Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CM3104
Module Title: Large Scale Databases
Lecturer: C.B. Jones
Assessment Title: Coursework Part B
Assessment Number: 1
Date Set: Week 4, Friday 6th November 2020
Submission Date and Time: Week 11, Wednesday 13th January 2021 at 9:30am
Return Date: Wednesday 10th February 2021

This assignment is worth 30 % of the total marks available for this module. If coursework is submitted late (and where there are no extenuating circumstances):

1 If the assessment is submitted no later than 24 hours after the deadline, the mark for the assessment will be capped at the minimum pass mark;
2 If the assessment is submitted more than 24 hours after the deadline, a mark of 0 will be given for the assessment.

Your submission must include the official Coursework Submission Cover sheet, which can be found here:

https://docs.cs.cf.ac.uk/downloads/coursework/Coversheet.pdf

Submission Instructions

Description		Type	Name
Cover sheet	**Compulsory**	One PDF (.pdf) file	[student number].pdf
PART B	**Compulsory**	One PDF (.pdf) file that is saved/exported from the provided SpatialDatabasesSolutionsTemplate file and contains all your answers to each question in exactly the format specified by that template file.	PartB_[student number].pdf
	Compulsory	Five Python scripts – one for each of questions 1, 2, 3, 4 and 7. It is acceptable to put these files into a single zip file.	PartB_[student number]-SQn.py where n in SQn is the question number, for example : PartB_c123456-SQ2.py Optional zip file name: PartB_c123456-SQpy.zip
	Compulsory	For Question 5, One text file that you used as input to QGIS or ArcGIS	PartB_[student number]-SQ5.txt

Any Python scripts submitted will be run on a Windows or MacOS command line with the command “python [filename].py” and SQL queries will be run in SQL Developer, and must be submitted as stipulated in the instructions above.
Any deviation from the submission instructions above (including the number and types of files submitted) will result in a mark of zero for the respective part of the assessment.

Staff reserve the right to invite students to a meeting to discuss coursework submissions.

Assignment

The CM3104 coursework consists of 2 parts; Part A and Part B, each worth 30% of the total coursework mark for this module. This document is Part B.

Spatial Databases Task

Spatial indexing of text documents with Named Entity Recognition (NER) and Geocoding.

Instructions for Spatial Databases Task

The aim of the assignment is to geocode (i.e. find the coordinates) of the place names (toponyms) in the provided set of photo captions and to use the resulting coordinates to create an Oracle Spatial Database, along with a GIS (QGIS or ArcGIS) project that can be used to map the data.

The task requires the use of the SpaCy named entity recognition function to detect the presence of the place names in each caption and the use of the GeoPy library to geocode the detected places. The final question allows you to use a different NER tool.

You are provided with a json file called json-capLatLong.json that lists the photo caption, the ground truth toponym, the guide coordinates as latitude and longitude, and the full address of the disambiguated toponym for each caption.

There are 7 questions for the Spatial Databases task.

Python scripts MUST be submitted for Questions 1, 2, 3, 4 and 7 of the Spatial Databases component and they must have names of the form PartB_[student number]-SQn.py where n in SQn is the number of the question and StudentNumber is your student number. Thus for spatial databases Question 1, if your student number is c123456 the script must be called PartB_c123456-SQ1.py
For questions 1, 2, 3 and 4 the only library installs that can be required for your Python scripts are spacy and geopy.

QUESTION 1 : Detect place names (toponyms) in the provided captions using the default SpaCy Named Entity Recognition interpretation of place names (i.e. as any FAC, GPE or LOC).

[5 marks]

Run the SpaCy NER software to detect all entities that are classified as some form of geographic feature, i.e. FAC, GPE or LOC.
Report on precision, recall and F1 relative to the provided ground truth recorded as property “ground truth toponym” in the file json-capLatLong.json.

Your program will need to output:
a) for each caption, with one line per caption:
the caption and the entity in it identified by SpaCy as a geographic feature;
b) the values you have computed for true positive, false positive and false negative;
c) the values you have computed for Precision, Recall and F1.

In the Solutions Template file you must insert these pieces of information as indicated AND provide a clearly legible screenshot of the above outputs.

