Practical Workbook Plan

BCA (6th Semester)

030010614: CC15 Essentials of Data and Text Processing

Course Objective: To familiarise the concept of data and text analysis, measurement levels and choose the relevant cleaning, transformation techniques to overcome data fallacies for effective representation, analysis for useful pattern identification.

Course Outcomes: Upon completion of the course, the student shall be able to

CO1: Describe the data types, quality measurement for data analytics.

CO2: Discuss importance and difference between Data Mining, and Data Science and Machine Learning for emerging domains.

CO3: Identify and use selected data acquisition techniques for data gathering through scraping data from specific data sources.

CO4: Summarise, describe and visualise data by utilising relevant data representation techniques.

CO5: Apply relevant data cleaning and transformation techniques to standardise the data for analytics along with dimension reduction.

CO6: Understand and apply approaches to text and document processing for statistical modelling and document summarization.

Programme Outcomes: The student will have

PO1: Ability to understand the concepts of key areas in computer science.

PO2: Ability to design and develop system, component or process as well as test and maintain it so as to provide promising solutions to industry and society.

PO3: Effective communication and presentation skill.

PO4: Ability to understand professional and ethical responsibility.

PO5: Recognition of the need for life-long learning.

Programme Outcomes and Course Outcomes mapping:

Course Outcomes	Programme Outcomes				
	PO1	PO2	PO3	PO4	PO5
CO1	√				
CO2	√				√
CO3	√	√			
CO4	√	√	√		
CO5	√	√			
CO6	√	√	√	√	√

Unit Number	Number of Questions	Time required to implement and debug the question (in hours)	Minimum required of workbook Certification	Submission Deadlines
1	2	9	2	4th Week of Semester
2	1	6	1	6th Week of Semester
3	3	12	3	8th Week of Semester
4	1	6	1	10th Week of Semester
5	1	6	1	12th Week of Semester
6	2	9	2	14th Week of Semester
TOTAL	10	48	10	-

Practical Definitions:

<table>
<tr><td>Practical No. 1</td><td>Enrollment No.</td></tr>
<tr><td>Practical Problem</td><td>A recent issue of Fortune Magazine reported that the following companies had the lowest sales per employee among the Fortune 500 companies.
<table>
<tr><th>Company</th><th>Sales per Employee in Dollar</th><th>Sales Rank</th></tr>
<tr><td>Seagate Technology</td><td>42.2</td><td>251</td></tr>
<tr><td>SSMC</td><td>42.19</td><td>414</td></tr>
<tr><td>Russell</td><td>41.99</td><td>410</td></tr>
<tr><td>Maxxam</td><td>41.99</td><td>459</td></tr>
<tr><td>Dibrell Brother</td><td>22.56</td><td>470</td></tr>
<tr><td>Ambi Care</td><td>52.43</td><td>513</td></tr>
</table>
1. How many elements are in the data set? Write down these elements.
2. How many variables are in the data set? Write down these variables.
3. How many observations are in the data set? Write down these observations.
4. Which of the above variables are qualitative and which are quantitative?</td></tr>
</table>

	5. Identify the scale of measurement used to store data in each of the above given variables and justify the same. 6. Write all steps to store this tabular data into CSV format. 7. Write the statement to read and view CSV file of above given data in R.
Objective(s)	To understand the types of data and identification of scale of measurement for each variable.
Pre-requisite	Basics of data and its characteristics
Duration for completion	5 hours
CO(s) to be achieved	CO1
Units mapped	1
Skill mapped	Analytical Skill , Technical Writing Skill
Nature of workbook submission	Handwritten practical solution with output
References for solving the problem	https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/scales-of-measurement/

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Understanding of Scale of Measurements				
Basic of R				
Signature & Date				

Practical No. 2	**Enrollment No:**
Practical Problem	Refer https://www.dataquest.io/blog/free-datasets-for-projects/ and download the dataset of your choice. Answer the following questions: 1. Summarize the dataset using R and write your interpretation about the dataset. 2. List the nominal, interval and ratio variable of this dataset, if any and justify the same. 3. List the ordinal variable of this dataset, if any and justify the same. 4. Download cancer dataset fromhttps://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 and solve the following questions. 5. Print first and third attribute from the data frame. 6. List observation that has malignant as target value. 7. Apply any one conditional query using subset. 8. Print the mean value of the area attribute.
Objective(s)	To apply subset operation on data frame and performing basic operation on data frame.
Pre-requisite	Scale of Measurement, Fundamental knowledge of Central Tendency

Duration for completion	4 hours
CO(s) to be achieved	CO1
Units mapped	1
Skill mapped	Technical Skill , Technical Writing Skill
Nature of workbook submission	Handwritten practical solution with output
References for solving the problem	

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Understanding of Scale of Measurements				
Operations in R				
Technical Skill				
Signature & Date				

