The University of Nottingham

SCHOOL OF COMPUTER SCIENCE

A LEVEL 3 MODULE, SPRING SEMESTER 2019-2020

FUNDAMENTALS OF INFORMATION VISUALIZATION (COMP3021)

Time allowed: 5 working days

Release date: 15th May 2020 15:00
Deadline for submission: 22th May 2020 15:00

Submission format: **PDF document**
Submission method: via Moodle
https://moodle.nottingham.ac.uk/course/view.php?id=96914

Open-book examination.

Answer all THREE QUESTIONS

Marks available for sections of questions are shown in brackets in the right-hand margin.
This open-book examination will be marked out of 100.

Please directly work on the WORD document provided and write your answer right after each question and submit your work in typed electronic format as a **single PDF file, with each page in the correct orientation, to the appropriate dropbox on the module's Moodle page. Hand-written submission is NOT accepted.**

Use the standard naming convention for your document: [Student ID]_[Module Code]. Write your student ID number at the top of each page of your answers. Do not include your name.

Although you may use any notes or resources you wish to help you complete this open-book examination, the academic misconduct policy that applies to your coursework also applies here. You must be careful to avoid plagiarism, collusion or false authorship. Please familiarise yourself with the *Faculty of Science Statement on Academic Integrity*. This statement refers to, and does not replace, the University policy which stipulates severe penalties for academic misconduct. Please check the box indicated on Moodle to confirm that you have read this statement and that you understand it.

Staff are not permitted to answer assessment or teaching queries during the period in which your open-book examination is live. If you spot what you think may be an error on the exam paper, note this in your submission but answer the question as written.

The standard University of Nottingham penalty (5% deduction per working day) will apply to any late submission of this work.
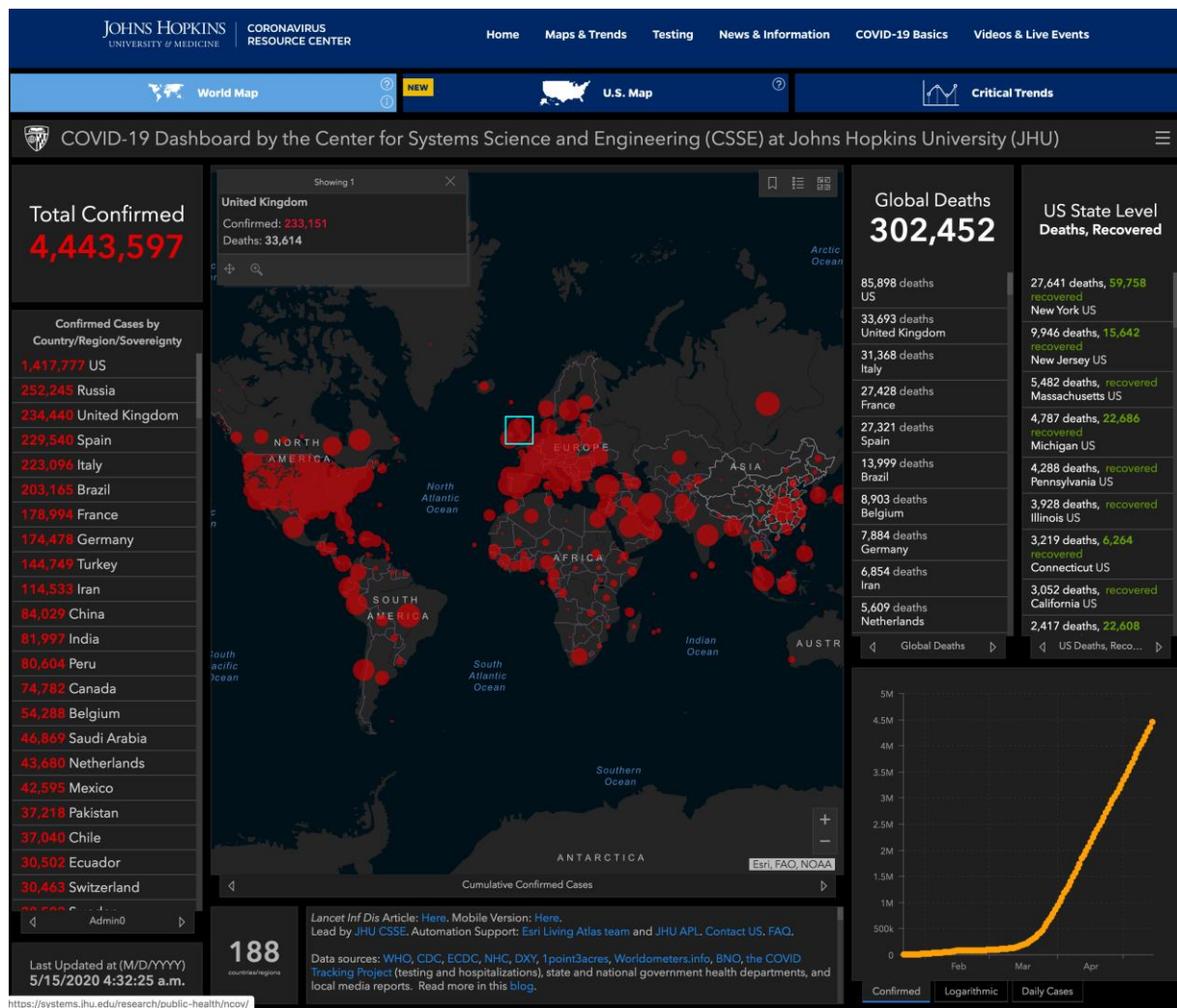
Question 1: Data, Visual Encodings and Interactions (Overall 28 Marks)

