732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

Laboratory work 6

Instructions

- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report; you are also recommended to show parts of the codes in the flowing text of the report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- Create a report to the lab solutions in RMarkdown. Make sure that it is can be compiled to HTML and that all paths in RMD file are relative to the current directory where the RMD file is located. **Reports that can not be compiled are returned without revision.**
- Put the RMD file and all supporting files into one ZIP archive when you submit it to LISAM.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1. Text Visualization of Amazon reviews

In this assignment you will analyze feedbacks given by customers for watches Casio AMW320R-1EV bought at www.amazon.com . Files Five.txt and OneTwo.txt contain feedbacks of the customers who were pleased and not pleased with their buy, respectively.

1. Visualize word clouds corresponding to Five.txt and OneTwo.txt and make sure that stop words are removed. Which words are mentioned most often?
2. Without filtering stop words, compute TF-IDF values for OneTwo.txt by aggregating each 10 lines into a separate “document”. Afterwards, compute mean TF-IDF values for each word over all documents and visualize them by the word cloud. Compare the plot with the corresponding plot from step 1. What do you think the reason is behind word “watch” being not emphasized in TF-IDF diagram while it is emphasized in the previous word clouds?
3. Aggregate data in chunks of 5 lines and compute sentiment values (by using “afinn” database) for respective chunks in Five.txt and for OneTwo.txt . Produce plots visualizing aggregated sentiment values versus chunk index and make a comparative analysis between these plots. Does sentiment analysis show a connection of the corresponding documents to the kinds of reviews we expect to see in them?
4. Create the phrase nets for Five.Txt and One.Txt with connector words
   - am, is, are, was, were
   - at

When you find an interesting connection between some words, use Word Trees https://www.jasondavies.com/wordtree/ to understand the context better. Note that this link might not work properly in Microsoft Edge (if you are using Windows 10) so use other browsers.

5. Based on the graphs obtained in step 4, comment on the most interesting findings, like:
   - Which properties of this watch are mentioned mostly often?
   - What are satisfied customers talking about?
   - What are unsatisfied customers talking about?
   - What are properties of the watch mentioned by both groups?
   - Can you understand watch characteristics (like size of display, features of the watches) by observing these graphs?

Assignment 2. Interactive analysis of Italian olive oils.

In this assignment, you will continue analyzing data **olive.csv** that you started working with in lab 2. These data contain information about contents of olive oils coming from different regions in Italy. Each observation contains information about

- Region (1=North, 2=South, 3= Sardinia island)
- Area (different Italian regions)

Different acids:

- Palmitic
- …
- Eicosenoic

ATTN: All diagrams that support your judgments should be included to the report

In this assignment, you are assumed to use Plotly without Shiny.

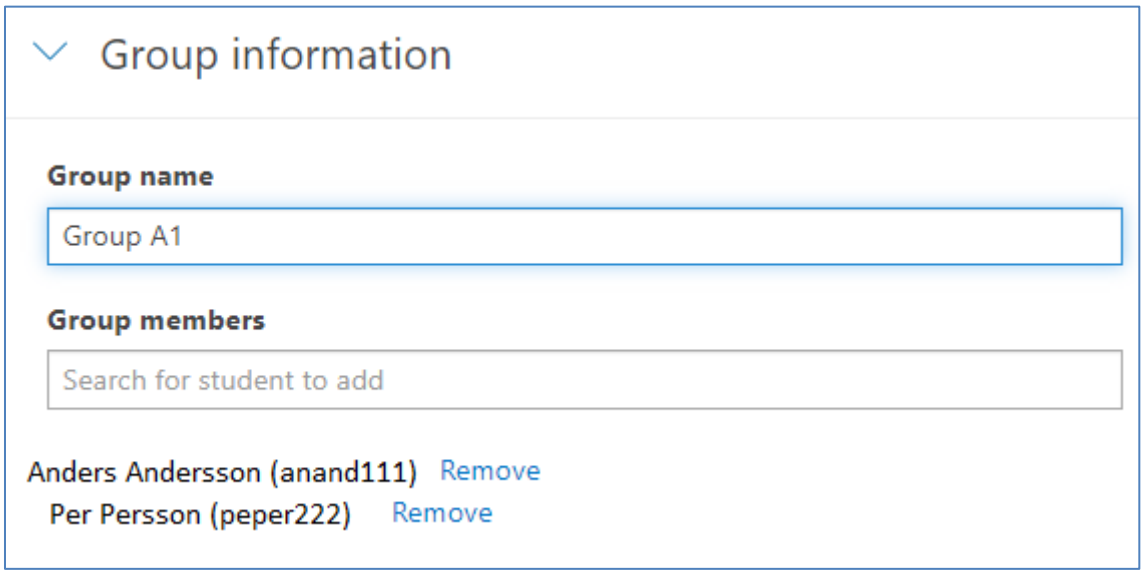
1. Create an interactive scatter plot of the eicosenoic against linoleic. You have probably found a group of observations having unusually low values of eicosenoic. Hover on these observations to find out the exact values of eicosenoic for these observations.
2. Link the scatterplot of (eicosenoic, linoleic) to a bar chart showing Region and a slider that allows to filter the data by the values of stearic. Use persistent brushing to identify the regions that correspond unusually low values of eicosenoic. Use the slider and describe what additional relationships in the data can be found by using it. Report which interaction operators were used in this step.
3. Create linked scatter plots eicosenoic against linoleic and arachidic against linolenic. Which outliers in (arachidic, linolenic) are also outliers in (eicosenoic, linoleic)? Are outliers grouped in some way? Use brushing to demonstrate your findings.
4. Create a parallel coordinate plot for the available eight acids, a linked 3d-scatter plot in which variables are selected by three additional drop boxes and a linked bar chart showing Regions. Use persistent brushing to mark each region by a different color.

