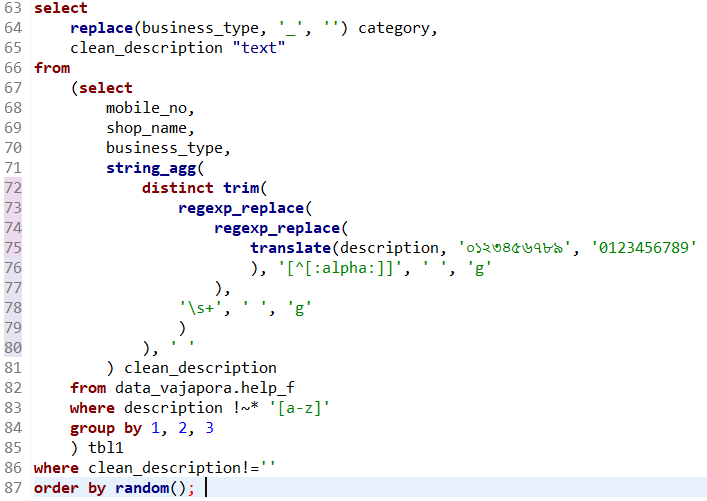
Prediction of Business Type from Journal Descriptions & Supplier Names Using Word2Vec Model

**Author (s):** Shithi Maitra, Atanu Das Bapon

**Supervisor:** Md. Nazrul Islam

**A.** Implementation Using Journal Descriptions

**Step-01:** Preparation of the Dataset



**a.** First, we replace all Bangla digits by their English counterparts.

**b.** Second, we ensure the strings contain all Bangla words (in version-01) and no punctuation marks.

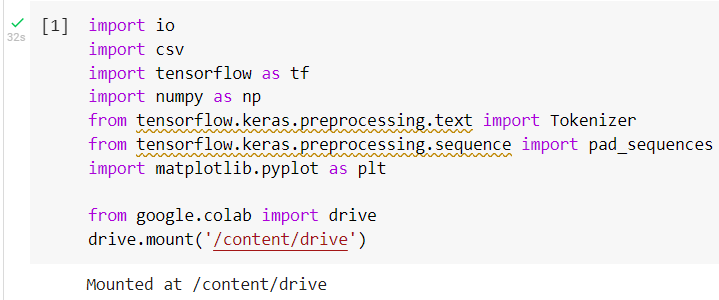
**c.** Third, we remove unnecessary whitespaces from the training strings.

**d.** Fourth, we aggregate the last 7 day's available descriptions against available merchants.

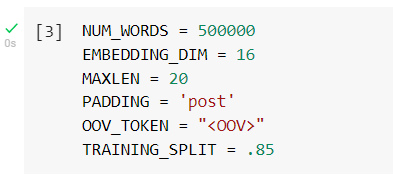
**e.** Finally, we randomly shuffle the dataset for a representative distribution in each sample.

**Step-02:** Fitting the Word2Vec Model

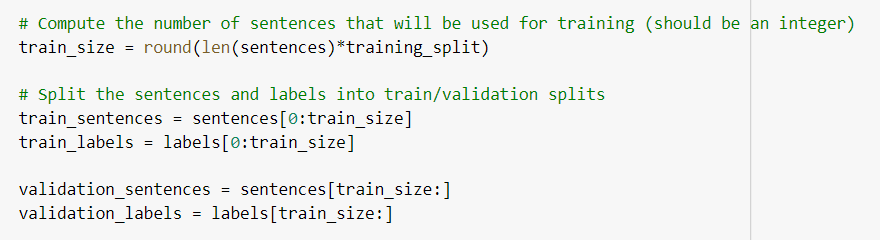
**a.** Use IDE: Google Colab and mount to Google Drive, for fetching training data



