Traditional ML Model for URL Classification

Documentation: Retraining & Using the Model

Dataset Format

- CSV file with the following format:
- • url: The full URL string
- • label: Target label (e.g., product, not_product)

Preprocessing Steps

- Steps before training the model:
- • Clean URLs (remove stop words, lowercase, etc.)
- • Tokenize or vectorize using TF-IDF or CountVectorizer
- • Encode labels to integers (LabelEncoder)

Model Training Code

- Steps to retrain the model:
- • Load the dataset
- • Split into train/test
- • Train using LogisticRegression / RandomForest / SVM
- • Evaluate using accuracy, F1, precision, recall

Saving and Loading the Model

- After training:
- • Save model: joblib.dump(model, 'model.pkl')
- • Save vectorizer: joblib.dump(vectorizer, 'vec.pkl')
- • Load later: joblib.load('model.pkl')

Batch Prediction Code

- To use model on a new CSV:
- • Load vectorizer and model
- • Preprocess the new URLs
- • Transform URLs using vectorizer
- • Use model.predict to get labels

Notes & Tips

- • Use GridSearchCV for tuning hyperparameters
- • Always save the label encoder too
- • Keep preprocessing steps consistent