Observe the parallel coordinate plot and state which three variables (let's call them influential variables) seem to be mostly reasonable to pick up if one wants to differentiate between the regions. Does the parallel coordinate plot demonstrate that there are clusters among the observations that belong to the same Region? Select the three influential variables in the drop boxes and observe in the 3d-plot whether each Region corresponds to one cluster.

5. Think about which interaction operators are available in step 4 and what interaction operands they are be applied to. Which additional interaction operators can be added to the visualization in step 4 to make it even more efficient/flexible? Based on the analysis in the previous steps, try to suggest a strategy (or, maybe, several strategies) that would use information about the level of acids to discover which regions different oils comes from.

Submission procedure

When submitting the report, remember to specify in LISAM your group name and all group members that wrote the report!



Assume that X is the current lab number, Y is your group number.

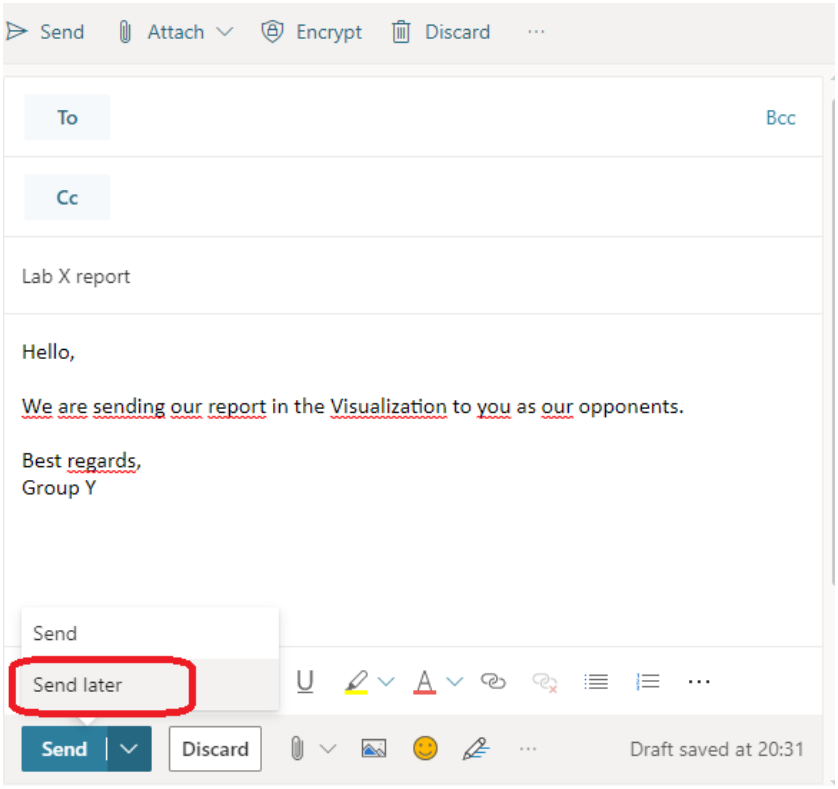
If you are neither speaker nor opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.

732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

If you are a speaker for this lab,

- Make sure that you or your group mate does the following before the deadline:
    1. submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Makes sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
    2. Goes to LISAM→Course Documents→Deadlines.PDF, finds the deadline (date and time) for the current lab.
    3. Goes to LISAM→Course Documents→Seminars.PDF and find the group number of your opponent group
    4. Goes to LISAM→Course Documents→Groups.PDF and finds email addresses of the students in the opponent group
    5. Go to LISAM→Outlook app and in the Outlook web client creates a new message where you
        - Specify Lab X report as a title (X is lab number)
        - Specify email addresses of the opponents in the "To:" field
        - Attach your RMD report and accompanying data files (Note: NOT HTML!)
        - **Important: Click on arrow next to "Send" button, choose "Send Later" and specify the lab deadline as the message delivery time stamp (see figure below)**



If you are opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.

- After the deadline for the lab has passed you should be able to receive the RMD report of the speakers per email. Compile it, read it carefully and prepare (in cooperation with your group comrade) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.