On this page

Build from scratch

Multiple different strategies can be used to build a machine learning model from scratch:

- "Bag-of-words" is one simple strategy. Refer to this Google CodeLab that walks you thorugh the process of building a simple model from scratch.
- A model based on the TFIDF Vectorizer strategy was contributed by the community into the project by Emmanuel Djaba. This is new base model would be our second approach to home growing Kindly's own model using traditional machine learning until such a time where the strategy would have to be changed due to new information.

Deep Learning vs Machine Learning

The Cardiff model from Huggingface uses the deep learning approach to identifying whereas the initial model created uses traditional ML. In simple terms, with deep learning, a lot of classified data is thrown at a neural network and it would figure out what is should look out for in predicting on its own. With traditional ML, a number of algorithmic steps are taken to try to point out features and what should be looked out for in the data, and prediction would be based off that.

Deep Learning involves using these packages or a combination of them to build "layers" of your model. A layer is a single filter or transformation of the full dataset you pass through the model. Sometimes a layer strips away some parts of the data, some other times too it adds to it or classifies and records it. This is the actual learning process. Multiple layers are created and combined to transform the input data such that the output of the model is your desired result or at least close to it.

Model Building

The general flow or process in building a model include. These can also be known as traditional ML steps:

1. Split
    i. This involves importing your full dataset,

   ii. Dividing it into two sets; one for testing/validating and the other for training. The validation set is usually smaller than the training set.
   iii. This split ratio can be 20% is to 80% depending on the total size of your data and can be adjusted as necessary.
   iv. Some also further break down the training dataset into smaller chunks because training requires a lot of computing resources and can take a long time.
   v. For Kindly, the initial dataset is very small so all of it can be used at once however, as more data is gathered over time, this would also have to be adjusted accordingly in the model.
2. Model
   i. This is the actual building of the layers and transforming and tokenizing the data.
   ii. Tokenization is a process where classified dataset entries are converted into a format that a neural network in the model can understand and work on.
   iii. The model can then be saved into a single binary file and be called or used by Pytorch or Tensorflow (depending on the strategy you're using) in the API endpoints for Kindly.
3. Train
   i. This is where the python package's training strategy is employed to "fit" the dataset into the model. Different packages call this function different things such as '.fit()' or '.train()'
4. Test / Validate

Jupyter Notebook

The main tool used to build models is Jupyter Notebook which is based on the Python programming language. The file extension of a Jupyter Notebook is ipynb. It enables you to use the Machine Learning packages namely Keras, Pytorch, Scikitlearn and Tensorflow.

Edit this page