QUESTION 2 : Improve on the default performance of the SpaCy NER tool for detecting place names in the caption text. **[2 marks]**

Attempt to improve on the precision, recall and F1 of Question 1 above using just the SpaCy NER tool (i.e. no other NER tool) in whatever way you wish, ***but you must use named entities identified by the SpaCy NER tool and not use any other NLP (natural language processing) tool.***

Your program will need to output:
a) for each caption, with one line per caption:
the caption and the entity identified by the SpaCy NER tool;
b) the values you have computed for true positive, false positive and false negative;
c) the values you have computed for Precision, Recall and F1 – a single set of based on applying whatever is the best method that you implemented.

In the Solutions Template file you must insert these pieces of information as indicated AND provide a clearly legible screenshot of the above outputs. Also explain briefly but very clearly the changes that you made to the previous script to improve the performance and discuss any limitations that your method might have with other text data.

QUESTION 3 : Geocoding the toponyms. **[4 marks]**

Take the above output of SpaCy from your best method from questions 1 and 2 and use it as input to GeoPy, with the Nominatim gazetteer (AND **NOT** WITH ANY OTHER GAZETTEER), and “limit=1” (i.e. you only report the first answer from GeoPy, even though there could be multiple possible answers).

Report the precision, recall and F1 of its performance in detecting the correct instance of the place name (toponym) that you have identified with the SpaCy NER tool.
Use a threshold distance of 20km between the detected place name instance and the corresponding “guide coordinates” to decide if the detected instance is correct. Guide coordinates are the photo subject coordinates (from the Geograph website) and they are in provided JSON file (json-capLatLong.json).

Your program will need to output:
a) for each geocoded caption, the full address of the Identified Toponym, as provided by Nominatim; its coordinates as latitude and longitude; and the distance in km units between the retrieved name and the guide coordinates;
b) the values you have computed for true positive, false positive and false negative;
c) the values you have computed for Precision, Recall and F1.

In the Solutions Template file you must insert these pieces of information as indicated AND provide a clearly legible screenshot of the above outputs.

QUESTION 4 : Improving the quality of geocoding the toponyms **[5 marks]**

This question requires you to consider all candidate toponyms retrieved by the Nominatim gazetteer (rather than just the first) and to use the caption “guide coordinates” to assist in choosing the correct instance of the candidate geocoded toponym. Note that the guide coordinates are close to the correct coordinates but they ARE NOT THE SAME.

Write a script, that again uses the SpaCY NER tool and the GeoPy geocoder to attempt to improve on the Precision, Recall and F1 in Question 3, using all of:
- the place names identified with your best SpaCY NER method (i.e. from questions 1 or 2); these must be the same place names used in Question 3;
- the GeoPy geocoder, with ONLY the Nominatim gazetteer, to generate multiple candidate place names and their coordinates, i.e. set the ‘limit’ parameter to a value higher than 1;
- the guide coordinate for each caption to assist in choosing the toponym instance.

Your program will need to output:
a) for each geocoded caption, the full address of the **single** selected Identified toponym, as provided by Nominatim; its coordinates as latitude and longitude; and the distance in km units between the retrieved name and the guide coordinates;
b) the values you have computed for true positive, false positive and false negative;
c) the values you have computed for Precision, Recall and F1.
In the Solutions Template file insert these pieces of information as indicated AND provide a clearly legible screenshot of the above outputs.
Also provide a description of the methods that you used.

QUESTION 5 : mapping the retrieved locations **[4 marks]**

Create a map in QGIS or ArcGIS of the caption locations that you found in Question 4 using the coordinates that you obtained there.

NOTE: **DO NOT** USE THE GUIDE COORDINATES PROVIDED TO YOU, i.e. you must use the coordinates that you retrieved from Nominatim using GeoPy.

As output in the Spatial Databases Solutions Template provide:

a) An explanation of how you created the data layer in QGIS or ArcGIS.
b) A listing of the text file that you used as input to QGIS or ArcGIS, i.e. with caption texts and coordinates that came from the output of Question 4.
c) A screen shot of a map of the point locations of the captions. Note that the points must be clearly visible, and ideally labelled with the corresponding caption text. You could also add some background map data for Britain to provide context.

