MAIS 202 - PROJECT DELIVERABLE 1

Choice of dataset

Stanford Question Answering Dataset (SQuAD)
https://rajpurkar.github.io/SQuAD-explorer/

We will use the Stanford Question Answering Dataset (SQuAD) for our machine learning project because it is a large-scale, diverse dataset containing over 100,000 questions and answers. It has been widely used and evaluated by the research community and is well-suited for training and evaluating models for question answering and machine reading comprehension tasks.

Project Goal

Question Answering Model: Building a supervised learning logistic regression model that can answer questions based on the information contained within SQuAD. The model could be trained on the questions and answers in the dataset, and then be used to answer new questions.

Methodology

Data Preprocessing

The data will undergo several preprocessing steps to ensure that it is suitable for the question-answering model. These steps include data cleaning, data transformation, and data encoding.

Data Cleaning: We will remove any duplicate data and missing or incorrect data. The data will also be formatted and normalized to ensure consistency. Typos and grammar mistakes will also be corrected to ensure that the model understands the questions' content. We will use imputers to fill in the gaps if any data needs to be added.

Data Transformation: The data will be transformed before being processed by the model. Continuous features will be discretized into ten equally sized buckets to make them easier to process. The features will then be grouped into 8 valuable categories: matching word frequencies, bigram frequencies, root match, lengths, span word frequencies, constituent labels, span POS tags, and lexicalized dependency tree paths. The last two categories are considered to be the most important.

Data Encoding: The data will be encoded during preprocessing to ensure that the model can easily process it.

Machine Learning Model

The machine learning model we are planning to use is Logistic Regression. This model was deemed by the creators of the SQuAD dataset to be the most efficient for classifying the type of question being asked and formulating an appropriate answer.

We have also considered other classification models, including KNN, Naive Bayes, Support Vector Machines (with various kernels), and Random Forest. However, after weighing the pros and cons of each model, we ultimately decided to use Logistic Regression for its efficiency in this specific task. However, the option to switch to another model is still open, if deemed necessary.

Evaluation Metric

When it comes to the Evaluation Metric we intend to use, since we’re using a classification model, it only makes sense to use a Confusion Matrix. However, we still have to learn how the BLEU score with brevity penalty could help us as it deals with text generation problems which is also what we work on.

Application

We hope to build a web application and provide a user-friendly interface that allows users to input their questions either through voice or text. This will allow for greater accessibility and convenience for users with different preferences.

The model will then provide its answer via text, which will then be voiced by an API. This will ensure that the user can receive the answer in their preferred format, whether they prefer to hear the answer or read it. The dual output format will also ensure that the bot's answer can be easily shared or recorded, making it more accessible for others to use.