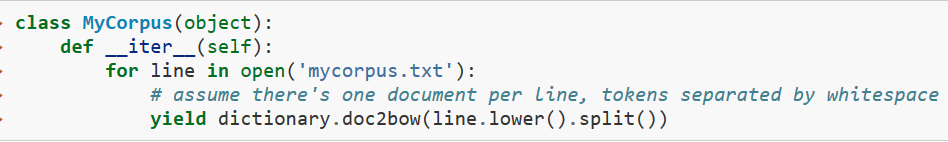
1. Get ‘documents’
2. Tokenize documents, remove common words, words that only appear once
   1. Self-determine common words, e.g ‘the’
   2. Take array of not common words and use a dictionary to determine their frequency (word: frequency as key: value)
   3. Save array of words that appear more than once (in all documents)
   4. Genism example splits by whitespace and lowercases everything as done in Deerwester et all’s article (simple but not the most efficient) - <https://radimrehurek.com/gensim/tut1.html#id3>
3. Use bag-of-words approach to convert documents to vectors
   1. Genism library contains a special dictionary method to create dictionary of token and frequency from list of lists, and another method to save
      1. >>> dictionary.save('/tmp/deerwester.dict') # store the dictionary, for future reference
   2. Also to save to disk:
      1. Corpora.mmCorpus.serilize(‘/tmp/deewester.mm’, corpus) where corpus = [dictionary.doc2bow(text) for text in texts]
         1. Tests = every list of not common words with repeats per document (list of lists)
         2. Test is the list of filtered words in that document

If memory needed for docments > RAM, have to save to disk



* There is one document per line in ‘mycorpus.txt’ (separated by newline)
  + Could change \_iter\_ to parse XML or whatever