Notes on the Spatial Databases part of the Large Scale Databases coursework

Counting true positives, false positives and false negatives:
For each individual caption in the data provided, note there is only one ground truth toponym. Thus for *an individual caption,* using the methods for computing precision and recall explained in the document `PrecisionRecall_2` (uploaded to the Week 8 folder in Learning Central Learning Materials, and to the `Files and Data for CM3104` folder of the coursework section on Learning Central):

-there cannot be more than one true positive;
-there cannot be more than one false negative;
-if there is a true positive then there cannot also be a false negative.

For the NER tasks (but not the geocoding tasks) there could be more than one false positive, if your classifier identified more than one of the non-ground truth items of text as a toponym.

Note that even though it could be argued that, in a couple of cases in the provided data, there might be other words/phrases that could regarded as a toponym, you have to work only with the ground truth data that are provided.

Question 1
Use the basic spacy NER tool with the default model `en_core_web_sm`

Question 2
There is no fixed correct answer to this question. Note that it is possible to obtain the marks without modifying the spacy NER tool that you used in Question 1. However, if you choose to modify the spacy NER tool (but keeping with spacy) that will be acceptable.
Your answer must make clear what you have chosen to do.
Note also that in Question 2, unlike in Question 1, there is no constraint on what values you select for the NER labels, thus different solutions become possible.

Questions 3 and 4
The toponyms that you are trying to geocode here must be ones that match the "ground truth toponym", as found in Q1 and Q2. Your inputs to the Geopy geocoder should therefore only be toponyms that match a 'ground truth toponym'. Hence there will only ever be a maximum of one toponym to be geocoded per caption.
These toponyms are ones that you found to be true positives in Q1 / Q2. If you did not find a true positive for a caption in those NER questions then for those captions you will not be geocoding any toponym.

Question 4
It is possible that you will find more than one instance of the geocoded toponym (that matches the ground truth toponym) that is within 20km of the guide coordinates. If so, choose just one of them.

Question 6
Where the question requests you to list the names of captioned places (all in your database and those within 400km of the "Farm track north of Aglionby") you should list the full text of the respective captions. Thus you are not being asked to extract the toponym from the caption (though that was requested in previous questions).

If you have not retrieved the coordinates for "Farm track north of Aglionby" you can insert a record for it in your database using the guide coordinates provided in the json file, and make a note of that in your solution to the question.

Question 7
The intention in this question, stated in the question, is that you use a different NER tool from that used in questions 1 and 2.

If you find that your use of a different python library does not improve on results you can obtain with spacy, explain this in the explanation part of your answer.

In the table for this question in the solutions template, report on the best results that you obtained for this question. Note that this could be worse than that provided in Question 4.