SEC 101: Data and Information Fundamentals Assessment Instruction

General Information

Competency Code: SEC-101
Competency Title: Data and Information Fundamentals
Competency Semester: Fall 2025
Instructor Name: Charnon Pattiyanon, Ph.D. (charnon@cmkl.ac.th)

Assessment Overview

Data is the most valuable asset in modern information systems. It plays a critical role in supporting business operations, ranging from user facilitation to management decision-making. In many cases, data can be sensitive and may expose personal matters. As future engineers in AI and Computer Engineering field, students must have a comprehensive understanding of data and information from every perspective and know the way to handle data and information properly.

In this assessment, students will have the opportunity to apply their knowledge and demonstrate their skills by analyzing and preprocessing existing datasets to fulfill their needs to develop a simple AI system.

Assessing Skills

- **[SEC-101:00010] Evaluate the relation of data acquisition, preparation, transformation and cleansing** - Successful students must understand the types, characteristics, and properties of data and information, and are able to evaluate the relation between data acquisition, preparation, transformation, and cleansing comprehensively.
- **[SEC-101:00020] Design a proper process of data acquisition, preparation, transformation, and cleansing** - Successful students must be able to design a proper process of data acquisition, preparation, transformation, and cleansing by integrating techniques, tools, and guidelines we have learned in the lecture.

Pre-cautions

- Please express your answers to each required outcome based on your own ideas and perspective. Plagiarism is unacceptable. If I find that either content or ideas are remarkably similar between two or more students without any sound reasons, scores of all students will be deducted as a consequence of illegal actions.
- Students are expected to demonstrate a deep understanding of the subject matter through critical analysis and original insights. Overreliance on AI-generated content without providing substantial original thought will negatively impact the assessment score.
- Justifications are critical explanations of why your answers or outcomes are expressed the way they are. They should be written in a "why" style. For example, "I believe that this privacy principle is possessed by the target system because . . . ." There will not be a one-size-fits-all solution or criticism for writing a justification. Your analysis skills will be evaluated through the clarity of your justifications.
- I encourage students to ask me any questions to address their curiosity about the assessment via email or the discussion page on Canvas. However, please do not submit your assessment report and ask for my feedback on it. I will consider it as a report submission.
- This assessment will provide some optional outcomes. They are not required to be included in the report. They will not affect the score of this assessment. However, they will be considered in cases where you receive a borderline score between two mastery levels or a low score on this assessment. The optional outcomes will not exceed 10% of the overall score, depending on the instructor's discretion.

Assessment Instruction

The total score for this assessment is **200 points**, with each assessed skill contributing 100 points. Please carefully follow the instruction below:

1. Search for an online open dataset that interests you. Two good sources are HuggingFace.co and Kaggle, though you are not limited to these platforms. **It is expected to select a dataset that has not been preprocessed yet**.
2. Design a data cleaning and preprocessing workflow to produce a clean and structured dataset suitable for AI model training.
3. Propose the development of an AI model that uses the selected dataset to train and deliver specific features or functionalities.
4. Submit a PDF report that follows the template provided below.

SEC-101: Data and Information Fundamentals (Assessment Report)

Name: [Replace with your name]
Nickname: [Replace with your nickname]
Email: [Replace with your email address]

Dataset Name: [Replace with the name of your selected dataset]
Dataset Link: [Replace with the link to your dataset]

Dataset Description: (*5 Points*)

Explain your selected dataset from various aspects. You must include at least the following points:

- How was the data collected or acquired by you or the dataset owner? Was it collected in a proper and unbiased way?
- How many records are included in the dataset? Is the dataset large enough to train an AI model?
- What was the purpose of collecting this dataset? How can this dataset be used?

Example of 10 Data Records: (*10 Points*)

Display at least 10 sample records from the selected dataset in a **table** format, and highlight all data quality issues in the dataset.

Dataset Analysis: (*45 Points*)

Write at least one paragraph to analyze the selected dataset. You must address the following questions:

- How many attributes are included in the selected dataset?
- Is this dataset clean and complete enough for AI model training?
- If not, what data quality issues did you identify in the dataset? Please list all of them in this section.
- If you believe it is clean and complete, provide a clear rationale to justify your decision. (Note: your score will be deducted if any discrepancies are found in the dataset that you failed to mention.)

Data Attribute Analysis: (40 Points)

Fill in the following table similar to what we did in Lab 1.

Attribute	Example	Type of Data	Representation	Sensitivity	Justification
Name	John Doe	QT-NM	Text	Low	Name is classified as QT-NM because ...
...	...	...	...	...	...

For the "Type of Data", "Representation", and "Sensitivity" columns, select your answers from the options below:

- Type of Data:
  - **QN-CT** - Quantitative continuous data
  - **QN-DC** - Quantitative discrete data
  - **QL-NM** - Qualitative nominal data
  - **QL-OD** - Qualitative ordinal data
- Representation:
  - **Text** - Data represented as a sequence of letters or symbols
  - **Number** - Data represented as numeric values
  - **Image** - Data represented as an image or a matrix of color pixels
  - **Video** - Data represented as a sequence of image frames and audio tracks
- Sensitivity:
  - **High** - Highly privileged data that must not be exposed to anyone.
  - **Moderate** - Data that should only be accessible to a permissioned group of individuals
  - **Low** - Non-privileged or public data

Data Acquisition, Cleaning, and Preprocessing Workflow: (60 Points)

Write a section explaining how you plan and design your data acquisition, data cleaning, and data preprocessing tasks. It must be described in a step-by-step manner and is expected to be simple enough for readers who know nothing about can understand easily. You should mention or refer to topics we discussed during the lecture in order to support your explanation. For example:

1. Data Acquisition: We aim to collect a dataset for training our AI model for classifying ....
   (a) We browsed in online dataset repositories from Huggingface.co and Kaggle, which are well-known and widely-used in research and practical communities. These repositories allow us to search with keywords and it will display datasets along with their description, data examples, and statistics.
   (b) We select three data candidates based on the purposes for classifying ....
   (c) ...

Example of 10 Data Records from the Preprocessed Dataset: (20 Points)

Display the same 10 sample records from the selected dataset in a **table** format, and highlight all the resolved or cleaned data records that are ready to use in the AI model training.

AI Model Development Explanation: (10 Points)

Write at least one paragraph to explain the AI model that you plan to train using the processed dataset. Your explanation must address the following questions:

- What is the purpose of the AI system? What are the expected outcomes?
- Which specific ML algorithms or deep learning approaches will you use to train the model on this dataset? Why do you believe these models are the most suitable for the processed dataset?
- What are the input and output formats of the model?

Model Development Details: (10 Points)

Provide screenshots of your model development source code and the outcome you achieved. Write at least one paragraph to explain your source code and the outcome.