Coursework 1 Instructions

Network Data Analysis Coursework 1 Instructions

The coursework consists of a series of connected tasks. In each task, you should consider and apply the ideas and techniques learnt in the module along with your technical skills as an urban data scientist. The tasks are not a set of options: **all four tasks must be completed by each student**. The submission will be a single written report documenting the results of these tasks.

Primary Data:

Please see the folder shared via the link provided on KEATS. You are each allocated a data file titled with your name contained within this folder. The data represents the similarities between various classical, pop/rock and jazz songs that are typically played in musical festivals in cities around the world. This data comes from various open music datasets (Schubert Winterreise [1], Isophonics [2], JAAH [3]); these have been further processed in the Polifonia project [4] to establish their *harmonic similarity*. The data has been split into different, equally representative parts, and each of you has a different subset. There is an accompanying dataset that adds further information about the songs, bands, authors, lyrics, and so on.

For this assignment, you should assume that the data is representative of typical songs that are played by bands in music festivals in cities around the world. You can consider extending the dataset with additional information (geographic, demographic, etc.) at your own discretion, if that can improve the depth of your analysis.

[1] https://dl.acm.org/doi/10.1145/3429743

[2] http://isophonics.net/datasets

[3] https://zenodo.org/record/1290737

[4] https://polifonia-project.eu/

Task A (network metrics):

You should first turn this data into a song similarity network for analysis, where we are considering a “cultural harmonic epidemy” that when a song is played in a music festival, chances are very high that a similar song will be played as well elsewhere. The nodes of the network will be the songs, and there will be an edge between each pair of songs if they hold enough harmonic similarity. Songs are harmonically similar if they share some sequence of chords. For example, “Crazy Little Thing Called Love” (Queen) and “P.S. I Love You” (The Beatles) share the harmonic pattern “Ab, Bb, C, Ab, Bb” – hence they have similarity 0.23 (please, note that harmonic similarity ranges in (0, 1], meaning that trivial similarity values of 0 (harmonically dissimilar songs) will not be provided in the dataset (to avoid O(n2) space complexity). You can consider ‘enough harmonic similarity’ between two songs as a threshold that you use as an experimental parameter; this threshold hence denotes sufficient similarity as for the song to become catchy and spark the playback of neighbouring, similar songs. Turning the data into a song similarity network will require thinking through the coding problem and using libraries such as pandas and networkx. You might find it helpful to look at the "Converting to and from other data formats" section of the networkx online documentation.

You should then answer the following questions:

- What are the characteristic properties of this song similarity network?
- How different is this network from a random network?
- If you take it to be a complete and representative description of songs played in music festivals and their tendency to become catchy, what does it tell you about the way songs are played and ignite its similar neighbours?

The documentation of this task should be snippets of code, outputs from running that code, visualisations as appropriate, and accompanying textual explanations for each of these, i.e. as you might see in a well documented Jupyter Notebook or like the code example pages provided on KEATS for this module. You do not need to include every line of code you have written, just those showing the most important and interesting steps.

Task B (epidemic models):

Consider a situation in which the city government and music rights institutions are concerned about massive copyright infringement. When a band plays repeatedly a song that is too similar to other songs, there is a risk to break those other song’s copyright. The overlooking institutions want to monitor for signs of song playback propagation and act quickly to avoid costly lawyer costs. To do so, two randomly chosen songs are selected every day and checked whether they have been played in any music festival. The city and copyright offices want to answer questions such as:

- If both songs have been played on the same day, how can the office use the network data to judge how plausible it is that its catchyness and trend have not propagated yet to neighbouring similar songs endangering their copyright?
- If one or both of the tested songs have indeed been played, then the office will start checking other songs as well, prioritising those with a higher chance of having been played. How should they use the data to come up with a priority list on what songs to check first?

To think about the above, you could consider shortest paths in networks, numbers of similar songs, etc. For the second question, the situation might be quite different depending on whether one song was played, or both were.

The documentation of this task should be first, a textual explanation of how you would tackle questions 1 and 2 and, second, snippets of code and outputs from running that code along with explanations of both as for Task A, to show examples (e.g. best case, an average case and worst case) taking two specific songs from the data as if one or both were played and show the answers to the questions being produced.

