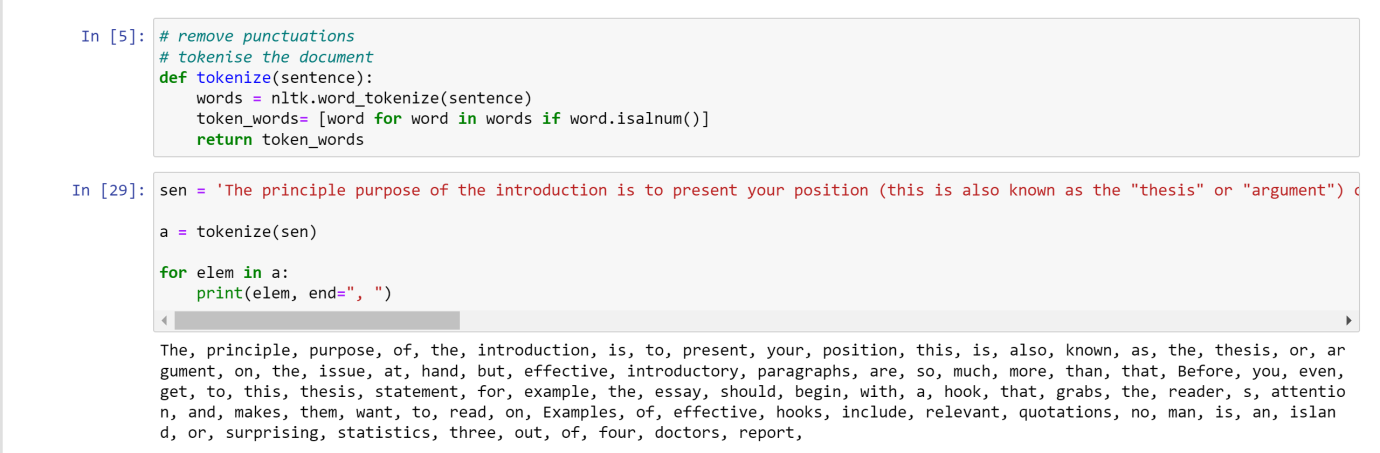
**Tokenizing and Removing Punctuation:**

The Input format of the below code is in the form of ‘**String**’.

The Output format of the below code is in the form of the ‘**List**’.



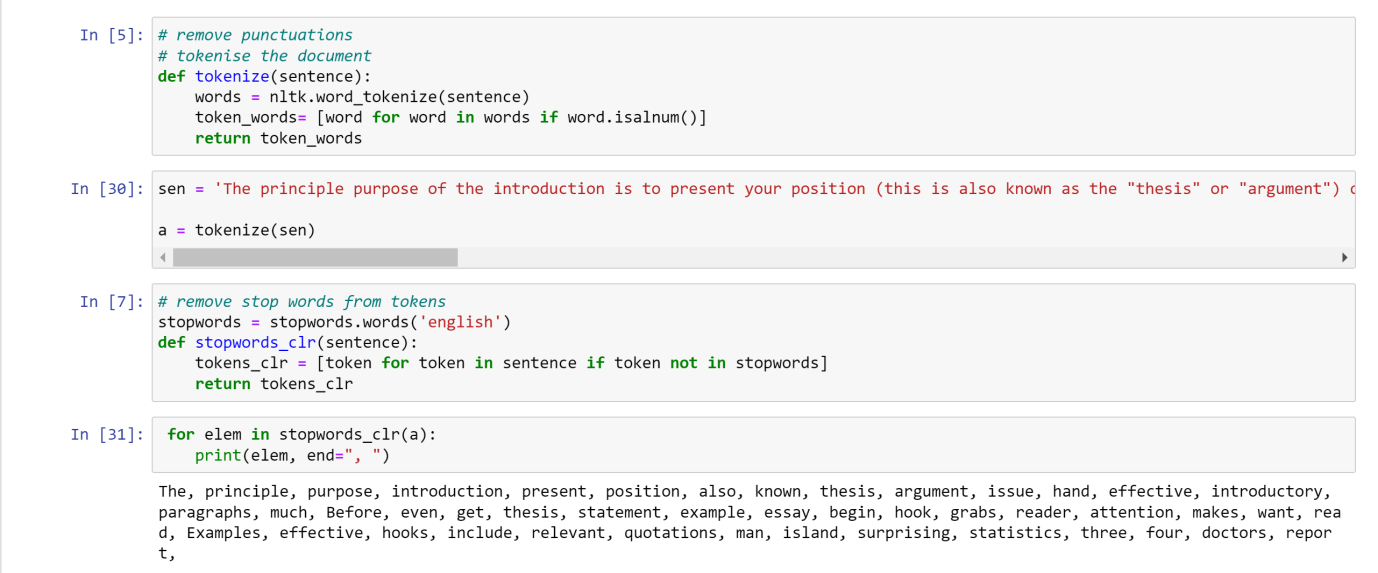
The above code tokenizes the given string, removes punctuation and outputs a ‘List’ which does not have punctuations.

