coursera

Lab Help

Out of vocabulary words (OOV)

Vocabulary

In the video about the out of vocabulary words, you saw that the first step in dealing with the unknown words is to decide which words belong to the vocabulary.

In the code assignment, you will try the method based on minimum frequency - all words appearing in the training set with frequency >= minimum frequency are added to the vocabulary.

Here is a code for the other method, where the target size of the vocabulary is known in advance and the vocabulary is filled with words based on their frequency in the training set.

In [1]:
```
# build the vocabulary from M most frequent words
# use Counter object from the collections library to find M most common
words
from collections import Counter

# the target size of the vocabulary
M = 3

# pre-calculated word counts
# Counter could be used to build this dictionary from the source corpus
word_counts = {'happy': 5, 'because': 3, 'i': 2, 'am': 2, 'learning':
3, '.': 1}

vocabulary = Counter(word_counts).most_common(M)

# remove the frequencies and leave just the words
vocabulary = [w[0] for w in vocabulary]

print(f"the new vocabulary containing {M} most frequent words: {vocabul
ary}\n")
```

```
the new vocabulary containing 3 most frequent words: ['happy', 'becau
se', 'learning']
```

Now that the vocabulary is ready, you can use it to replace the OOV words with $< UNK >$ as you saw in the lecture.

In [2]:
```
# test if words in the input sentences are in the vocabulary, if OOV, p
rint <UNK>
sentence = ['am', 'i', 'learning']
output_sentence = []
print(f"input sentence: {sentence}")
```