Tokenization in Python - Detailed Explanation

1. Importing the regex library:

import re: This imports the 're' module, which provides support for regular expressions in Python.

2. Defining Token Types:

TOKEN_TYPES = { ... }: This dictionary defines different token types, where the key is the token type name and the value is the regex pattern to match that type.

3. Combining patterns:

TOKEN_REGEX = '|'.join(...): This combines all token patterns into a single regex using '|' as the OR operator.

4. Defining the tokenize function:

def tokenize(code):: This defines the function that will tokenize the input code.

5. Initializing the token list:

tokens = []: This initializes an empty list to hold the extracted tokens.

6. Iterating through each line:

for line in code.splitlines():: This loops through each line of the input code.

7. Stripping whitespace:

line = line.strip(): This removes any leading and trailing whitespace from the line.

8. Skipping empty lines:

if not line:: This checks if the line is empty and skips it if true.

9. Finding matches:

for match in re.finditer(TOKEN_REGEX, line):: This uses finditer to search for all matches in the line using the combined regex.

10. Getting token type:

token_type = match.lastgroup: This retrieves the type of the matched token.

11. Getting token value:

token_value = match.group(token_type): This retrieves the actual value of the matched token.

12. Avoiding whitespace:

if token_type != 'WHITESPACE': This checks to avoid adding whitespace tokens to the list.

13. Returning the tokens:

return tokens: This returns the list of extracted tokens.

14. Main program block:

if __name__ == '__main__': This checks if the script is being run directly.

15. User input:

code_input = input('Enter your code: '): This prompts the user to enter their code.

16. Tokenization call:

tokens = tokenize(code_input): This calls the tokenize function to analyze the input code.

17. Printing results:

for token_type, token_value in tokens:: This iterates through the extracted tokens and prints their types and values.