LabTwin Data Science Challenge

Thank you for taking the time to take the LabTwin Data Science challenge! Below are two questions, please answer both of them. For both questions, clear information is provided on the expected format of the answer.

Part 1: Data Normalisation

At LabTwin, we are working on a speech-to-text model for life-science specific vocabulary. In order to train our transcription model, we are collecting life science content from various sources, for example from the US National Library of Medicine National Institutes of Health. Attached you will find the *Methods and Materials* part of 363 papers dealing with genetics. A relevant step in preparing this content for subsequent model training is data normalisation.

Please write a script in a programming language of your choice that normalises the provided papers in a uniform way, and aggregates them into a single document; please add comments to your script, explain your procedures and provides clear instructions on how to run the script. What other techniques would you apply to use these papers in the best way to train a transcription model?

Minimum requirements:

- plain text using only the ASCII printable character set
- lower-case text
- removing all punctuation
    - expand numbers to spoken form (333 -> *three hundred and thirty three, 0.1 -> zero point one*)

remove special character and symbols, mathematical and chemical formulae can be ignored

Part 2: Model Performance Measurement

The Data Science team at LabTwin has to constantly monitor performance of the speech-to-text transcription model to ensure good transcription accuracy and therefore a good user experience. Selecting and preparing domain relevant content is highly important for the success of our transcription model. We are adding constantly new content to our model to cover a broad range of domain specific vocabulary and expressions.

How exactly would you keep track of the overall performance of the transcription model over time? Which KPIs are relevant and what kind of (acoustic) test data would you generate?

Additionally, how could user generated content be integrated into the model in order to improve the transcription accuracy? What sources of bias can occur and how would you deal with it? Please provide your answer in no more than 300 words.