https://coronavirus.jhu.edu/map.html

Please check the above visualization carefully through the URL provided above and answer the following questions.

a. List and explain all the visual encodings and their corresponding data types in this visualization.

(6 Marks)

Each country or region in the visualization is defined by its position on the world map. Since each country or region is unique and there is no order, the data type is classed as nominal. Each category (with the default being cumulative confirmed cases) correlates to a size of circle in the visualization. This size represents a zero-fixed number in each case meaning that the data type is classed as quantitative ratio. Finally, the category of data in the visualization is represented by the color hue of circles. The colors are unique with no order making the data type nominal.

b. Given the 7 categories of interactions for information visualization based on user intent, explain all the interaction techniques used in this visualization.

(6 Marks)

The user is able to select a circle by clicking on it to see more detailed information about that specific region, with the previously unseen attributed then appearing in a new panel. The user is also able to explore the visualization by dragging it with their mouse cursor to examine a different subset of the data. This shows new regions of the world map, overcoming the limitation of the size display. Using the visualization's zoom feature the user can abstract/elaborate the data by only seeing specific regions of the map. This reduces the visual overlapping of the circles and makes comparing them easier.

c. Describe how "multiple coordinated views" and "focus and context" are incorporated in this visualization and the corresponding merits.

(6 Marks)

Multiple coordinated views are used to show where the cases are on the world map and the actual numbers for each country in a panel on the left-hand side. This is used to good effect, as it is hard to distinguish between the sizes of the circles on the map but would be a bad encoding choice to use the numbers directly

on the map. The visualization also generates new views when a circle is clicked, creating a new panel which shows the information of a region in more detail - which avoids over encoding the initial visualization.

Focus is used to allow the user to select a region by clicking it, to see greater detail about the given category. Surrounding circles can then be clicked to give context to a value, allowing for easy comparison of data points. Context is also given by the size of the surrounding circles when the map is zoomed in, allowing the user to quickly see how one region compares to another. The region focus finds merit because only countries are listed in the left-hand panel, as including all the regions would give too many data points which are scattered throughout the list. This would make the list harder to use, and the user would not be able to compare surrounding regions due to the list only being displayed in descending order.

d. Explain two problems you find in this visualization and how you would solve them.

(10 Marks)

When the map is completely zoomed out, there is a large amount of overlap between the circles, and much of the usefulness of their size is lost. This could be solved by grouping regions which overlap in this zoomed out state, allowing the user to see the values of a combined region. This would add more value to the visualization, with larger areas of the map being comparable to give further insight into the data.

The size of the circle is used to show the quantitative data on the map. This is ok, but it can be hard to distinguish between some of the medium and smaller sized circles from a distance. Furthermore, the ranges for the sizes are very large and two regions of the same size could have vastly different numbers. To solve this, I propose that a value of the color is also used to help distinguish between similar sized regions. A lighter colored would be used for a region with a small number of cases and a darker colored for a region with a larger number of cases. Where regions have the same size of circle, the one with the darker color can easily be distinguished as the one with a greater value.

