

IDC410: Machine Learning & AI Programming Exercise-4

(Question identification using NLP)

An EdTech company has hired you as a machine learning engineer. The first task they provide you is to build a model that can identify questions within transcripts of video lectures. You are given the attached data set which has been painstakingly annotated, with each sentence being classified as a question or a sentence. During your NLP course at university you had learned various methods for vectorization of text and numerous classification algorithms in your Machine Learning course. You had also learned about methods for evaluating models and selecting the best model.

Using all these skills, build a pipeline that does various cleaning of the text sentences, use all vectorization approaches you have been introduced to (both semantic and syntactic), build models for discriminating between questions and general sentences and evaluate the models using ROC curves.

Some highlights:

1. Use tfidf, LSA, LDA and one word embedding and one sentence embedding method to vectorize the texts.
2. Use 4 modelling techniques from the ones that have been taught to you for building models.
3. Evaluate using area under the ROC curve.
4. Create a Restful API.
5. Dockerize it.