**Removing Stop words:**

The input of the below code is in the form of ‘**List**’.

Output is also in the form of a ‘**List**’.

Here we are using the output of the above ‘Tokenizing’ function and passed the output as parameter to ‘stopwords\_clr’ function.



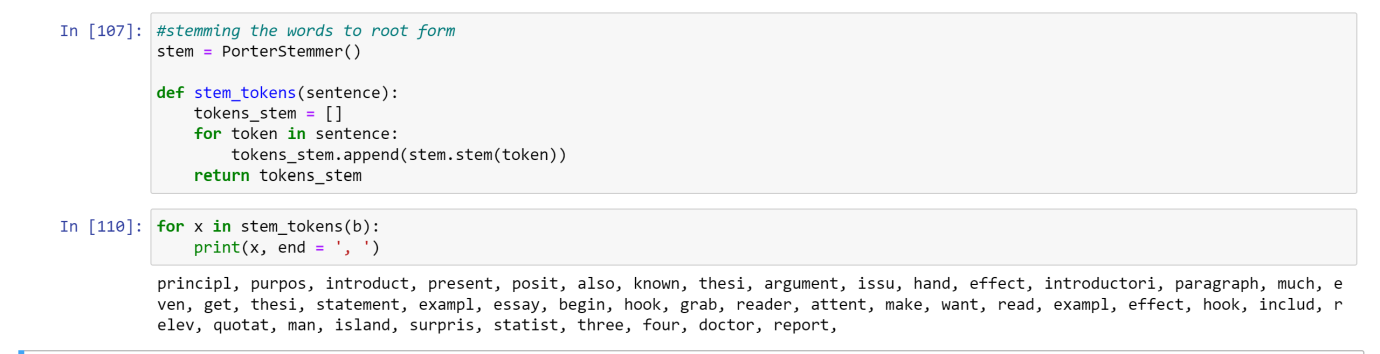
The above function removes all the stop words and outputs a ‘List’ which doesn’t contain stop words.

**Stemming the document:**

Here the input is given in the form **‘List**’.

Output is given in the form of ‘**List**’.

Here we used the output of ‘stopwords\_clr’ function above and passed it as a parameter to ‘stem\_tokens’



The output of the above function is a ‘List’ of words which are stemmed.

**Inverted Index:**

Here the input is in the form of a set of documents (List) and index of the document (Int).

