// Documentation

Amazon ML Challenge

2023

// Problem

To build a machine learning model that can predict product length from catalog metadata.

// Dataset

Column Name	Description
PRODUCT_ID	Represents a unique identification of a product
TITLE	Represents the title of the product
DESCRIPTION	Represents the description of the product
BULLET_POINTS	Represents the bullet points about the product
PRODUCT_TYPE_ID	Represents the product type
PRODUCT_LENGTH	Represents the length of the product

Download ›

// Team

ML Architects

- Pranjal Agarwal
- Akanksha Mishra
- Prajesh Pratap Singh

// Approach

- Imported Training and Testing dataset using `pandas dataframe`
- Applied data pre-processing
    - Removed `Nan` values from columns ['BULLET_POINTS', 'DESCRIPTION'] of training and testing data
    - Merged columns ['TITLE', 'BULLET_POINTS', 'DESCRIPTION'] into one column ['INFO'] in training and testing data
    - Data cleaning: cleaned training and testing data of ['INFO'] column
        - Replaced newline with space
        - Replaced tab with space
        - Replaced quotes with space
        - Converted to lowercase
        - Removed punctuation marks
        - Removed apostrophe s
        - Re-structured data
            - Training data
                - ['Title', 'PRODUCT_LENGTH']
            - Testing Data
                - ['Title', 'PRODUCT_ID']
        - Separated data into training and testing set
        - Performed vectorization
            - Imported `TF-IDF` (Term Frequency - Inverse Document Frequency) vectorizer
            - Removed stopwords using `NLTK`
            - Transformed dataset and fit into vectorizer model
        - Performed data encoding to feed into `Logistic Regression` model
- Training Model
    - Providing `x_train` and `Y_train` as input
    - Predicting using Model

- Writing data into File
    - Initializing predicted [`'PRODUCT_ID', 'PRODUCT_LENGTH'`] to new dataframe variable
    - Writing this variable to the file `` `submission.csv` `` and saving it.

// Feature Engineering

- Extracted all the columns necessary for prediction
- Merged these columns into one column
- Removed
    - Whitespaces
    - Newline
    - Punctuation
    - Apostrophe
    - Stopwords
- Converted to lowercase
- Performed `` `TF-IDF` `` vectorization

// Tools

- **Sklearn**: Scikit-learn or Sklearn is a popular Python library for machine learning. It provides a wide range of tools and algorithms for various machine learning tasks, such as classification, regression, clustering, and dimensionality reduction.
- **Pandas**: Pandas is a popular open-source Python library used for data manipulation, analysis, and cleaning. It provides data structures for effectively storing and manipulating large and complex datasets.
- **Numpy**: NumPy (short for Numerical Python) is a Python library used for working with arrays and matrices of numerical data. NumPy provides a set of numerical and mathematical functions for fast operations on these arrays.
- **Csv**: CSV stands for Comma Separated Values. It is a simple file format used to store tabular data, such as a spreadsheet or a database.

- **Nltk**: NLTK (Natural Language Toolkit) is a popular Python library used for natural language processing (NLP) tasks such as tokenization, stemming, lemmatization, part-of-speech tagging, parsing, and machine learning. It provides a set of tools and resources for working with human language data in Python.
- **Re**: re is a built-in Python module used for working with regular expressions. Regular expressions are a sequence of characters that define a search pattern, used for matching and manipulating strings.
- **TF-IDF**: TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical measure used to evaluate the importance of a word in a document in a collection of documents or corpus.