# 09 Prepare: Text Files

Most computers permanently store lots of data on devices such as hard drives, solid state drives, and thumb drives. The data that is stored on these devices is organized into files. Just as a human can write words on a paper, a computer can store words and other data in a file. During this lesson, you will learn how to write Python code that reads text from text files.

## Concepts

Broadly speaking, there are two types of files: text files and binary files. A **text file** stores words and numbers as human readable text. A **binary file** stores pictures, diagrams, sounds, music, movies, and other media as numbers in a format that is not directly readable by humans.

### Text Files

In order to read data from a text file, the file must exist on one of the computer's drives, and your program must do these three things:

1. Open the file for reading text
2. Read from the file, usually one line of text at a time
3. Close the file

The built-in `open` function opens a file for reading or writing. Here is an excerpt from the official documentation for the `open` function:

```
open(filename, mode="rt")
```

> Open a file and return a corresponding file object.
>
> *filename* is the name of the file to be opened.
>
> *mode* is an optional string that specifies the mode in which the file will be opened. It defaults to `"rt"` which means open for reading in text mode. Other common values are `"wt"` for writing a text file (truncating the file if it already exists), and `"at"` for appending to the end of a text file.

Example 1 contains a program that opens a text file named `plants.txt` for reading at line 26. At line 30 there is a `for` loop that reads the text in the file one line at a time and repeats the body of the for loop once `for` each line of text in the file. In the body of the `for` loop at lines 32–38, the code removes surrounding white space, if there is any, from each line of text and then stores each line of text in a list.

```
1  # Example 1
2
3  def main():
```

```
    # Read the contents of a text file
    # named plants.txt into a list.
    text_list = read_list("plants.txt")

    # Print the entire list.
    print(text_list)


def read_list(filename):
    """Read the contents of a text file into a list and
    return the list. Each element in the list will contain
    one line of text from the text file.

    Parameter filename: the name of the text file to read
    Return: a list of strings
    """
    # Create an empty list that will store
    # the lines of text from the text file.
    text_list = []

    # Open the text file for reading and store a reference
    # to the opened file in a variable named text_file.
    with open(filename, "rt") as text_file:

        # Read the contents of the text
        # file one line at a time.
        for line in text_file:

            # Remove white space, if there is any,
            # from the beginning and end of the line.
            clean_line = line.strip()

            # Append the clean line of text
            # onto the end of the list.
            text_list.append(clean_line)

    # Return the list that contains the lines of text.
    return text_list


# Call main to start this program.
if __name__ == "__main__":
    main()
```

```
> python example_1.py
['baobab', 'kangaroo paw', 'eucalyptus', 'heliconia', 'tulip',
'chupasangre cactus', 'prickly pear cactus', 'ginkgo biloba']
```

After the body of a `for` loop that reads from a file, we can write a call to the `file.close` method. However, when calling the `open` function, most programmers use a `with` block as shown in example 1 at line 26 and nest the `for` loop inside the `with` block as shown at lines 30–38. When the `with` block ends, the computer will automatically close the file, so that the programmer doesn't have to write a call to the `file.close` method.

## CSV Files

Many computer systems import and export data in CSV files. CSV is an acronym for comma separated values. A **CSV file** is a text file that contains tabular data with each row on a separate line of the file and each cell (column) separated by a

comma. The following example shows the contents of a CSV file named hymns.csv that stores data about religious songs. Notice that the first row of the file contains column headings, the next four rows contain data about four hymns, and each row contains three columns separated by commas.

```
Title,Author,Composer
O Holy Night,John Dwight,Adolphe Adam
Away in a Manger,Anonymous,William Kirkpatrick
Joy to the World,Isaac Watts,George Handel
With Wondering Awe,Anonymous,Anonymous
```

Python has a standard module named `csv` that includes functionality to read from and write to CSV files. The program in example 2 shows how to open a CSV file and use the `csv` module to read the data and print it to a terminal window. In example 2 at line 8, there is a call to the Python built-in `open` function, which opens the `hymns.csv` file for reading. At line 12, the program creates a `csv.reader` object that will read from the `hymns.csv` file. Within the `for` loop at lines 16 and 17 the `csv.reader` reads and prints each row from the CSV file.

```
 1  # Example 2
 2
 3  import csv
 4
 5  def main():
 6      # Open the CSV file for reading and store a reference
 7      # to the opened file in a variable named csv_file.
 8      with open("hymns.csv", "rt") as csv_file:
 9
10          # Use the csv module to create a reader object
11          # that will read from the opened CSV file.
12          reader = csv.reader(csv_file)
13
14          # Read the rows in the CSV file one row at a time.
15          # The reader object returns each row as a list.
16          for row_list in reader:
17              print(row_list)
18
19
20  # Call main to start this program.
21  if __name__ == "__main__":
22      main()
```

```
> python example_2.py
['Title', 'Author', 'Composer']
['O Holy Night', 'John Dwight', 'Adolphe Adam']
['Away in a Manger', 'Anonymous', 'William Kirkpatrick']
['Joy to the World', 'Isaac Watts', 'George Handel']
['With Wondering Awe', 'Anonymous', 'Anonymous']
```

When a `csv.reader` reads a row from a CSV file, the reader returns the row as a list of strings. The output from example 2 shows that a `csv.reader` returns a list of strings. In the output, notice the five lists of strings, (strings surrounded by square brackets [ ... ]) that were printed by the print statement at line 17. Notice also that the reader reads all the rows from a CSV file, including the first row, which contains column headings.