Question 2: Evaluation (Overall **32 Marks)**
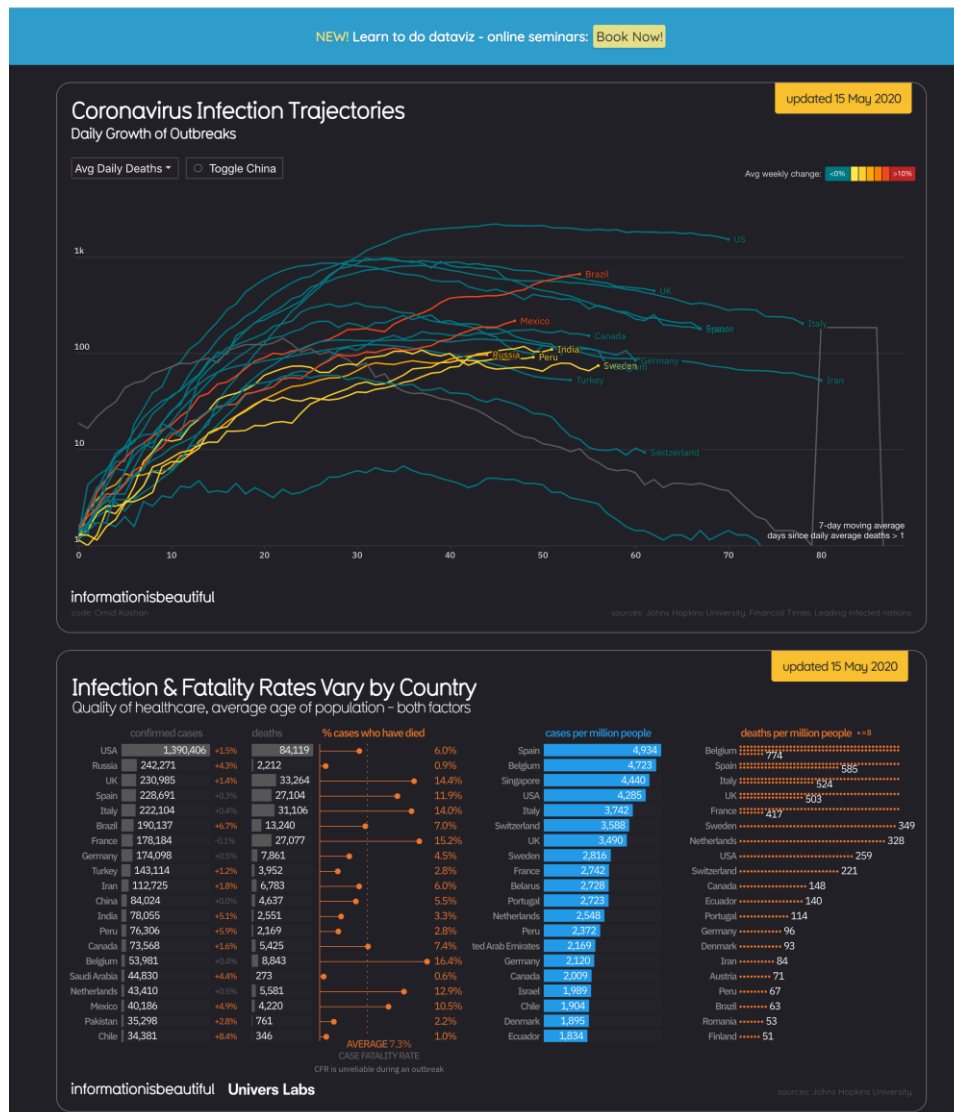
In addition to the visualization shown in Q1, **please check carefully the second visualization through the URL above and answer the following questions**.

a. Evaluating information visualization is important in order to understand whether the developed visualization systems are useful. Design an evaluation study that can enable evaluating both visualizations. Be specific about your evaluation goals, subjects, testing process, and how to record outcomes.

(20 Marks)

My study proposes a quantitative evaluation through a controlled experiment. The goal is to determine which evaluation presents the information the clearest manner, by asking test subjects to answer set questions and recording the speed and accuracy of their responses. There will also be questions at the end to evaluate each user's subjective ratings for each visualization. Subjects will be placed in groups of 3 and asked to answer 3 questions relating to information available in both the visualizations. The subjects should not have a large amount of prior knowledge about COVID-19 statistics, as this could bias the speed and accuracy of the questions. A presenter will demonstrate by answering an example question pertaining to similar data, to outline how the users should go about answering their questions. The 3 questions are as follows:

1. Which European country has the most confirmed cases?
2. Out of Mexico, Peru, and Brazil, which country has the highest number of active cases?
3. What is the case fatality ratio of Iran?

The countries in each question are in different parts of the world to prevent the users from becoming too familiar with finding information from a certain region, to make the study as fair as possible. There is also not a complete overlap of data between the visualizations, hence why certain countries are picked in some questions. The speed and accuracy of each question will be averaged for each visualization and so that they can be easily compared to determine which visualization was the most effective. After answering these questions, the participants will be asked which visualization they enjoyed using the most, and which they found to be the most helpful. This will allow us to compare the effectiveness of each visualization and whether the most effective was also the most enjoyable to use.

b. After exploring both visualizations, what have you learned about COVID-19 which you did not know previously? Why do you think visualization is important?

(12 Marks)

From the first visualization I liked how, where possible, each country's data was broken down into regions allowing me to see which were the most afflicted. This is not data I had previously seen or thought to look for, but through my exploration process of the world map I became intrigued and found it very useful to put

surrounding areas into context. This visualization also showed me the true global scale of the pandemic, which I normally do not follow as I mainly focus on what is going on in the United Kingdom (UK). For me this embodies the importance of visualization – to show people new information or perspectives they otherwise would not have seen.

