programming language.

Subsequently,

1. Prepare a poster presentation for your group project. Clearly communicate your results. Use visualization and accessible explanation. Please note: also the form of presentation is part of the grading. Any group member should be able to answer questions regarding the poster. Use the poster template provided via LearnWeb.
2. Polish (i.e. comment and structure) your code and hand it in together with your poster (e.g., Jupyter Notebook).
3. Submit both poster (as `.pdf`) and code (as `.py` or `.ipynb`) by **January 26th at 11:59 PM** via the Learnweb. Submissions in other formats will not be graded.
4. Present your poster (as a group) in the Leo 18 lecture hall on 30 January at 12:15 pm. Participation is mandatory! The poster will be printed by us. We, other faculty members, and possibly other students may ask questions about your work. The ability of group members to answer these will be a major part of the grade.