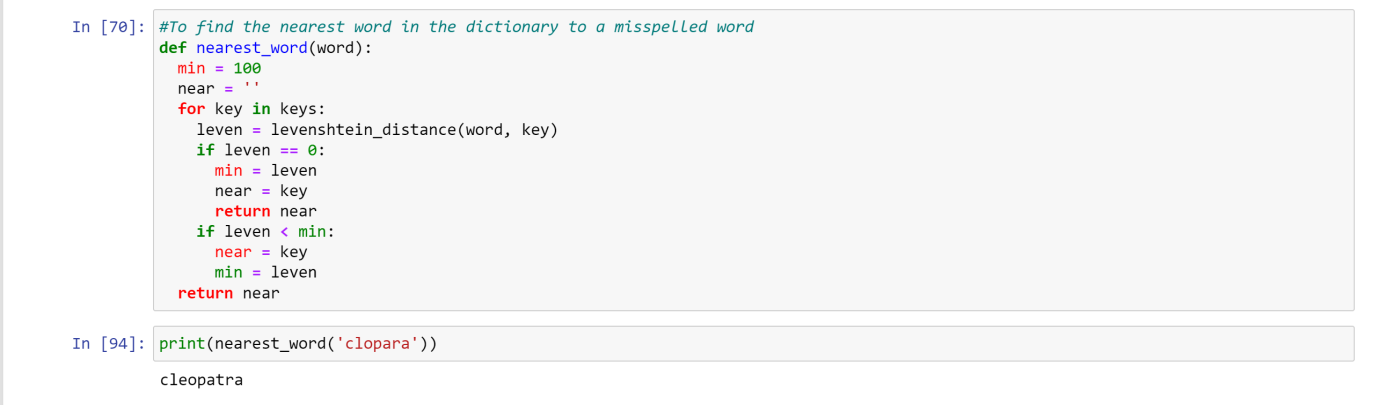
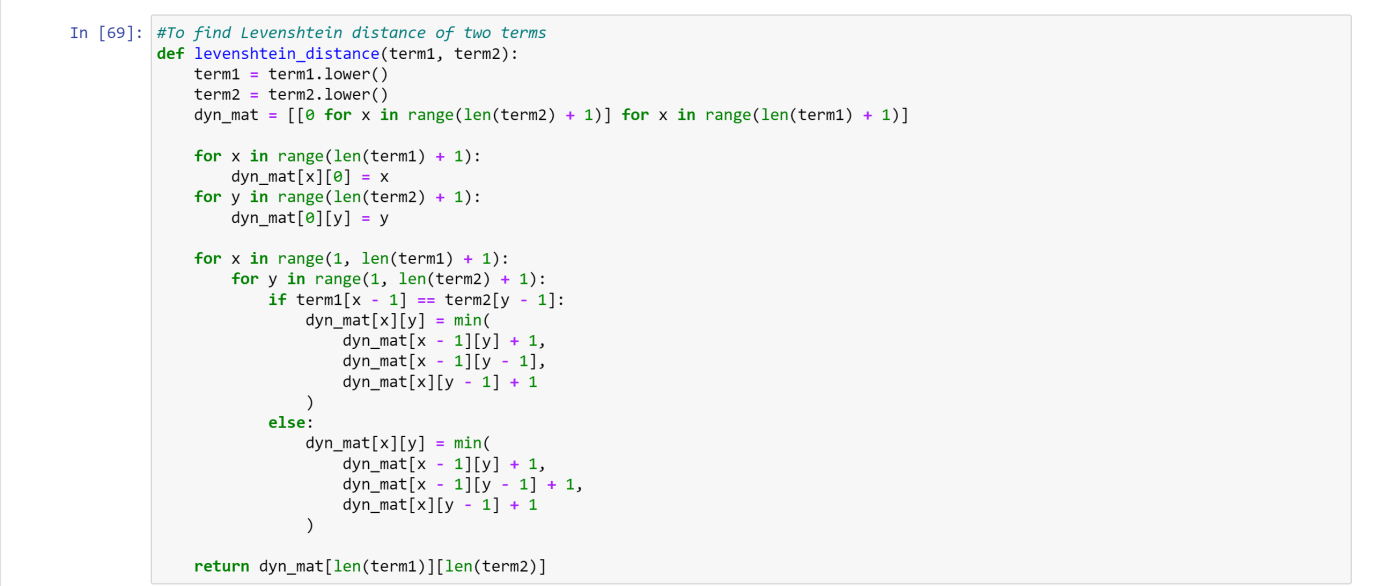
The output we get is a **dictionary** which have key-value pairs, where words of the documents are keys and posting list of the keys are the values of the dictionary respectively.



In the above code we passed a folder containing 4 documents. All the documents are first tokenized, stemmed and the stop words are removed. We got a dictionary containing key-value pairs of words of document and their respective posting lists.

**Spelling Correction:**

Here the below function, ‘nearest\_word’ takes the input as a **String** for which we used Levenshtein distance to find the nearest word in the document if any.



