Homework due Jul 13, 2021 22:00 +06

Exercise 1
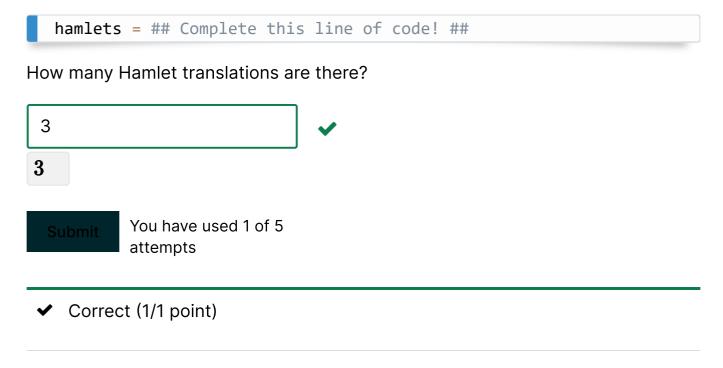
1/1 point (graded)

Note that `book_titles` is a nested dictionary, containing book titles within authors within languages, all of which are strings. These books are all stored online, and are accessed throughout this case study. In Exercise 1, we will first read in and store each translation of Hamlet.

Instructions

Read in the data as a pandas dataframe using `pd.read_csv`. Use the `index_col` argument to set the first column in the csv file as the index for the dataframe. The data can be found at <u>this link within the courseware</u> ⧉, and at <u>this link when coming from outside the courseware</u>.

Complete the following line of code to read in the data:

```
hamlets = ## Complete this line of code! ##
```

How many Hamlet translations are there?

```
3
```
✔

3

Submit   You have used 1 of 5 attempts

✔ Correct (1/1 point)

Exercise 2

1/1 point (graded)

In Exercise 2, we will summarize the text for a single translation of Hamlet in a `pandas` dataframe.

Instructions

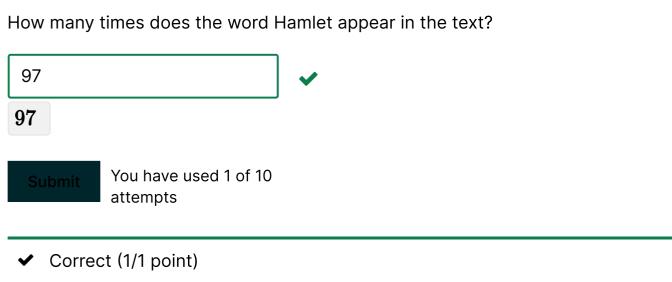Find the dictionary of word frequency in `text` by calling `count_words_fast()`. Store this as `counted_text`.

Create a `pandas` dataframe named `data`.

Using `counted_text`, define two columns in `data`:

- `word`, consisting of each unique word in `text`.

- `count`, consisting of the number of times each word in `word` is included in the text.

Here's the code to get you started:

```
    language, text = hamlets.iloc[0]

# Enter your code here.
```

How many times does the word Hamlet appear in the text?

```
97
```
✔

97

Submit   You have used 1 of 10 attempts

✔ Correct (1/1 point)

Exercise 3

1/1 point (graded)

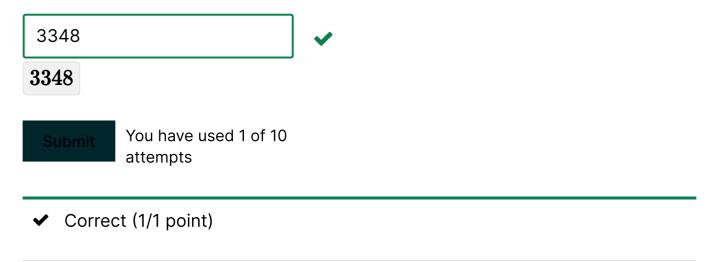In Exercise 3, we will continue to define summary statistics for a single translation of Hamlet.

Instructions

Add a column to `data` named `length`, defined as the length of each word.

Add another column named `frequency`, which is defined as follows for each word in `data`:

- If `count > 10`, `frequency` is "frequent".
- If `1 < count <= 10`, `frequency` is "infrequent".
- If `count == 1`, `frequency` is "unique".

How many unique words appear in the text?

| 3348 | ✔ |

3348

| Submit | You have used 1 of 10 attempts |

✔ Correct (1/1 point)

Exercise 4

1/1 point (graded)
In Exercise 4, we will summarize the statistics in `data` into a smaller `pandas` dataframe.
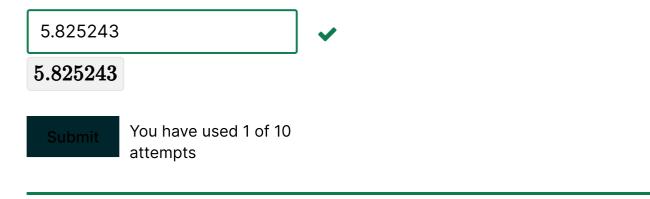
Instructions
Create a `pandas` dataframe named `sub_data` including the following columns:

- `language`, which is the language of the text (defined in Exercise 2).

- `frequency`, which is a list containing the strings "frequent", "infrequent", and "unique".

- `mean_word_length`, which is the mean word length of each value in `frequency`.

- `num_words`, which is the total number of words in each frequency category.

What is the average word length of the infrequent words?

5.825243

✔

5.825243

Submit   You have used 1 of 10
attempts

✔ Correct (1/1 point)