

Group Work Assignment Part 2: **Analyzing unstructured data from your company**

Scenario

In this assignment, you identify data and use cases from your own company (i.e., one group member's company) and create value for stakeholders in the company by conceptualizing and implementing an analytical solution.

Before you start, please discuss the chosen use case with the lecturer. Also use the coaching sessions during the lectures.

Task II: Variety

For the second part of the assignment, please identify a **large text corpus** that company stakeholders need to gain insights. These can be large Excel or csv files containing many text-based fields, large number of E-mails, Text from Websites or social media, or collections of pdf files that can be processed with Natural Language Processing (NLP).

- Identify and document the **analytical requirements** of stakeholders regarding your chosen text corpus. Please describe the chosen use case along the following aspects
 - **Stakeholders:** who are the people that have the analytical/information needs?
 - **Trigger:** in which context(s) will stakeholders need to access the corpus?
 - **Questions:** what analytical needs or questions do the stakeholders have in the described context? Please formulate these needs in the form of questions!
 - **Expected result:** which type of interaction facility and output will satisfy the stakeholders' information needs?

If your use case can be related to searching for information, getting meaningful insights, text mining, clustering of information, efficient processing or filtering of large amount of text or applications of Retrieval Augmented Generation (RAG) with the help of Large Language Models (LLMs).

- **Deliverable 1:**
An implementation of a **prototype**, which can be in the form of a Jupyter notebook or in form of an Orange Workflow or in form of other NLP frameworks such as LangChain.
- **Deliverable 2:**
A **slide set** that documents your solution. (Presentation 15min, 5min Q&A).

Remarks

- Please prepare a live demo that you can show during your presentation and please include some screenshots of the demo in your slides (as additional slides and backup).
- Make sure that your code is executable (e.g. use relative paths to the data files, specify pip and import statements)

Assessment Criteria:

- **Completeness of analytical requirements:** how well have you understood the needs of stakeholders and how well have you translated them into a specification of analytical processing?
- **Implementation:** justification of your technical solutions (illustration of the way it satisfies the stakeholder needs as well as what could be improved).
- **Complexity:** diversity of aspects included, scope of the scenario, number of technologies or methods, decomposition of tasks
- **Innovation potential:** enhancing communication, automation potential, analytical insights, improved understanding, innovative human machine interactions.
- **Presentation:** how well have you justified and visualized your case and your design choices?

Be creative, work together and decompose the project into several tasks, which you then can combine and present as a team.