Assignment 6

Data Files: Alexa Reviews.csv

The **Alexa Reviews.csv** file consists of 3,150 records with Amazon customers' reviews (text), star ratings (integer 1 to 5), date of review and variant of various Alexa products like Alexa Echo, Echo dot, etc.

The file uses 'cp1252' encoding. Thus, if you are using a mac, you might need to specify encoding='cp1252' when you use the open() function to open the file for reading:

```
open("alexa reviews.csv","r", encoding='cp1252')
```

Question 1.

Write code that creates a new csv file with a headers row and a record per type of device (values in *variation* field), with the device name, the number of reviews, and the average rating for that type of device. You should not hardcode the current values in the variation field; that is your program will work correctly even if later other types of devices are added to the Alexa Reviews.csv file.

Hint. As you loop over the records, start populating a dictionary where the keys are the variations and the corresponding values are lists with the count and average. The latter will need to be updated every time you find a record for a variation already in the dictionary.

Question 2.

Write a program that reads the Alexa Reviews.csv file (here need to use a DictReader), examines sentences from the reviews' text that include the exact word "sound" (not case sensitive), and determines what are the 10 most common words, with at least 4 characters, in those sentences (do not include "sound" as one of them) and the respective counts. Your program should write the output to a new csv file with two fields, *Word* and *Count*. Include a headers row in the output csv file.

Examining words should not be case sensitive, and you should remove punctuations from end of words so that "love!" and "Love" would be treated as same word.

Notes.

1. When you loop over the reviews, you will need to first split a review's text to a list of sentences. Sentences are separated by ".", "?", or "!", but the split() method takes only one character as its argument, so think how to do this correctly. **Hint.** You will need to first use the replace() method.
2. When you loop over the sentences from a given review text, you will need to split each sentence to a list of words and then check if "sound" is in that list. If so, you will need to do more work with this list of words. **Hint.** See the example we did in class for creating a dictionary with words and the frequencies of those words.
3. When you are done going over all the reviews, you should have a single dictionary with words as keys and counts as values. Now you need to find the top 10 used words. It is not possible to sort a dictionary. But, you can get a list of the keys sorted according to the values. Try the following

```
Ages={"Hila":44,"Pete":6,"Roy" :50}
Names=sorted(Ages ,key=Ages.__getitem__, reverse=True ) #two underscores on each side
```

```
print(Names)
```

Names will be the list [‘Roy’, ‘Hila’, ’Pete’] since Roy has the largest value, 50 , and Pete the smallest value, 6.

Once you get a sorted list of the words, you can use a slice operator to get a sublist with just the first 10 items from the complete sorted list. Walla- you have the top 10 most used words in sentences with the word “sound”. Now think how to write each of those words, with its corresponding value from your dictionary with all the words and count, to a csv file.