Statistical Machine Translation Lab Exercise 2: Tokenisation and Cleaning

Please use Python as your programming language for this lab
You can use the same data as last weeks lab - located in Loop

1. Write a program that takes in a file, and outputs a tokenised file.
E.g.
Input: They're over John's bridge!
Output: They ' re over John ' s bridge !

Compare it to your tokenised file from last week - are there any differences?

2. Write a program that removes sentences from pairs of files over a certain length.
E.g., given 2 tokenised files train.en train.fr and the number 4, any sentences in train.en that are over 4 tokens in length should be removed, along with the corresponding sentence in train.fr, and vice versa.

Before cleaning:

train.en	train.fr
he is a doctor	c ' est un docteur
Where are you from ?	D ' où êtes - vous ?
I am Claude	Je suis Claude
The sun	Le soleil

After cleaning:

train.en	train.fr
I am Claude	Je suis Claude
The sun	Le soleil

Next steps: alter the program so that it takes a minimum length and a maximum length. E.g. given 2 and 10, any sentences with less than 2 tokens or greater than 10 tokens should be removed.