GEOG0051 Mining Social and Geographic Datasets (2020-2021) Coursework Instructions

Stephen Law[1,*] **and Nikki Tanu**[1,**]

[1]Department of Geography, University College London, London,UK
[*]stephen.law@ucl.ac.uk
[**]n.tanu@ucl.ac.uk

1 Overview of Tasks

The coursework for the module consists of two separate tasks. The first concerns analysing the *Gowalla_Cambridge_mobility_patterns GC* dataset and the second concerns a machine learning task analysing a *venue-review* dataset. Although each of these tasks will have sub-prompts to be answered, your responses to each of them should be in the form of a coherent report addressing all of these prompts, rather than discrete paragraphs specifically answering individual prompts. Finally, any datasets that you require for these tasks will be uploaded on the Assessment tab on the course Moodle page.

1.1 Submission format

Students should submit a report through Turnitin on the course Moodle page, under the 'Assessment' tab, containing a description and analysis of the methods taken and results obtained,

- in a PDF document with text of font size 11 or 12 and written fully in complete sentences, e.g. not using bullet points,
- of a maximum length of 2,500 words which you are free to divide in any way between your responses for the two tasks.
- The word count includes the title, headings and sub-headings, introduction, conclusion and captions of figures or tables, but excludes the coursework cover page and bibliography (list of references) at the end of the document.
- The maximum number of figures is 10 in total (multiple sub-figures used to make the same point are allowed) and the relevance of these figures should be explained in your write-up.

The code developed by the student should be submitted using a separate submission link available on the course Moodle page in **a single ZIP (compressed) file**. The code can be submitted as either Jupyter notebook(s), i.e. .ipynb files, or as a .py files, but they must be contained within one ZIP file.

The submission deadline is **noon on the 26th of April, 2021**. Further details on the submission procedures will be available on Moodle.

Note: FAILURE TO INCLUDE YOUR FULL NOTEBOOKS/CODE WILL INCUR A 10-POINT PENALTY.

1.2 Queries

A sub-channel has been created specifically for queries about the coursework to be asked in. All related queries **must be** posted in this sub-channel; this is largely to address a likely overlap in questions that students may have and so that all students will benefit from any clarification that is given.

Questions seeking clarification about, for instance, the wording of the task briefs or format of submission will be answered. However, as this is an assessed piece of work, you **may not** ask about questions that pertain directly to the coursework itself, e.g. *"Is analysis X the best way to answer question 1a?"* Because of the same reason, any collaboration or discussion of the coursework with anyone is strictly prohibited. The rules for plagiarism apply and any cases of suspected plagiarism of other works, published or not, will be taken very seriously.

The deadline for any questions to be asked and answered is **noon on the 19th of April, 2021**, i.e. 1 week before submission deadline (26th of April, 2021).

2 Mobility Patterns Analysis in Cambridge

For the first task, you will be analysing the mobility patterns of users from *Gowalla*, a now-defunct online geo-social network from a decade ago. On Gowalla, users were able to check in at different locations across the course of the day. The dataset that is provided to you (available on Moodle) is a subset of Gowalla users located in Cambridge, UK and, although with some personal identifiers of the users removed, you could trace the movements of particular individuals on certain days, according to their check-ins.

For further information, the entire dataset is available at https://snap.stanford.edu/data/loc-gowalla.html.

2.1 Format of Data

The variables contained in the dataset (which should be self-explanatory), provided in a .csv file, are:

- **User_ID**, or the unique identifier of the user, e.g. 196514
- **check-in-date**, e.g. 2010-07-24
- **check-in-time**, e.g. 13:45:06
- **latitude**, e.g. 53.3648119
- **longitude**, e.g. -2.2723465833
- **loc_id**, or the unique identifier of the location, e.g. 145064

2.2 Analysis Prompts

2.2.1 Visualise individual check-in locations

Visualise the check-in locations of the GC dataset for users with **User_IDs [75027] and [102829]** using the Folium library. Comment briefly on your findings of the locations visited by the 2 users, using any library that enables mapping. You should also comment briefly on the privacy implications of this type of analysis. **[Note: This task primarily serves to help familiarise you with the dataset; we advise not to spend too long on it.]**

2.2.2 Provide Characterisation of the Gowalla dataset

Provide a characterisation of the data available for the **user [75027] on 30/01/2010** and for **user [102829] on 11/05/2010**, by visualising the paths for both users using the OSMnx library. Then, summarising your answers in a table in your report and compute, for each user:

- the maximum displacement (i.e. maximum distance between two consecutive locations they moved between);
- the average displacement (i.e. average distance between two consecutive locations/check-ins);
- the total distance travelled on the day;
- ****Note:** All distances should be described in network distance, i.e. the distances of paths along the street networks, rather than geographical distances without consideration of the street paths.

2.2.3 Comparative analysis of check-in frequencies and network centrality

Describe the general pattern of user check-ins in the Gowalla dataset in relation to closeness centrality measure for the City of Cambridge, UK, using whatever visual aids you see as fitting to your analysis. Comment on any observable trends which you find most noticeable and/or interesting.

