MEDICAL INFORMATION ANALYTICS

NLP LIVE CHALLENGE

COMPOUND NOUNS

DESCRIPTION

In healthcare, the International Classification of Diseases (ICD) is a widely used catalog of clinical codes for coding diseases in medical documents, such as doctor letters. The catalog exists in language-specific variations for many countries. For Germany, the DIMDI provides a localized version of all clinically relevant diagnosis. It contains textual descriptions for every known code, e.g.

```
I77.2;Arterienriss
N30.9;Harnblaseninfektion
S42.00;Klaviculafraktur
N83.2;Ovarialzyste
S00.95;Schädelprellung
S72.00;Schenkelhalsfraktur
C01;Zungengrundkarzinom
```

The first column is the ICD code, the second column the corresponding textual description.

Words like „Artierienriss“ or „Zungengrundkarzinom“ are *compound nouns,* i.e. nouns that consist of more than one word. German compound nouns are usually written out as a single word, which makes them challenging to process with NLP tools. In real world medical documents, you'll rarely find the original descriptions literally in the text. Instead, you face lots of variations which also should be matched to the same code. E.g. “Riss der Arterie“ has the same meaning as „Artierienriss“ and should be assigned I77.2. Here are a few more real world examples for split-up variants of the descriptions above:

Arterienriss - Riss der Arterie
Harnblaseninfektion - Infektion der Harnblase
Klaviculafraktur - Fraktur der Klavicula
Ovarialzyste - Zyste des Ovars
Schädelprellung - Prellung des Schädels
Schenkelhalsfraktur - Fraktur des Schenkelhalses
Zungengrundkarzinom - Karzinom des Zungengrundes

TASK

Find a solution to match *compound noun* descriptions such as the list above against split-up variants of those. Your solution should be able to assign a single ICD code to a given textual input. As your input data, assume the small ”ICD catalog” provided in the beginning of description above. Optionally, you can also use the full catalog provided as an extra .zip file.

Input:

`Zungengrundkarzinom`

Expected Output:

`C01`

Input:

`Karzinom des Zungengrundes`

Expected Output:

`C01`

You will work in a pairing mode, with a data scientist from MIA in the navigator role and you taking the driver role most of the time. The navigator will make sure you don’t get stuck on details, carefully direct you to the next steps if necessary and also help you keeping the big picture in mind.

SOLUTION CRITERIA

The goal of this exercise to see if you can come up with an idea to tackle an open NLP challenge and write maintainable and well-tested code that will make it easy to work with you in a team. The goal is not to create the most elegant or fastest solution.

Use python with any libraries of your choice. For this live challenge, it is not necessary to use git (you can do this of course if you are used to work with it in your usual workflow). A readme, build script, docker file etc. are not required. Your implementation should be executable without crashing. Your solution should be covered by automatic tests.

You are not expected to finish the task completely, but your solution should at least be able to read the input and produce a partially correct result without throwing any exceptions for some of the given examples. Good luck and have fun! ☺