You might recall that in CSE 110, you wrote a program that reads from a CSV file without using a `csv.reader`. That program split each row of text from the CSV file using the string `split` method. Unfortunately, using the `split` method will not work for all CSV files. Consider the following `hymns.csv` file that contains rows for the hymns "Far, Far Way on Judea's Plains" and "Oh, Come, All Ye Faithful". Both of these hymns have commas in their titles. If we use the string `split` method to separate the columns in this CSV file, the hymn titles will be split. A `csv.reader` will correctly split rows in all valid CSV files.

```
Title,Author,Composer
"Far, Far Way on Judea's Plains",John Mcfarlane,John Mcfarlane
"Oh, Come, All Ye Faithful",John Wade,John Wade
"Christ the Lord is Risen Today",Charles Wesley,Anonymous
```

## Processing Each Row in a CSV File

After reading each row from a CSV file, the `for` loop in the previous example simply prints the row list to a terminal window. Of course, a `for` loop can do much more than simply print each row. Consider the following CSV file named dentists.csv that stores data about dental offices. Notice that the first row of the file contains column headings, the next four rows contain data about four dental offices, and each row contains five columns separated by commas.

```
Company Name,Address,Phone Number,Employees,Patients
Eagle Rock Dental Care,556 Trejo Suite C,208-359-2224,7,1205
Apple Tree Dental,33 Winn Drive Suite 2,208-359-1500,10,1520
Rockhouse Dentistry,106 E 1st N,208-356-5600,12,1982
Cornerstone Family Dental,44 S Center Street,208-356-4240,8,1453
```

The program in example 3 processes each row in the `dentists.csv` file to determine which dental office has the most patients per employee. Notice that the first row of the `dentists.csv` file contains column headings. The headings contain no numbers and aren't needed for the calculations, so the program skips the first row by calling the built-in `next` function at line 25.

```
 1  # Example 3
 2
 3  import csv
 4
 5  # Indexes of some of the columns
 6  # in the dentists.csv file.
 7  COMPANY_NAME_INDEX = 0
 8  NUM_EMPS_INDEX = 3
 9  NUM_PATIENTS_INDEX = 4
10
11
12  def main():
13      # Open a file named dentists.csv and store a reference
14      # to the opened file in a variable named dentists_file.
15      with open("dentists.csv", "rt") as dentists_file:
16
17          # Use the csv module to create a reader
18          # object that will read from the opened file.
19          reader = csv.reader(dentists_file)
20
21          # The first row of the CSV file contains column
22          # headings and not data about a dental office,
```

```
        # so this statement skips the first row of the
        # CSV file.
        next(reader)

        running_max = 0
        most_office = None

        # Read each row in the CSV file one at a time.
        # The reader object returns each row as a list.
        for row_list in reader:

            # For the current row, retrieve the
            # values in columns 0, 3, and 4.
            company = row_list[COMPANY_NAME_INDEX]
            num_employees = int(row_list[NUM_EMPS_INDEX])
            num_patients = int(row_list[NUM_PATIENTS_INDEX])

            # Compute the number of patients per
            # employee for the current dental office.
            patients_per_emp = num_patients / num_employees

            # If the current dental office has more
            # patients per employee than the running
            # maximum, assign running_max and most_office
            # to be the current dental office.
            if patients_per_emp > running_max:
                running_max = patients_per_emp
                most_office = company

    # Print the results for the user to see.
    print(f"{most_office} has {running_max:.1f}"
          " patients per employee")


# Call main to start this program.
if __name__ == "__main__":
    main()
```

```
> python example_3.py
Cornerstone Family Dental has 181.6 patients per employee
```