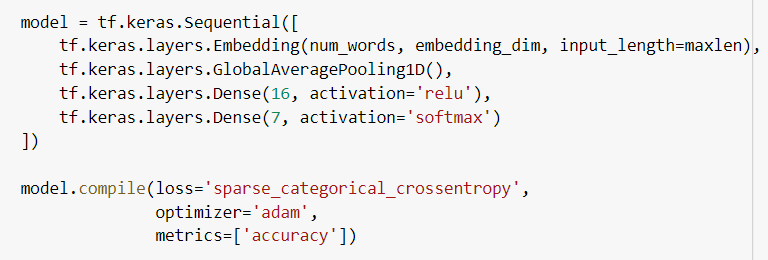
**b.** Set hyperparameters for Word2Vec model



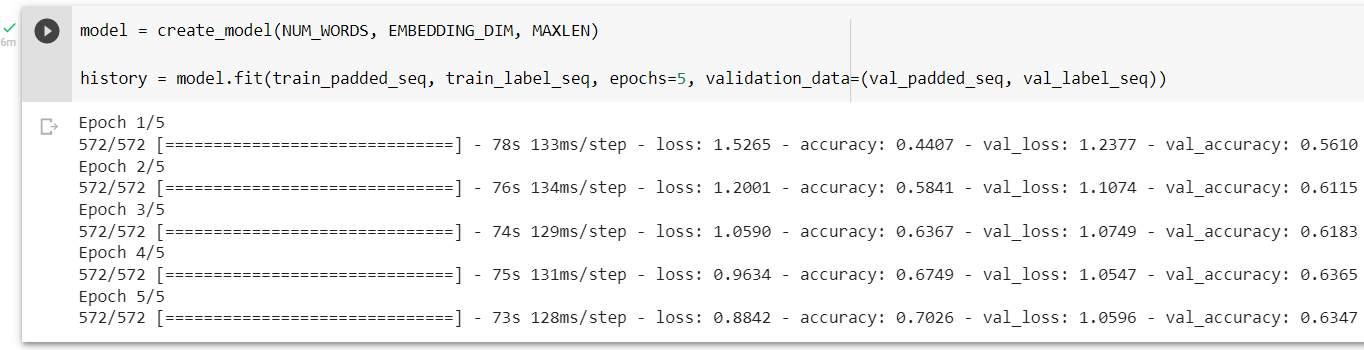
**c.** Split full dataset into training and dev sets



