**Tutorial 7 – OCR**

A Biology teacher is teaching her students about the influenza virus. She has been explaining that the most common 4 symptoms of this virus are:

* Headache
* Fever
* Fatigue
* Cough

In an assignment the teacher has asked the students to mention the most common symptoms of influenza. The students submitted the assignment as a hard copy.

The teacher is trying to automate the correction process of this particular question. She wants to award a student 1 mark for every correct symptom he mentions (thus a maximum of 4 marks). The data has been given to you in the form of images (each image containing the answer of 1 student).

The following steps will help you:

1. Create an array containing the 4 correct symptoms
2. Copy this function into your script. The job of this function is to give you back a list containing only the common words found in two lists given as input.

**def** intersection(lst1, lst2):  
 lst3 = [value **for** value **in** lst1 **if** value **in** lst2]  
 **return** lst3

1. In a loop, read each image and use OCR to convert the student’s answer to text.
2. Use the wordpunct nltk tokeniser to tokenise the student’s answer.
3. Use the intersection function how many of the 4 symptoms he/she mentioned.
4. For each student, print in the console the answer he/she gave, as well as the marks he/she got.

Expected Output:

