Code

1. Longest Matching Tokenizer

```
# Từ điển là một tập hợp

dict = {"tôi":0, "là":1, "công": 2, "dân": 3, "nước": 4, "việt": 5, "nam": 6, "công dân": 7, "việt nam": 8 }

# Xây dựng hàm tách từ

def tokenizer (text, dict, is_show=False):

#in câu đầu vào

  print ("input:", text)

  print()

#tách câu đầu vào thành các âm tiết

  input = text.split(" ")

#words là mảng gồm các từ được tách ra từ câu

  words = []

  s = 0

  while True:

    e = len(input)

    while e>s:

      tmp_word = input[s:e]

      is_word = ""

      for item in tmp_word:

        is_word += item + " "

      is_word = is_word[:-1]

      e -= 1

      #print (is_word)

      if is_word.lower() in dict:
```

```
        words.append(is_word)
        break
      if e==s:
        words.append(is_word)
        break
    if e>= len(input):
      break
    if is_show:
      print("s =", s)
      print("e =", e)
      print(words[len(words)-1])
      print("-"*100)
    s = e + 1
  output = ""
# In kết quả
  for item in words:
    output += item.replace(" ", "_")
    output += " "
  output = output[:-1]
  return output
text1 = "Tôi là công dân nước Việt Nam"
test1 = tokenizer(text1, dict, False)
print("output:", test1)
```