Document Classifying Algorithm

Learning:

Document Learning Module:

The module takes in > Directory to a Class & **Class** Name

A **class** is created using the **Document Class** module

Get the list of files of in the directory passed to the module

For Loop > for each file in the directory > create a document > using the **document** module

Each file is read using the **read_doc** module

> **Read_doc**: Text file is read > all words set to lower > each word Is added to the bag of words of that particular document
>
> Each document has a **Bag Of Words.**

Using the **__add__** operator the files in the directory are joined and added to the **class**

The **class** is added to the list of learned_classes

The number of documents in that class is recorded and the location of that class is saved

Classifying:

Load text file chosen

Probability List created > stores the probability of the text file given a class

1st For Loop > For each learned class pass the chosen text file and class to the Probability module

Get sum of the words in the learned class

Read the chosen text file

2nd For Loop > For each learned class > get sum of words in that learned class

3rd For Loop > For each word in the chosen text file get the frequency of the word in the learned classes

Calculating Probability : The probability is returned and added to the list of probabilities

```
for word_in_doc in test_document.Words():

    if word_in_doc not in removable_words_symbols:

        freq_of_words_given_docClass = 1 + self.learned_classes[document_class].WordFreq(word_in_doc)
        freq_of_words_in_class = 1 + self.learned_classes[_class].WordFreq(word_in_doc)

        result = freq_of_words_in_class * sum_of_words_in_doc_class / (freq_of_words_given_docClass * sum_of_words_in_class)
        product *= result

prob += product * self.learned_classes[_class].NumberOfDocuments() / self.learned_classes[document_class].NumberOfDocuments()
```

```
1st For Loop
Learned Class:  heart
2nd For Loop
    Class:  heart
    Class:  volcanoes
    Class:  mongol

1st For Loop
Learned Class:  volcanoes
2nd For Loop
    Class:  heart
    Class:  volcanoes
    Class:  mongol

1st For Loop
Learned Class:  mongol
2nd For Loop
    Class:  heart
    Class:  volcanoes
    Class:  mongol


[['volcanoes', 0.9930928364385707], ['mongol', 0.00690715540362863], ['heart', 8.157800575740672e-09]]
```

Each class is loaded twice every time the probability is calculated for a class