

## Term Project: Heterogeneous Data - Group Submission - **Returned**

Title	Term Project: Heterogeneous Data
Groups	Group B'''
Students	Nikita Kapoor (nikitakapoor), Mona Malik (monamalik27), Sannath Vemula (sannathreddyv)
Submitted Date	Dec 15, 2019 10:18 PM
Grade	<b>3.90 (max 4.00)</b>

### Instructions

#### Objective

At this stage, you have a broad understanding of and a solid toolkit for four of the most fundamental types of data: relational, spatio-temporal, graph-based, and textual. Unfortunately, these rarely exist in isolation like how we have been studying them. Indeed, the co-presence of multiple data types, and the need to think concurrently in multiple data models, i.e., the existence of *heterogeneous data*, is what most people face most of the time. It's one of the biggest challenges in Data Science, whether industry or academia.

This project thrusts you right into that challenge. You need to *synthesise* concepts *across* the course.

#### Data

You are provided with an attachment that is an archived [data dump](#) of an entire StackExchange community (incidentally, [the one on Data Science](#)). StackExchange is a question and answer (Q&A) forum, where users contribute "posts" (i.e., questions and answers) which they *upvote*. There is a tagging system to organise posts and a badge system to incentivise users. You can use any subset of it that you want, but it must include at least two files from the attached archive.

The data dump contains a broad selection of the elements we have been looking at previously: the posts themselves are (marked up) temporal text, with revision histories; the votes resemble our previous MovieLens dataset; there are myriad opportunities to construct graph representations of elements of the data. *You are unrestrained!* The data itself is described in the following readme: <https://ia800107.us.archive.org/27/items/stackexchange/readme.txt>.

#### Implementation

It is expected that you will use Python, as it is the language of choice in Data Science. Moreover, it provides many visualisation libraries that will help you achieve more than you realise you can! Also, this should allow you to recycle components from previous assignments.

#### The Task

This time you are to formulate three explicit questions (hypotheses, if you like) that you are trying to evaluate with the data. Your visualisations should answer those questions, or at least shed

meaningful insight into them. You will be partly evaluated on the complexity of these questions and their inherent dependence on data heterogeneity (i.e., the need to use multiple types of data).

### Oral Examination

There is no in-class demo for the term project. Instead, you have an oral examination, worth 50% of the project grade. Each group will give a ten-minute presentation in the style they prefer (e.g., live demo, slideshow). Thereafter, each member of the group (privately, if preferred) will have a circa ten minute individual oral exam. You will be provided with much information about this in December.

Each group can self-select a slot for their presentation on 18-19 December (Doodle to be released later). It is possible to arrange an earlier date by special arrangement, but it would then have to be before 7 December and your report would need to be submitted at least 3 days in advance.

### Submission

You should submit a short technical report (5-10 pages) in pdf format. The report will describe:

- the questions/hypotheses pursued and how they each engage multiple modules of the course
- the insights revealed by the visualisation/application
- the design choices in the implementation/visualisation
- how the project *synthesises* concepts from multiple modules
- challenges encountered (if relevant)
- how this submission meets the requirements set out in the rubric (i.e., a justified self-evaluation)
- other details that you consider relevant

### Grading

The entire group will receive a grade for the written component based on how well the submission adheres to the *first* rubric below, worth 50% of the grade. Note that your report provides the opportunity to persuade the markers, but that they will ultimately grade according to the rubric. This will be established *prior* to the oral exam and released here on connex.

Each group member will then receive an individual grade for the oral exam, based on performance relative to the *second* rubric below. The grade will be issued *immediately after the oral exam*. This comprises the other 50% of the grade and is not included in connex.

