Introduction: In this project, we aim to train a model to accurately classify question IDs, questions, answer IDs, and answers from a dataset containing both portions in Excel and PDF formats. We utilize the spaCy library, a popular open-source library for natural language processing (NLP) in Python, to build and train our model. Problem Statement: The dataset consists of two portions: one containing exam questions and the other containing their corresponding answers. Our task is to train a model on this dataset to accurately classify question IDs, questions, answer IDs, and answers. Model Architecture: The proposed model architecture utilizes spaCy, a Python library for NLP. We leverage spaCy's capabilities for tokenization, entity recognition, and training a custom Named Entity Recognition (NER) model. The model architecture involves preprocessing the text, creating training examples, initializing and training the NER model, and testing the trained model. Key Processes: Data Loading: The code begins by loading the dataset from Excel files containing question and answer data. Preprocessing: Text preprocessing involves stripping whitespace and any other necessary preprocessing steps to clean the text data. Training Data Creation: The code creates training examples by pairing questions with their corresponding answers and annotating them with the question ID, question text, answer ID, and answer text. Named Entity Recognition (NER): We initialize a NER model in spaCy and add a custom label for identifying questions in the text. Training: The NER model is trained using the created training examples. We utilize spaCy's training capabilities to update the model weights and optimize it for entity recognition. Model Testing: Finally, the trained model is tested by providing a sample question text, and the extracted entities (question) along with the associated answer ID and answer text are printed.