2.2.4 Urban Planning Application Question

Imagine that you were taking the role of a consultant to the authorities in Cambridge responsible for urban planning. Choose one of the following urban features and propose a new location where you would build that feature: museum, shopping mall, fire station, community park. Use the outputs of your analysis from the task above (2.2.3) and any relevant knowledge of the local area to justify your decision. **[Note: You do not have to do any further analysis/ visualisation by code. However, if you feel like your response could benefit from further analysis, you can choose to briefly describe what accompanying analysis you would undertake.]**

3 Machine Learning Analysis with Venue Review Data in Calgary, Canada

For this second task, we would like you to analyse a dataset that contains review data of different venues in the city of Calgary, Canada. With the help of several machine learning techniques that we have learnt in the course, you will be tasked to distill insights from this social media dataset. Two of its notable features are the geocoding of every reviewed venues and the availability of a considerable amount of text data in it, which lend to its ability to be processed using spatial and text analysis techniques respectively.

As a prelude to the analysis prompts below, have a brief think about some of these questions: What can we discover about the venue review data? Are there any spatial patterns that can be extracted from the data? Can we build a machine learning model that predicts review rating for unseen data points using the text of the reviews?

3.1 Format of Data

The variables contained in the dataset provided in a .csv file, are:

- **'business_id'**, unique identifier of the premise
- **'Name'**, name of premise
- **'latitude', 'longitude'**, i.e. the locational attributes of the venue
- **'review_count'**, or the number of reviews the venue has been given
- **'categories'** general category of establishment that a venue falls under *(Note: this variable is rather messy and requires cleaning to be used)*
- **'hours'**, or the opening hours of the venue
- **'review_id'**, unique identifier of the review
- **'user_id'**, unique identifier of the individual who left the review
- **'stars_y'** individual ratings of the venue
- **'useful', 'funny', 'cool', 'text'**, i.e. tags for the review (similar to " of Likes" for a review.)
- **'date'**, i.e. the date of the review

3.2 Analysis Prompts

3.2.1 Loading and cleaning the textual dataset

In a realistic context, most text datasets are messy in their raw forms. They require considerable data cleaning before any analysis can be conducted and, not unlike data cleaning for non-textual datasets, this would include the removal of invalid data, missing values, and outliers. In this first prompt you will be required to complete the tasks stated below to prepare the dataset for subsequent analysis.

- Load and understand the dataset.
- Think about which attributes you will use / focus on (in subsequent prompts) and check its data distribution.
- Pre-process the text review data and create a new column in the data frame which will hold the cleaned review data.
- Some of the steps to consider are: removal of numbers, punctuation, short words, stopwords, lemmatise words, etc.

Note that while there are no immediate outputs from this prompt that you will be assessed on, you will be assessed on the process of data cleaning that you detail in your report. Furthermore, the quality of your data clean for a text analysis task will strongly impact your outputs and thus you should spend a reasonable proportion of your time on this task.

3.2.2 Build a supervised learning model for text analysis

The objective of this sub-task is to build a supervised learning model that predicts the star ratings of the venue data, based on the different features of each review included in the dataset. You can choose a subset of venues to review for example based on a *general category*. You can use a combination of text and non-text features, and below are some guidelines that you could follow:

- Firstly, vectorise the preprocessed review text data to give text features you can used in your model.
- Split your dataset into a train and test-set and train $K \geq 2$ machine learning models to predict the star rating varying either features used (E.g. bag of words features vs TF-IDF features) or choice of models.
- Report the model test results.
- Discuss and interpret the results you obtained.

3.2.3 Geospatial analysis and visualisation of review data

Having explored the dataset, its constituent variables and coverage above, the objective of this sub-task is for you to visualise any of the spatial patterns that emerge from the data that you find interesting. This task is intentionally open-ended and leaves you with some choice. To achieve this, you should:

- Choose 1 or 2 variables *(including any variables you generated from 3.2.2)* that you wish to explore and from the list of variables available in the dataset
- Use either or both of the geopandas and folium libraries in Python to produce up to 3 visualisations
- Comment on the spatial distributions of the 1-2 variables you chose, any trends or outliers that emerge and if they have any notable implications.
- **Note:** You may use any subset of the dataset instead of the entire dataset, but comment on why you chose this subset.

3.2.4 Extra task (Optional)

For extra marks, you could choose **1** of **EITHER**:

- **(a)** Use a pretrained neural word embedding method (ie. word2vec) for the supervised learning task and compare the results with the bag of words features, **OR**,
- **(b)** Apply topic modelling (eg. LDA) on the text data and give a characterisation of each of the topics that your topic model generates. Comment briefly on whether these characterisations were roughly what you expected before, **OR**,
- **(c)** Run a lexicon-based sentiment analysis (eg. NLTK Vader Sentiment Analyser) on the textual data, then report and discuss the results. Does the lexicon sentiment score associate with the venue ratings provided by the users?