# adding context dependent stopwords to independent stopwords

c\_keys = colloc\_dict.keys()

for unit\_code in c\_keys:

for word in colloc\_dict[unit\_code]:

word\_count = 0

for unit\_code2 in c\_keys:

if word in set(colloc\_dict[unit\_code2]):

word\_count += 1

if (word\_count / len(colloc\_dict) < 0.05) or (word\_count / len(colloc\_dict) > 0.95):

stopwords.add(word)

print(len(stopwords))

# Total stopwords: 4123

# stemming

stemmed\_dict = {}

stemmer = PorterStemmer()

final\_tokens = []

# each word is stemmed in colloc\_dict and added to stemmed\_dict

for k,v in colloc\_dict.items():

stemmed\_dict[k] = [stemmer.stem(each) for each in v]

# vocab is updated

vocab = set(list(chain.from\_iterable(stemmed\_dict.values())))

len(vocab)

# 225