Note that an Ordnance Survey GeoTiff map of Britain can be found at https://osdatahub.os.uk/downloads/open/GBOverviewMaps

QUESTION 6 : create an Oracle database of the retrieved disambiguated locations. **[5 marks]**

Create a table in Oracle containing only the captions and point locations (coordinates) that you found in Question 4 above (i.e. the same data used for input to QGIS or ArcGIS in Question 5) and perform two SQL queries:
1) to list the names of all captioned places in your database;
2) to find the names of all captioned places that are within 400 Km of the place with caption “Farm track north of Aglionby”.

As output in the Spatial Databases Solutions Template:

a) provide an explanation of how you created the Oracle table, i.e. converting from the data generated in Question 4;
b) write the two SQL queries and their outputs (not screenshots);
c) provide screenshots of the two queries and their outputs.

QUESTION 7 : improving on the NER process and hence the resulting geocoded caption locations. **[5 marks]**

Attempt to improve on the performance that you obtained in Question 4, making use of any other Python NER (Named Entity Recognition) tool you wish, but you must stay with exactly the same GeoPy geocoding functionality (with Nominatim) that you used in Question 4.

[Question continued on the next page]

As output, in the Spatial Solutions Template, provide:

a) a clear explanation of what Python NER tool you used (with a link to the Python software and its required installs and imports in Python) and of how you used it;

b) for each geocoded toponym, the retrieved location address, its coordinates and the distance from the guide coordinates;

c) the precision, recall and F1 of your method;

d) a screen shot of running the program and its printed output which should consist only, for each geocoded caption, of the full geocoded location address and its latitude and longitude coordinates, and its distance from the guide coordinates.

Learning Outcomes Assessed

1. Demonstrate an appreciation of applications of large-scale databases in a variety of commercial, scientific and professional contexts;
2. Be able to choose and develop a non-relational database solution suitable for the type of data and application considered;

Criteria for assessment

Credit will be awarded against the following criteria.

Spatial Databases Task

The criteria against which the questions will be assessed, and which vary in importance between the questions, are:

1. Accuracy of your solutions with regard to how well they identify toponyms and disambiguated toponyms relative to the best results that can be expected for each question, including correct computation of precision, recall and F1 relative to the ground truth data.

2. Correctness of the Oracle SQL query formulation and answers.

3. Clarity of explanation of the methods used where this is requested. Brief but clear and informative explanations will be preferred over more convoluted, lengthy explanations.

4. Clarity / legibility of the Python script output printouts as provided in screenshots.

5. Insight into and appropriateness of the methods used, as reflected in the explanations and any discussion provided.

Application of criteria to the spatial databases questions:

Component & Contribution	***Fail***	***Pass (40-49%)***	***2(2) (50 –59%)***	***2(1) (60 – 69%)***	***First (>= 70%)***
Questions 1,2,3,4	*No solution, or no screenshots, or no python script, or solution matches none or very little of best expected solution.*	*Some progress towards correctly identified elements, and some explanation or discussion.*	*Reasonable progress to correct answers but some errors in identified elements, or problems with explanation or discussion.*	*Most of the identified elements are correct. Or all correct but limitations in the explanations or discussion.*	*Results as good as can be expected and their explanation and presentation clear.*
Question 5	*No solution or no screenshot or map locations are wrong.*	*Basic map with adequate explanation of methods.*	*Map locations appear correct but map clarity poor or problems in explanation of methods used*	*Map locations correct but some limitations in presentation and/or explanation.*	*Clear very well presented map with correct locations and good explanation of methods used.*
Question 6	*No solution or no screenshots or SQL code or outputs mostly wrong.*	*Data loaded to Oracle and some progress in SQL queries.*	*Problems in some of the queries and solutions, or mostly correct but inadequate explanation of how table created.*	*Queries and solutions good and adequate explanation of how table created, or some error in SQL.*	*Queries and solutions good and clear and appropriate explanation of how the table was created.*
Question 7	*No solution or no screenshots, or no python script, or lack of appropriate use of alternative NER tools.*	*Some progress in use of alternative NER tools with some explanation.*	*Moderate evidence of progress in using alternative NER tool appropriately and reasonable explanation of method.*	*Good progress towards improving the solution with possible limitations in quality of solution or of explanation.*	*Good improvement in performance with very clear explanation of NER methods.*

Undergraduate

1st (70-100%)
2.1 (60-69%)
2.2 (50-59%)
3rd (40-49)
Fail (0-39%)

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned within 4 weeks via Learning Central.