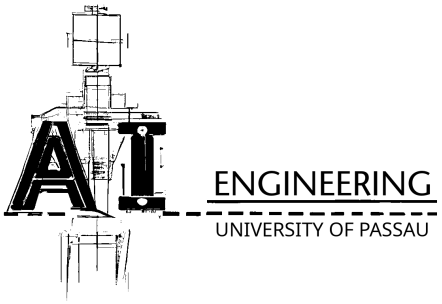
Principles of AI Engineering

Exercise 2

Prof. Dr. Steffen Herbold



Author: Lukas Schulte

Project task

Your coworker provided a first dataset, which you can use to train your AI model (`sample1.csv.gz` in StudIP). Load the data into a Jupyter Notebook, inspect it and implement appropriate text pre-processing steps. Collect at least one example which shows what each of your pre-processing steps does.

Hint: Next week's exercise will deal with creating the model. The model (a random forest model) will be built with sklearn[1].

Questions

1. Analyze the dataset.
   - Which fields are relevant for the prediction in the given project scenario?
   - Is the dataset even suitable for the task at hand? Do you spot potential problems?
2. Define the following text pre-processing steps:
   - Tokenization
   - Normalization
   - Noise removal
   - Stemming
   - Lemmatization
   - Stop-word removal
3. Explain concept drift and how it would impact your application. Describe how you would mitigate the effects.
4. Describe what could be considered a feedback loop in your application and why that can be a problem. Explain how you can measure the effect.

[1] `https://scikit-learn.org`