Time

- 7 days to develop
- Recommended time for all tasks: up to 8 hours

Results delivery

- Link to code repository with the solution
- Classification results description
- Short Readme with instructions on how to obtain preprocessed dataset, train the model and deploy/test the API.
- We will assess whole workflow, not only the performance of the model

Dataset

- Based on ‘IMDB dataset’ (attached in zip file)

Task

- Data preprocessing
    - Preprocess provided dataset so that it can be used for modeling. Suggested steps:
        - Remove non - digit and non - letter parts.
        - Remove stop words (but keep the word 'from').
        - Remove standalone numbers (e.g., remove '100' from '100 pieces').
        - Any other text cleaning and standard processing.
- Classification
    - Using the preprocessed data, build a classifier model. Target variable is ‘sentiment’.
- Deployment
    - Build a sample Rest API in Python for your model.
    - Create a Dockerfile that can be used to build an image that will serve your model and API.
        - Provide script / command that can be used to build the image and run the container.
        - Add script / command that can be used to test the API

- Notes
    - Data preprocessing and Classification steps should be delivered in reproducible form (.py files / Jupyter notebooks) so that we can recreate the steps and build the model.
    - Decision on whether to train the model inside of a container or not is up to you. We just need a way to recreate it on our end.
    - Please follow PEP8 and general programming best practices.
    - Choice of Python packages and algorithms is up to you.