COMP4650 Assignment 2 Answers

Jieli Zheng

u6579712

Question 1 A simple linear classifier

Generally I make three major changes on the classic Logistic Regression model.
The first major change I make in preprocessor is using SnowballStemmer with language=english. SnowballStemmer is a nltk package built-in stemmer that can transform any english words back to their stems which is easier for tokenizer to find the right tokens from the raw text.
The seconde major change I make is using tokenizer to WordPunctTokenizer instead of WhitespaceTokenizer. The WordPunctTokenizer can split sentence into a bunch of words and single punctunations. It is stronger than the original setting with WhitespaceTokenizer because WhitespaceTokenizer can't split punctunations from word. It definitely has different tokens like "word," "word." which should be the same meaning in human's perspective.
Last major change I make in CountVectorizer is setting the parameter max features to 5000. The advantage of extracting more features is that model can have larger observation on low frequency words with explicit tendency.

Question 2 Embedding based classifier

Code in Pycharm

Question 3 Tuning a pytorch model