In the second visualization it was interesting to see that the UK had such a high number of daily deaths in the global community, given that its population is much smaller than many other countries. The addition of the bar charts below the main line graph gave exact number for the deaths per million, which made it easy to see how the UK was fairing compared to its European neighbors. Using this information, we can see that other countries performed much better in handling the pandemic, which is important to know so that the UK can improve its capability to respond to a pandemic in the future.

**Question 3: Design and Text Visualization (Overall 40 Marks)**

Suppose you are monitoring the development of COVID-19 on a daily basis by reading news articles published online. However, soon you find that there are too many news articles published (over 100 daily) and it is too overwhelming to read all of them.

It would be great to have a visualization that can summarize those related news articles of the past (February, March and April) and help you understand the development. You can assume there is already a repository that stores all the news articles in the text format over the given time period you are interested in.

a. Design a text-based visualization that can achieve the goal of tracking the development of COVID-19 through published news articles on a daily basis. Describe your design in details and provide a corresponding drawing (either in electronic format or in scanned copy of your hand drawing) of your design.

(20 Marks)

I propose a word tree which uses all the collected articles in the repository to track the development of COVID-19. Each word from the repository will be counted, with the most frequent words being displayed in a larger font in the tree. These words will be ordered alphabetically from top to bottom. The words will then link to all the phrases that follow it, with each following word branching until a unique phrase is reached. An individual word or phrase can be clicked to remove other words above and below in the tree, to declutter the view. This will provide context to the words and allow the reader to see how different articles are using said words and the topic of the use. A special click (Ctrl + click) will also be used to allow the user the select the tree of a word in the current tree they are looking at. If the reader then wants to see the complete article, the unique phrase can be clicked to open the article, with the clicked phrase being highlighted in the text. This visualization would allow the reader to focus on one topic, for examples vaccines, and see all the information from multiple articles stemming from that one word. Assuming the repository is updated daily, the reader will be able to see any new additions surrounding a word (with new additions being highlighted in green) so that they can keep up to date with the latest information.
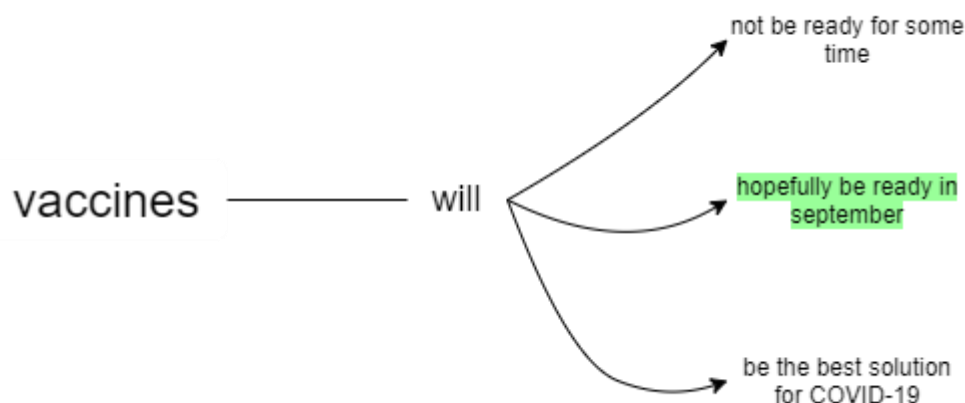
b. Given your design above, describe the procedure and techniques you would like to use to process the data (news articles in the textual format) to achieve your designed visualization. Be specific about the text processing techniques you will use. Note that the intention here is not to write specific codes to process this, but to provide an overview of all the procedures and techniques.

(20 Marks)

To process the data into a usable format, first all the text from each article needs to be broken down into sentences, with full stops being used to distinguish these. Then stop words at the start of sentences will be removed - if not, there will be too many branches coming from a stop word tree for the entire word tree to be useful. This is due to the size of the repository and the number of articles that will be stored in it, increasing daily. The root of each tree will be a noun, as I believe this will be the most useful way to see track information surrounding a topic. All the nouns from each article need to be counted so that their font size can be adjusted appropriately when the visualization is complete. All the individual sentences will need to be stored in a relational database, with each noun having its own table with entries showing the remainder of the sentence they originate from. The first word from each sentence remainder will be analyzed in a recursive manner, creating a new table for each word that branches one or more times, again storing the remainder of the sentence. This will be repeated until a word does not branch, meaning that the remaining phrase is unique, and this branch of the tree is complete. This final table will also store the link to the original article, which as described will be used to allow the user to see the original sentence in the article. At runtime, a stack will be used to store the user's exploration process, so that they can go back at any point to a previous word they were looking at and continue exploring from there. This is needed for the special click function where a completely new tree can be selected from the current one.