Matters needing attention

1. The corpus has been **lowercased white-space tokenized**. Therefore, you don't need to invoke an external tokenizer.

2. Please maintain a vocabulary with a size of **33,278**. We have preprocessed a vocabulary file (https://nlp.cs.hku.hk/comp3361/vocab.txt) for your convenience.

3. The provided corpus is the only dataset source for your model. You cannot use pre-trained parameters to initialize your model.

4. You are allowed to use packages of existing models, such as the logistic regression classifier of scikit-learn.