**d.** Define the model

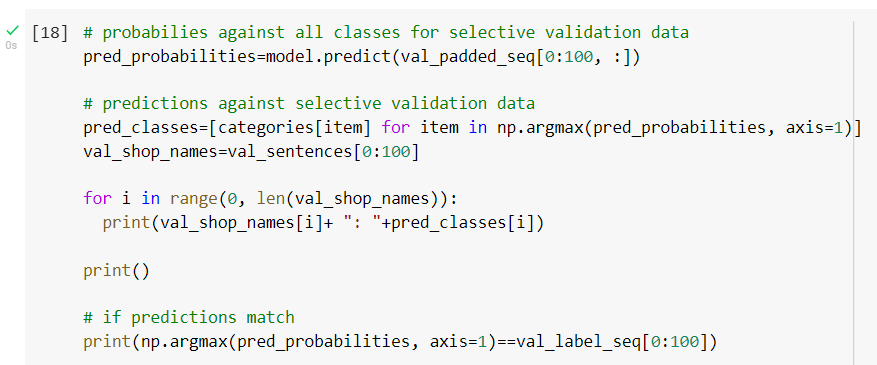


**e.** Train the model

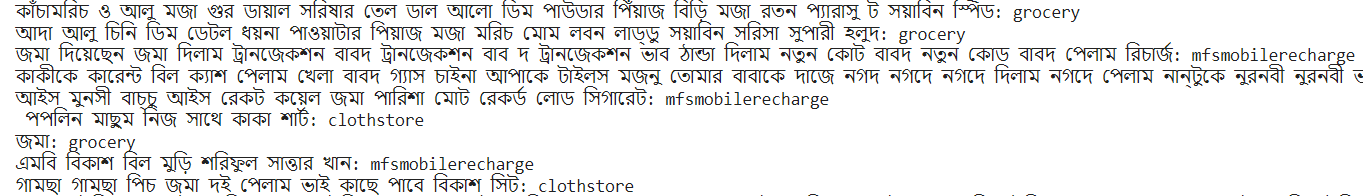


**Step-03**: Generate/Inspect Predictions

**a.** Generate output as per the highest probable class

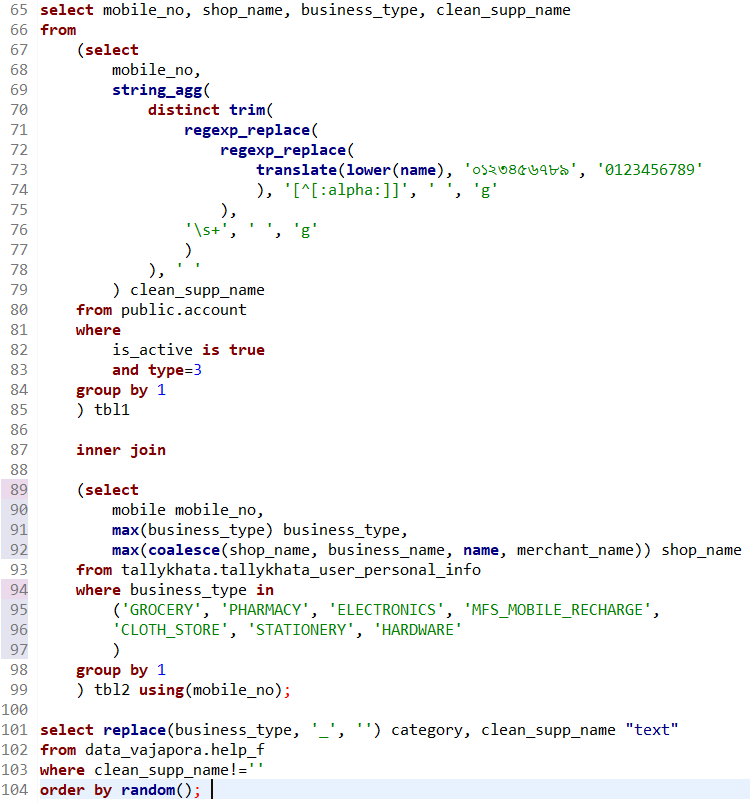


**b.** Inspect outputs



**B.** Implementation Using Supplier Names

**Step-01:** Preparation of the Dataset



All steps are the same as before, except that:

**a.** We consider both Bangla and English names.

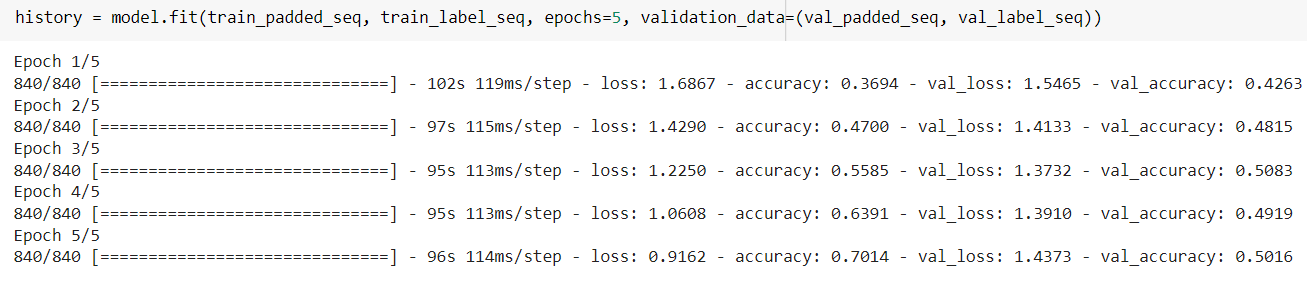
**b.** For English names, we convert all the letters to lowercase for the sake of cleanliness of data.

**c.** Additionally, we classify on 7 classes and remove separators (e.g., \_ ) from class names.

**Step-02:** Fitting the Word2Vec Model

Steps - **a, b, c, d** are as before.

**e.** Train the Model



**Step-03:** Generate/Inspect Predictions

**a.** Generate output as per the highest probable class

**b.** Inspect outputs

