Report

Chaiyasait Prachaseree

March 16, 2023

1 How to run EDA and Sentiment Classifer Training

1.1 Environment Creation

Assuming Anaconda or Miniconda has been downloaded and installed :
Create and activate conda environment named *aisg_task2*:

```
$ conda create -n aisg_task2 python=3.9
$ conda activate aisg_task2
```

Install dependencies:

```
$ pip install scikit-learn pandas numpy pyvi jupyter
```

1.2 Running the script

In root directory and after activating the conda environment, run :
Run jupyter notebook by :

```
$ jupyter notebook scripts/EDA_vietnamesedataset.ipynb
```

Train SVM classifier by :

```
$ python scripts/task_2.py
```

Train Naive Bayes classifier by :

```
$ python scripts/task_2.py nb
```

1.3 EDA

There are a total of 5914 reviews, 4 of which have null labels. These 4 labels are dropped for the following analysis.

Overall, each review has an average of 79.04 words. There are 467147 total words, with only 10978 unique words in the dataset.

There are more positive (47.0 %) and negative (40.9 %) reviews than neutral ones (12.1 %). The average rating for negative, neutral, and positive is 3.3, 4.8, and 9.2 respectively. Negative sentiment reviews have more words on average, while neutral reviews have the least.

The jupyter notebook also shows top 20 frequent words overall and in each sentiment.

1.4 Sentiment Classifier Training

The selected model used for this task is a Support Vector Machine (SVM) with TF-IDF as features. TF-IDF gives the importance of a word relative to the frequency it appears between a single review and the overall word frequency. A SVM classifies by finding a set of hyperplanes that seperates vectors into their classes. Thus, new samples will lie within some plane that is mapped to a single label. SVMs performs slightly better than Naive Bayes for this specific task in some runs, but more experiments are needed.

The code splits the given Vietnamese reviews dataset by 85/15 training randomly and trains the classifier on the training split. Then, it shows the accuracy, confusion matrix, and the precision-recall score on the validation set.

1.5 Flask App

This Flask App uses a SVM trained fully on the given dataset to test on Vietnamese reviews. With more time, more exceptions such as non-Vietnamese languages can be added.

1.5.1 How to Run

From root directory :

```
$ cd scripts/Flask-Docker-App
$ sudo docker build --tag python-docker .
$ sudo docker run -d -p 5000:5000 python-docker
```

Go to localhost:5000 on any web browser