School of Computing and Information Technology

Student to complete:

Family name	Sharan shah
Other names	N/A
Student number	7081315
Table number	N/A

CSCI446/CSCI946
Big Data Analytics
South Western Sydney and Wollongong

Examination Paper
Spring 2021

Exam duration	2 hours + 15 minutes for the submission of Part B
Weighting	40%
Items permitted by examiner	*Open book, lecture notes, calculator*
Aids supplied	None

Directions to students

Part A: 10 multiple choice questions to be answered in Moodle.
Part B: 2 questions. Use this document as the template for answering the questions in part B. Your answers to Part B:

- must be typeset (i.e. using MS-word, LibreOffice, OpenOffice).
- must be inserted into this document in the relevant placeholders.

Save the completed document as a PDF file. One PDF file is to be submitted before due time via the corresponding submission link provided on Moodle.

Answer all questions in Part A and Part B.

This is part B of the exam. Insert your answers in the provided placeholders that are marked by the text "<Insert your answer here>". There is no word limit and no page limit. All answers must be typed text. Quotations, citations, and references are permitted. Images, drawings, diagrams, etc. are not permitted. All of your answers must be typeset. Your answer will be checked for plagiarism. Plagiarism of any part of your answer will result in zero marks for the whole of part-B of the exam. Do not include any material in your answer that has been copied from sources other than this exam paper.

Question B-1 (8 marks)

(a) Explain the following concepts in hypothesis testing: significance level, *p*-value, *t*-statistic, and confidence interval. **(4 marks)**

<Significance level: In statistics significance means something that has a probability if it has a high significance it means it has a high probability where as if the significance level is low it means the probability is less. It describes the probability of wrong elimination of null hypothesis when it is actually true it is stated as a probability of type-I error, significance helps us to decide if the null hypothesis is assumed to be accepted or rejected. Significance is denoted by Greek symbol **α** (alpha).

P-value: it is the probability of obtaining the result as extreme as the observed result of a hypothesis test assuming the null hypothesis is true.it is a measure of the probability that an observed difference can occur by random chance. The lower the p-value is the significance level is higher of the observed difference. The p-value can be used as an alternative or an addition to pre-selected confidence level for hypothesis testing. P-value is calculated using the deviation between the observed value and a selected reference value.

t-statistic: It is used when we want to decide if we want to support or reject a null hypothesis. It is used when there is a small sample size or when we don't know the population standard deviation. T statistic is used with the p value. The p value tells you what the odds are that the result could have occurred by random chance, example if there is a group of adults who consider professional cricket or are those who score by luck, so here finding the t statistic and p value will give us a good idea, getting this values will give us a evidence of significant difference between the teams mean and population mean which is everyone.

Confidence interval: It gives the range of probable values of the estimated parameter. For example, a 80% confidence interval gives the range of values of the actual parameter values with 80% confidence based on the sample data.>

(b) Describe the situation for which you will prefer to use Student's t-test, Welch's t-test, and Wilcoxon Rank-Sum test to conduct hypothesis testing, respectively. **(4 marks)**

<Student t-test: it is a method of testing hypothesis of mean of small sample drawn from normally distributed population when the standard deviation is unknown.it tells us how significant the differences can be between groups it helps to compare mean of two different population.

Example: Let's say I have a headache and I take a allopathic medicine and the headache lasts for a couple of days the next time I have a headache I try ayurvedic medicine and the headache lasts for a week, then I survey all my friends and ask which of the headache were shorter duration when they took the allopathic medicine or the ayurvedic medicine. What we really want to see is if the results are repeatable, the test helps us by comparing the means of two groups and tell us the probability of happening

Welch’s t-test: This test is designed for unequal population variance but the assumption of normality is maintained.

Wilcoxon Rank-Sum test: It is used to test if two populations have the same mean without making the assumption that both distributions have a normal distribution. Its like Welch’s test, the two distributions may have unequal standard deviations.>

Question B-2 (22 marks)

(a) A company would like to monitor what is being said about its products in social media. The company is interested in 1) whether people mention its products and 2) what is being said, good or bad. Describe your plan as data scientist for this task. **(7 marks)**

<Text analysis and sentiment analysis can be used to track product reviews on social media or in a corpus provided by the company.
Data collection and data cleaning:
I would begin by gathering the raw data provided by the company. Because the data is gathered from social media, it will be unstructured and contain mixed feelings.
We must normalise the text, which can be accomplished using tokenization. This means I'll have to separate the words from the sentence's body separately. I can accomplish these using approaches like case folding, lemmatization etc
Data modelling:
Then we'll apply the Bag of Words technique. This programme analyses the document and displays the words in a vector format based on their frequency in the corpus. This method determines the information content of a corpus, or the significance of a phrase. However, it is unable to analyse unstructured material that is constantly changing. As a result, we employ TFIDF. We then move on to Inverse Document Frequency, commonly known as Term Frequency (TFIDF).

It detects the use of words and can be updated on a regular basis at our discretion. It contains a graph-based scoring system that identifies a term and gives it a higher score if it appears more frequently in a document than in a corpus.
After that, we'll move on to subject modelling. It's a group of words with similar meanings that regularly appear in the same document. Each of these words has a certain amount of weight. It makes use of LDA (Latent Dirichlet Allocation). The documents were created through a generative approach.

Now we'll look into sentiment analysis. We can use this to figure out what the word's theme is. This method will be used to label the data. We'll need to train the data, classify it, and then test it on a sample.
We will see the outcomes by using word clouds or graphs to see how the product had worked overall.>

(b) Explain Graph Neural Networks. Also explain why graph modelling is increasingly important to Big Data Analytics. **(8 marks)**

<Graph neural networks are a type of ML algorithm that can extract important information from graphs and make good predictions, and because

graphs are becoming better and more detailed with time the GNN have become more powerful for many applications

Every graph is composed of nodes and edges, example for a food chain node can represent food and its product requirement, while edges can represent the relation between the combos the food chain is offering, if the graph is more complex it can also show the location of the chains the staff members etc.

Graph neural network can be created like any other neural network using fully connected layers. The number of layers depend on the type and complexity of graph data and the output we want.

The GNN receives a formatted graph data as a input and gives a vector of numeric values that represent the apt information, this representation of vector is also known as graph embedding these are used in ML to transform complex data into a structure so that it is easy to be separated and learned, example NLP system use word embedding to create a numerical representation of word and their relations.

when the graph data is passed to the GNN the feature of each node is combined with neighbouring nodes, if there are more than one layer the process is repeated in each layer and are connected accordingly, for example in a social network the first layer will connect the user to the friends and the next layer would add data from the friend's friend and like this it continues.

Few applications of GNN are:

- node classification
- edge prediction
- clustering >

(c) Explain Deep Convolutional Neural Networks and their role in Big Data Analytics. . **(7 marks)**

<A subset of neural networks are deep convolutional neural networks. It's a tool for analysing images. It's a tool for extracting features from an image. CNN is a network that broadcasts content in real time. They're utilised to see patterns in images like handwritten words, OCR, and drawings. It's a deep learning architecture with multiple stages. It has a filter for creating an image feature map, and it can have any number of filters. It use pooling as a means of reducing image size and thus pixels. The last layer's output is sent to a fully connected MLP. Which results in a result. For CNN, we normally utilise a pre-trained model because training from scratch takes a long time, Because they can reduce dimension while extracting features and providing output with the Fully connected layer, pre-trained models are widely employed in pooling layers. Because there is less noise, feature extraction is easier and more precise.>

- END OF EXAMINATION -