# Reading a CSV File into a Compound List

The program in example 3 reads and processes each row in a CSV file. That program needs to access the data in each row once only. If a program needs to access the contents of a CSV file multiple times, the program can read the contents of the file into a compound list and then access the data from the list. The program in example 4 contains a function named `read_compound_list` that reads the contents of a CSV file into a compound list.

```
# Example 4

import csv

def main():
    # Read the contents of the dentists.csv file
    # into a compound list.
    dentists_list = read_compound_list("dentists.csv")

    # Print the entire list.
    print(dentists_list)
```

```
12
13
14  def read_compound_list(filename):
15      """Read the contents of a CSV file into a compound
16      list and return the list. Each element in the
17      compound list will be a small list that contains
18      the values from one row of the CSV file.
19
20      Parameter filename: the name of the CSV file to read
21      Return: a list of lists that contain strings
22      """
23      # Create an empty list that will
24      # store the data from the CSV file.
25      compound_list = []
26
27      # Open the CSV file for reading and store a reference
28      # to the opened file in a variable named csv_file.
29      with open(filename, "rt") as csv_file:
30
31          # Use the csv module to create a reader object
32          # that will read from the opened CSV file.
33          reader = csv.reader(csv_file)
34
35          # Read the rows in the CSV file one row at a time.
36          # The reader object returns each row as a list.
37          for row_list in reader:
38
39              # If the current row is not blank,
40              # append it to the compound_list.
41              if len(row_list) != 0:
42
43                  # Append one row from the CSV
44                  # file to the compound list.
45                  compound_list.append(row_list)
46
47      # Return the compound list.
48      return compound_list
49
50
51  # Call main to start this program.
52  if __name__ == "__main__":
53      main()
```

```
> python example_4.py
[['Company Name', 'Address', 'Phone Number', 'Employees',
'Patients'], ['Eagle Rock Dental Care', '556 Trejo Suite C',
'208-359-2224', '7', '1205'], ['Apple Tree Dental',
'33 Winn Drive Suite 2', '208-359-1500', '10', '1520'],
['Rockhouse Dentistry', '106 E 1st N', '208-356-5600', '12',
'1982'], ['Cornerstone Family Dental', '44 S Center Street',
'208-356-4240', '8', '1453']]
```

# Reading a CSV File into a Compound Dictionary

If the values in one of the columns of a CSV file are unique, then a progam can read the contents of a CSV file into a compound dictionary and then use the dictionary to quickly find data. Recall that each item in a dictionary is a key value pair. The values from the unique column in a CSV file will be the keys in the dictionary. The program in example 5 shows how to read the data from a CSV file into a compound dictionary. Notice in example 5, because of lines 9, 14,

58, and 62, that the program uses the dental office phone numbers as the keys in the dictionary.

```
# Example 5

import csv


def main():
    # Index of the phone number column
    # in the dentists.csv file.
    PHONE_INDEX = 2

    # Read the contents of the dentists.csv into a
    # compound dictionary named dentists_dict. Use
    # the phone numbers as the keys in the dictionary.
    dentists_dict = read_dictionary("dentists.csv", PHONE_INDEX)

    # Print the dentists compound dictionary.
    print(dentists_dict)


def read_dictionary(filename, key_column_index):
    """Read the contents of a CSV file into a compound
    dictionary and return the dictionary.

    Parameters
        filename: the name of the CSV file to read.
        key_column_index: the index of the column
            to use as the keys in the dictionary.
    Return: a compound dictionary that contains
        the contents of the CSV file.
    """
    # Create an empty dictionary that will
    # store the data from the CSV file.
    dictionary = {}

    # Open the CSV file for reading and store a reference
    # to the opened file in a variable named csv_file.
    with open(filename, "rt") as csv_file:

        # Use the csv module to create a reader object
        # that will read from the opened CSV file.
        reader = csv.reader(csv_file)

        # The first row of the CSV file contains column
        # headings and not data, so this statement skips
        # the first row of the CSV file.
        next(reader)

        # Read the rows in the CSV file one row at a time.
        # The reader object returns each row as a list.
        for row_list in reader:

            # If the current row is not blank, add the
            # data from the current to the dictionary.
            if len(row_list) != 0:

                # From the current row, retrieve the data
                # from the column that contains the key.
                key = row_list[key_column_index]

                # Store the data from the current
                # row into the dictionary.
                dictionary[key] = row_list
```

```
63
64      # Return the dictionary.
65      return dictionary
66
67
68  # Call main to start this program.
69  if __name__ == "__main__":
70      main()
```

```
> python example_5.py
{'208-359-2224': ['Eagle Rock Dental Care', '556 Trejo Suite C', '208-35
'208-359-1500': ['Apple Tree Dental', '33 Winn Drive Suite 2', '208-359-
'208-356-5600': ['Rockhouse Dentistry', '106 E 1st N', '208-356-5600', 1
'208-356-4240': ['Cornerstone Family Dental', '44 S Center Street', '208
```

# Additional Information

The following tutorials contain additional information that you may find helpful. You are not required to read these tutorials.

- Python "for" Loops
- Writing Text Files in Python

# Summary

A text file stores words and numbers as human readable text. During this lesson, you are learning how to write Python code to read from text files. To read from a text file, your program must first open the file by calling the built-in `open` function. You should write the code to open a file in a Python `with` block because the computer will automatically close the file when the `with` block ends, and you won't need to remember to write code to close the file.

A CSV file is a text file that contains rows and columns of data. CSV is an acronym that stands for comma separated values. Within each row in a CSV file, the data values are separated by commas. Python includes a standard module named `csv` that helps us easily write code to read from CSV files. Sometimes a program simply needs to use the values in a CSV file in calculations, so we write Python code to perform calculations for each row. Other times, we write Python code to read the contents of a CSV file into a compound list or compound dictionary.