HW#2 EDIT DISTANCE

Name: Rudraksh Sugandhi

Observation Criteria:

For spell\_checker() function:

1. In this assignment we are trying to find the miss spelled word in each text on the basis of dictionary created in the howework1 assignment.
2. For the same we are taking simple input from user in text format, if the text format is contains any upper or lower case elements, typecasting to lower case using .lower() python built function.
3. As per the instruction if the, the input provide by the user is “quit” the code returns “Goodbye” message otherwise continue.
4. Since the punctuation and special character doesn’t play any role in determining the spellings, defined regex just to capture the word in the sentence using regex [^\w\s], which means only accept character which have word in starting and then space at the end.
5. Used split() function and saved the file in list to iterate over it later.
6. Since the stop word is removed from dictionary in assignment 1, its required to remove it here otherwise there will be ambiguity for the stop word in this assignment.
7. Opening the dictionary in read mode and storing it in list to compare the words from the text sentence list, if there is some which word which are not in dictionary it should considered as miss-spelled.
8. For checking if the miss- spelled word is not available i.e. is if the length of the list is zero we are just printing the “ No misspellings detected! .
9. Here I have created two dictionaries, the first dictionary is used to take one word from the sentence which is miss-spelled and calculate the distance of that word with each word in the dictionary.
10. The second dictionary is looping through for each word in sentence and storing the minimum distance of word in sentence to that of in the dictionary.
11. If we considered few example sentence where the function working correctly and having issues in some:

|  |  |  |
| --- | --- | --- |
| **Sentence** | **Output** | **Reason** |
| “The computer is on table, with chair and equipment.” | No misspellings detected! | Removed the stop word so word like ‘The’,’is’, ‘on’ ‘and’ and ‘with’  is already removed.  Words ‘computer’, ‘table’,‘chair’ and ‘equipment’ is available in dictionary. |
| The raen in Spain, fals mainly on the plain. | Misspelling - Suggestion  raen - ran  spain - pain  fals - falls  mainly - mail | The punctuation is not having any effect on the miss-spelling of the word and the algorithm randomly predict the nearest word to raen that is ran, similarly for the others. |
| The name of series is Sherlock Holmes | Misspelling - Suggestion  sherlock - lock  holmes - homes | These words are correct but because this word is not available in dictionary the code will convert it into word which is having smallest distance. |
| My email address is rudraksh.sugandhi1@gmail.com | Misspelling - Suggestion  rudrakshsugandhi1gmailcom - animal | There is the ambuity that it is trying to convert the email address we can add extra code to make sure its doesn’t change the email address as this address can contain any special character and numbers. |

1. The possible way of resolving the problem with unknown word is providing more dataset which will create a dictionary or bag of word which contain all the letters on internet. In that way output will be more logically correct and accurate.
2. Finally looping through each key and value pair in final dictionary to print the miss-spelled word.

For the Min\_distance() function:

1. Most important part of this is to implement the formula which we discussed in class.
2. As the code available everywhere for reference on Wikipedia is providing the cost of 1 when we change one letter or character with another letter or character.
3. But according to the professor its counted as 2, that’s a kind of unique thing which was highlighted in the assignment documentation.
4. In this code we basically need to create matrix which is of the length M\*N where M is the length of first word and N is the length of other word.
5. Creating zeros matrices with NumPy library, and initialize the column and rows to the numbers so where the (I,0) we replaced with I and for (j,0) its replace with J.
6. Implementing the formula for words+1 because if we end it at word it will not capture the last character in the word.
7. There are various way to implement this algorithm we selected the formula which was displayed in the class.
8. For the cost I have used “temp” variable to update exchange one character from another and assigned the value 2 for that exchange of any character in the word.
9. Finally return the last value of the matrix value the value which give the total value or distance of the word or the cost of changing the whole word to new word.
10. The few problems which I faced with creating matrices is understanding the end case condition where and when to stop the loops and to understand the whole matrices. Manually trying to create the metrics on pen and paper to verify correctability of code.