Text Processing

Sources:
“Think Python” by A. Downey

Modified for 219114/115: Paruj Ratanaworabhan

Open a Text File

- Download the file words.txt from the course’s Google Classroom

The built-in function open takes the name of the file as a parameter and returns a file object you can use to read the file.

```
>>> fin = open('words.txt')
```

fin is a common name for a file object used for input. The file object provides several methods for reading, including readline, which reads characters from the file until it gets to a newline and returns the result as a string:

```
>>> fin.readline()
'aa\n'
```

The file object keeps track of where it is in the file, so if you call readline again, you get the next word:

```
>>> fin.readline()
'aah\n'
```

Open a Text File

We can get rid of the new line with the string method strip:

```
>>> line = fin.readline()
>>> word = line.strip()
>>> word
'aahed'
```

You can also use a file object as part of a for loop. This program reads words.txt and prints each word, one per line:

```
fin = open('words.txt')
for line in fin:
    word = line.strip()
    print(word)
```

In-class Exercise

- Find and print only the words with more than 20 characters (not counting space)

- Find and print only the words that have no “e” and compute the percentage of the words in the list that have no “e”

- Prompt the user to enter a string of forbidden letters; then print the number of words that don’t contain any of them and show no more than 5 sample of such words

- Find out how many words are there that use all the vowels aeiou and show 10 sample of such words

- If the letters in a word appear in alphabetical order (double letters are ok), the word is abecedarian. Find out how many abecedarian words are there and show 10 sample of such words

Tackling the Exercise Problems

- Complete and run doctest on the file lesson_8.py; add more test cases if you want to

- lesson_8.py will become a python module that contains functions related to processing English words

- lesson_8 module will be imported to lesson_8_run.py where you will put your code to solve all the exercise problems there

More Text Processing

- Download the file emma.txt from the course’s Google Classroom

- Solve the following problem:

Write a program that:

1. Reads a file, breaks each line into words, strips whitespace and punctuation from the words, and converts them to lowercase.

   Hint: The string module provides a string the following definitions:

   ```
   >>> import string

   >>> string.punctuation

   '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'

   >>> string.whitespace

   ' \t\n\r\x0b\x0c'

   >>> string.ascii_lowercase

   'abcdefghijklmnopqrstuvwxyz'
   ```

   Also, you might consider using the string methods strip, replace and translate.

2. Count the total number of words in the book, and the number of times each word is used; print the total number of words and the 20 most frequently used words in the book.

   Hint: Use Python’s Dictionary

3. Find and print all the words in emma.txt that are not in words.txt