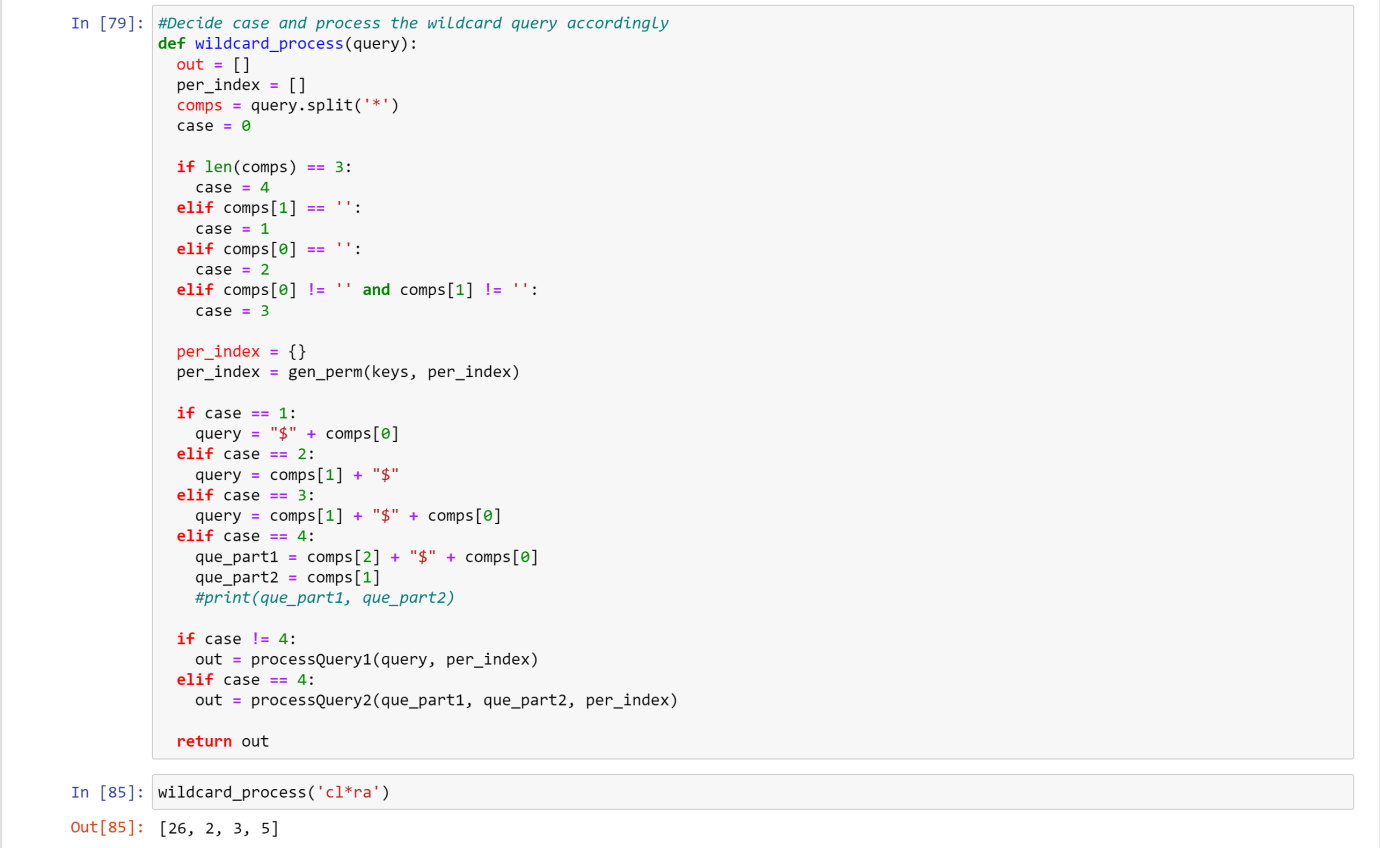
Output of the above function is a **String** which is the closest word in the document.