Task C (interventions):

Consider the perspective from the city/copyright office. They would like to ask festival organisers to check the playback of songs in-situ as described above to assess copyright infringements. However, festival organisers charge a fee for checking, and thus this a cost for the office; any suing to bands or cancellation of concerts will harm the festival's and city's income, so the festival organisers have incentive not to participate rigorously nor to share the data they have. There is a social network between festival organisers, as they can share premises (e.g. event terrains), trade with each other, be on shared online forums etc. Consider the following questions:

- How might the office use this social network to encourage the adoption of robust checking of played songs?
- What interventions could it consider?
- How might it use simulation to test these interventions before implementing them in practice?

This task does not (necessarily) involve any coding. The documentation of this task should be a textual description suggesting network intervention ideas to the specific problem described.

Task D (network metric-based quality -reading club):

Finally, consider the fact that despite the accuracy of the harmonic similarity algorithm, there is a subjective and social aspect to song similarity. This can make obvious algorithmic links in the network not so obvious (and less prone to propagation and copyright infringement) to the human ear. We want to use network metrics on the song similarity network to assess the quality of its links. Consider the following questions:

- What kind of problems (bias, identity, etc.) can impact the quality of the network?
- What network metrics, and combinations thereof, could be effective for assessing link quality and resolving identity?

To answer these questions, use the papers and the discussion we had about them in the reading club. Reflect on the following question: how useful was the reading club session to answer these questions?

This task does not (necessarily) involve any coding. The documentation of this task should be a textual description suggesting how the ideas discussed in the reading club could be applied to the specific problem described.

Assessment. It is good to apply ideas and methods from any of the modules on your programme (Introduction to Urban Analytics, Telling Stories with Data, Spatial Data Analysis etc.) but the focus should be on networks and network data analysis.

The following criteria will be used to determine the mark for each submission:

- Demonstrated understanding of network data analysis concepts and how they can apply to the questions in the coursework tasks.
- Technical ability in using programming to tackle a data analytics problem, showing ability to research and apply data manipulation techniques as required for the problem.

- Creative thinking about the problems described in the coursework and specifically the network-related aspects.
- Clarity of explanation of what code does and why and what results mean, plus good use of visualisations and presentation.
- Succinctness of reporting, i.e. conveying a lot of substance clearly in a short amount of space. Note that marks will be deducted for exceeding the page limit.

Submission instructions

Deadline. You have approximately 3 weeks between release of the coursework and submission. The strict deadline is **4pm UK time on Friday 25 February**.

Size limits. The report should be at most 6 pages in length including references and figures. You are welcome to add an appendix beyond the 6 pages if you want to document more work you have done but this appendix will not be considered by the markers. The amount of space used per task may vary depending on how much you find to say on each, but expect each task to fill between 1 and 3 pages. Use font size 11 or larger for readability. Otherwise, there is no restriction on the format.

Submission format. The report should be submitted on KEATS as a single PDF document. Ensure you are clear well in advance of the deadline how you will turn your working document(s) into a single PDF file.

Plagiarism, collusion and technical support. You are not allowed to submit anyone else's work as your own (plagiarism) or conduct the coursework in collaboration with other students (collusion). Both are serious matters of misconduct.

However, you are allowed and even encouraged to discuss specific technical problems you face with the coursework and how to solve them with the rest of the class via the KEATS discussion forum. For full guidance on what is acceptable to ask the class and what must be done individually see the 'Coursework questions and collusion' page in the Assessment section of the KEATS page and feel free to ask clarifying questions in the tutorials or the discussion forum.

Last modified: Monday, 7 February 2022, 5:51 PM

◄ Coursework questions and collusion

Jump to...

Datasets for Coursework 1 ►

Administration

Course administration

NMES Student Tech Support

- NMES Tech Support Wiki
- Support Portal

Activities

- Assignments
- External tools
- Feedback
- Forums
- Lessons
- Resources

King's Resources

- Email
- Student Services
- King's Academic Skills for Learning
- Equality, Diversity & Inclusion
- Disability Support
- IT Services
- Library Guides
- LinkedIn Learning
- Health & Wellbeing