Word counts with bag-of-words

Puteaux, Fall/Winter 2020-2021

```
##############################################
##                                          ##
##  Natural Language Processing in Python   ##
##                                          ##
##############################################
```

§1 Introduction to Natural Language Processing in Python

§1.2 Simple topic identification

1 Word counts with bag-of-words

1.1 What is bag-of-words?

- It is a basic method for finding topics in a text.
- Need first to create tokens using tokenization.
- And then count up all the tokens.
- The more frequent a word, the more important it might be.
- It can be a great way to determine the significant words in a text.

1.2 Code of bag-of-words in Python:

[1]:
```
from nltk.tokenize import word_tokenize
from collections import Counter

Counter(
    word_tokenize("""The cat is in the box. The cat likes the box. \
The box is over the cat."""))
```

[1]:
```
Counter({'The': 3,
         'cat': 3,
         'is': 2,
         'in': 1,
         'the': 3,
         'box': 3,
         '.': 3,
```

```
         'likes': 1,
         'over': 1})
```

[2]:
```
counter = Counter(
    word_tokenize("""The cat is in the box. The cat likes the box. \
    The box is over the cat."""))
counter.most_common(2)
```

[2]:
```
[('The', 3), ('cat', 3)]
```

1.3 Practice question for bag-of-words picker:

- It's time for a quick check on the understanding of bag-of-words. Which of the below options, with basic NLTK tokenization, map the bag-of-words for the following text?

  "The cat is in the box. The cat box."

  ☐ `('the', 3), ('box.', 2), ('cat', 2), ('is', 1).`

  ☐ `('The', 3), ('box', 2), ('cat', 2), ('is', 1), ('in', 1), ('.', 1).`

  ☐ `('the', 3), ('cat box', 1), ('cat', 1), ('box', 1), ('is', 1), ('in', 1).`

  ☒ `('The', 2), ('box', 2), ('.', 2), ('cat', 2), ('is', 1), ('in', 1), ('the', 1).`

▶ Question-solving method:

[3]:
```
from nltk.tokenize import word_tokenize
from collections import Counter

Counter(word_tokenize("The cat is in the box. The cat box."))
```

[3]:
```
Counter({'The': 2, 'cat': 2, 'is': 1, 'in': 1, 'the': 1, 'box': 2, '.': 2})
```

1.4 Practice exercises for word counts with bag-of-words:

▶ Package pre-loading:

[4]:
```
from nltk import word_tokenize
```

▶ Data pre-loading:

[5]:
```
article = open('ref1. Wikipedia article - Debugging.txt').read()
```

▶ Bag-of-words `Counter` building practice:

[6]:
```
# Import Counter
from collections import Counter

# Tokenize the article: tokens
```

```
tokens = word_tokenize(article)

# Convert the tokens into lowercase: lower_tokens
lower_tokens = [t.lower() for t in tokens]

# Create a Counter with the lowercase tokens: bow_simple
bow_simple = Counter(lower_tokens)

# Print the 10 most common tokens
print(bow_simple.most_common(10))
```

```
[(',', 151), ('the', 150), ('.', 89), ('of', 81), ("'", 66), ('to', 63), ('a', 60), ('``', 47), ('in', 44), ('and', 41)]
```