**Querying:**

**1.Wildcard Query:**

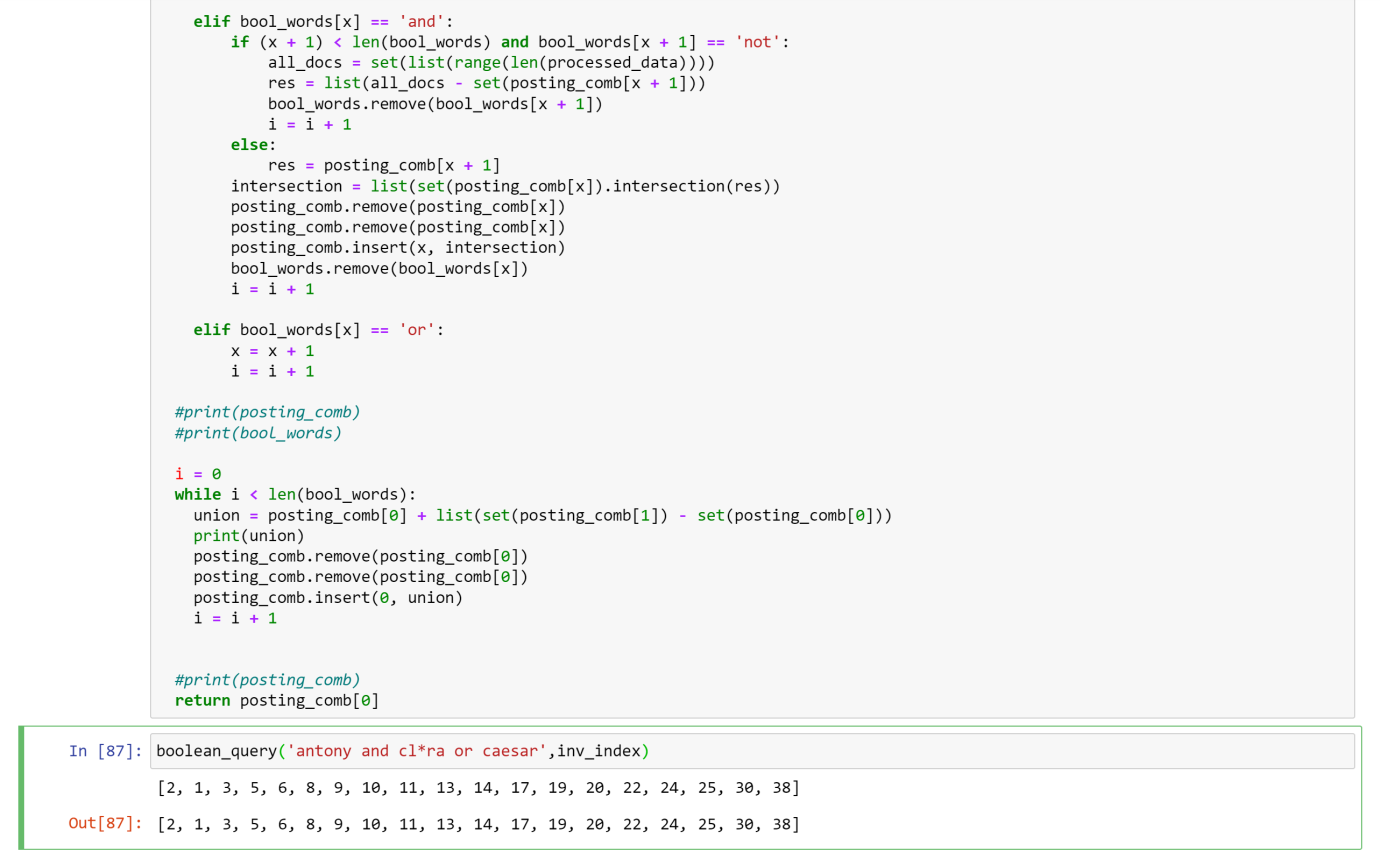
Here the below function, ‘wildcard\_process’ takes the input in the form of **String**.

It gives Output of **list** of the documents in which the wildcard is present.



**2.Boolean Query:**

Here the below function, ‘boolean\_query’ takes the input in the form of a Query (**String**) and the Inverted Index (**Dictionary**).



The above function gives output of ‘**List**’ of documents which satisfy the query.

Final Project Input :

Final Project Output :