1. **Group number: 31**

2. **Names of students in group?**
   Nikolaj Fu Jie Lin, Marie Skovbye, Mathias Hedegaard Udsen, Elena Gritskevich

3. **What is your research question?**
   Complexity categorization of articles from public newspaper.
   Build a model to predict/categorize text complexity of the public newspaper article.

4. **What kind of data are you planning on using? How will you get access to these data?**
   Web-scraping of public newspapers (e.g. TV2/ DR/ Ekstrabladet). Publicly available open resources.

5. **What will your data analysis be like? Will you use machine learning? How?**
   We scrap articles of different categories (politics, sport, culture, etc).
   Each article is assessed for complexity (target parameter 'y' to be predicted) using LIX number[1]:

   $$\mathrm{LIX} = \frac{A}{B} + \frac{C \cdot 100}{A},$$, where

   A is the number of words,
   B is the number of periods (defined by period, colon or capital first letter), and
   C is the number of long words (more than 6 letters).
   Scores usually range from 20 ("very easy") to 60 ("very difficult")

   Article parameters are used as 'x' variables (e.g. author characteristics, time of publication, category of the article, etc) to predict complexity. In terms of machine learning, we plan to use the elastic net model because we expect high multicollinearity between our variables. For example, we can imagine that the same author often writes in the politics section, because the author is a politics expert.

   Preliminary data analysis of all collected data will be made to analyze correlation of variables.

6. **Have you already identified other papers within this area that you can use in a literature review? If so, name a few and explain what they do in one sentence only.**
   - **Collins-Thompson, K.**
     (2014) Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2), 97–135. 10.1075/itl.165.2.01col https://doi.org/10.1075/itl.165.2.01col [Google Scholar]
   - **Marina Santini, Arne Jönsson**
     (2020) Pinning down text complexity. An Exploratory Study on the Registers of the Stockholm-Umeå Corpus (SUC). *Register Studies,* Volume 2, Issue 2, Dec 2020, p. 306 – 349
     https://doi.org/10.1075/rs.19005.san

   Above articles made research on various aspects of text readability including LIX number parameter, included in our project. However, given our project timeframes, we are not aiming to challenge or compare our approach with the ones used in the above mentioned articles, instead observe the work in the text readability assessment area.

7. **How do you 'contribute' to the literature?**
   The main focus of the research project is to analyze if article characteristics can provide a robust estimates of article text complexity. Additional goal is to demonstrate practical knowledge of Python tools gained at the course 'Introduction to Social Data Science – summer school 2023'

[1] https://en.wikipedia.org/wiki/Lix_(readability_test)