Term Project Rubric: Written Component

Component	Weight	4	3	2	1
<b>Transforming Raw Data to Insights</b>	40 %	Visualisation is very informative and tells a story about the data; visualisations are clear and easy to interpret without aide.	Visualisation is informative, albeit possibly with some support from text	Visualisation is complete; information is presented, but lacks complexity or depth.	Very difficult to extract any insights from the visualisation, or no visualisation at all
<b>Synthesis</b>	40 %	The project effectively <i>blends</i> concepts from different modules of the course	Concepts from different modules of the course are utilised, but perhaps in juxtaposition; they are not integrated, but they support each other. Interesting questions are formed; their answers may leverage multiple types of data; potentially only one is needed to arrive at a satisfactory answer.	Multiple modules of the course are engaged, but in a fully independent manner. Some interesting questions are formed, but they do not fulfill the objectives of the assignment.	Project only makes use of one data model. Questions are unclear or very trivial.
<b>Problem Selection</b>	10 %	Interesting questions are formed; their answers clearly depend on multiple types of data.	Submission makes effective use of ideas from any of the research papers presented in class or otherwise goes beyond the core curriculum in a manner that adds substantial values to the project.	Some attempt is made to engage with research papers presented in class, but it adds limited value to the project	Does not engage with research papers presented in class nor other extension concepts.
<b>Extension of Course Concepts</b>	10 %	Submission makes effective use of ideas from multiple research papers presented in class or otherwise goes beyond the core curriculum in a manner that adds substantial values to the project.	Submission makes effective use of ideas from any of the research papers presented in class or otherwise goes beyond the core curriculum in a manner that adds values to the project.	Some attempt is made to engage with research papers presented in class, but it adds limited value to the project	Does not engage with research papers presented in class nor other extension concepts.

Term Project Rubric: Oral Component

Component Weight		4	3	2	1
<b>Conceptual Overview</b>	30 %	Understanding of the broad concept and overall goals of the project submission is well demonstrated.	Understanding of the overall project submission is mostly understood; some gaps may affect independent reproducibility.	Overall project objectives can be described but are not fully understood.	Minimal awareness of how one's own components fit into the larger project objectives.
		Given sufficient time and resources, the student could adequately delegate tasks to reproduce the entire project.			
<b>Individual Contribution</b>	50 %	In a non-relative manner, the student has contributed an important component of the project and can discuss that component is substantial depth.	In a non-relative manner, the student has contributed an important component of the project; the student can independently describe this component.	The contribution of the student appears to be important, but they are not able to describe what they have done.	No clear important contribution is evident.
<b>Link to Curriculum</b>	20 %	Especially with respect to an individual component, the relationship to multiple modules of core course content is excellently described.	Especially with respect to an individual component, the relationship to core course content is described well.	Relationship to core course content is not well described	No clear relationship to core course content is evident.

## Tips

1) Try to load the data into a usable structure early (first few days). You may find it more difficult than you expect to model the raw datasets as workable data structures.

2) Ascend quickly to a working visualization and add complexity later; i.e., first build a minimal viable product (MVP). You will be exposed to new research and new ideas in class as you work on the assignment, so you want to be *agile* with your development patterns. At the same time, even in groups, it may take longer than you expect to go from raw data to a working visualization; so, you don't want to leave this until the last few days before the demo or you could end up with nothing to show at all!

3) Use libraries prolifically.



4) Consider addressing each rubric component individually in your tech report.

5) Reflect closely on the feedback for Assignments 1-4. This is meant to guide you towards a higher grade on subsequent assignments.

## Additional resources for assignment

-  [datascience.stackexchange.com.7z](https://datascience.stackexchange.com.7z) ( 39 MB; Oct 23, 2019 9:17 pm )
- 

## Submitted Attachments

-  [Final Project Report CSC - 501 .pdf](#) ( 1 MB; Dec 15, 2019 10:17 pm )
-  [CourseProjectFiles.zip](#) ( 3 MB; Dec 15, 2019 10:17 pm )

## Additional instructor's comments about your submission

### Overall

4/4/4/3 = 3.9

### Insights

4

+ Cohesive set of questions related to post recommendation, first by relatedness to other posts and then by relatedness to user's self-expressed and latent (tag-based) interests  
- Visualisations themselves could be more informative; e.g., is the graph in question 1 more informative than the simple html response below?

### Synthesis

4

+ Excellent blending of graph and text concepts in the project design  
- ParentId seems to be missing from Posts in the ER diagram; RelatedPostId should be an FK relationship  
- In Step 1 of question 2, it looks like you might have forgotten to exclude self-matches? Have you accidentally paired every user with themselves?

### Problem

4

+ Interesting questions that leverage both graph and text by deriving with one model and refining with the other

### Research

3

+ Good use of extension ideas like recommender systems and clustering

- No explicit relationship to state-of-the-art techniques for recommender systems nor clustering; inventive approach, but not yet scientifically validated (citations would improve score here)

**Questions to Presentation (5-10 slides)**

1. Which is your favourite result in the report and why?
2. In question 1, what do you do if the TF-IDF score is 0 for every related post? In question 2, what if the "About Me" is empty? (This relates to a general concept called "the cold start problem.") What are the limitations of the use of TF-IDF in your project?
3. How would you balance informativeness and data privacy in question 3?