Practical No. 3	Enrollment No:
Practical Problem	1. Create a sitemap to extract the web data from at least two websites. 2. Scrape the web data from multiple webpages using created sitemap. 3. Write all the necessary steps to complete Question 1 and 2. 4. Determine whether the scraped data containing missing values or not. 5. Display total number of missing values in scraped data. 6. Calculate missing values percentage in scraped data. 7. Remove observations which contains the missing values. 8. Remove variables which contains the missing values. 9. Fill the Quantitative variable using constant 1000 and also write the command to fill the missing values by its means. 10. Impute at least three values for missing data in all available categorical variable. 11. Fill the categorical missing values in dataset with first impute value.
Objective(s)	To experience data scraping and understand that how to handle with missing values in data set.
Pre-requisite	Basics of Data Scraping and Data Pre-processing

Duration for completion	6 hours
CO(s) to be achieved	CO1, CO3, CO4
Units mapped	1,2,4
Skill mapped	Technical Skill , Analysis Skill
Nature of workbook submission	Handwritten practical solution with output
References for solving the problem	Han, J. and Kamber, M. - Data Mining: Concepts & Techniques - Morgan Kaufmann Publishers

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Knowledge of Data Pre-processing				
Technical Knowledge				
Technical Writing				
Signature & Date				

Practical No. 4	Enrollment No:
Practical Problem	Consider the following dataset to solve the below given questions.

Brand	Model	Price	Consumer Ratings(1 refers to lowest and 5 refers to highest)
Sony	Ready Smart 75cm	35000	5
Sony	32LM Dark Iron 80cm	32000	5
Samsung	Wondertainment series 100 cm	54000	4
Toshiba	VW 80cm	41000	3
OnePlus	Y series 108cm	50000	3
Vu	Vu 108 72cm	36000	1
Mi	Mi TV 4 100 cm	490	1

- Apply data smoothing on the TV Price attribute using equal frequency bins, bin means and bin boundaries.
- Analyse the outlier if any after applying data smoothing techniques and write an interpretation for the same.

	• Perform data generalization on rating attribute with the category High, Average and Low ratings. • Normalize the price attribute in such a way that falls in the range of 0.0 to 1.0. and identify the noise from the value of price attribute and provide the interpretation from the result.
Objective(s)	Student shall be able to understand the data representation and data summarization by applying its various techniques.
Pre-requisite	Fundamentals of Central Tendency
Duration for completion	6 hours
CO(s) to be achieved	CO1, CO4
Units mapped	1, 4
Skill mapped	Technical Skill , Analysis Skill, Technical Writing Skill
Nature of workbook submission	Handwritten practical solution with output
References for solving the problem	Han, J. and Kamber, M. - Data Mining: Concepts & Techniques - Morgan Kaufmann Publishers

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Data Representation				
Data Summarisation				
Signature & Date				

Practical No. 5	Enrollment No:
Practical Problem	Consider the following dataset to solve the following questions.

Payment Method	Coupon Applied	Product Category	Region	Price	Units	Sales
Master Card	Yes	P2	East	$19.95	2	39.90
Master Card	Yes	P3	West	$22.95	1	22.95
Master Card	No	P4	East	$19.95	1	19.95
Master Card	No	P1	North	$22.95	5	114.75
Visa	No	P1	West	$22.95	1	22.95
Visa	No	P1	East	$19.95	3	59.85
Paypal	No	P1	South	$22.95	2	45.90
Paypal	No	P1	South	$22.95	1	22.95
American Express	Yes	P2	Mid-West	$19.95	1	19.95
American Express	Yes	P2	South	$22.95	1	22.95
Visa	Yes	P2	Mid-West	$19.95	2	39.90
Paypal	Yes	P3	South	$22.95	2	45.90

- Analyse the following data and discuss all the representation problems with this data.
- Suggest at least two techniques to solve the representation problems for the given data.
- Perform Cross-Tabulation by taking at least three attributes and write all the steps for the same.
- Create Pivot Table to represent the given data using Cross-Tabulation technique.
- Write all the steps to perform One-Hot Encoding techniques on at least two attributes form the given data.
- Perform One-Hot Encoding in R and write a snippet for the same.
- Analyse the result of both the methods, Cross- Tabulation and One-Hot Encoding and suggest the optimal technique which is best suited to represent the given data with proper justification.

Objective(s)	To understand the importance of data representation and to use the techniques for the same.
Pre-requisite	Basic R Operations, Excel Operations
Duration for completion	4 hours
PO(s) to be achieved	PO1 & PO2
CO(s) to be achieved	CO3
Units mapped	3
Skill mapped	Technical Skill , Technical Writing Skill, Analysis Skill
Nature of workbook submission	Practical Solution with Output
References for solving the problem	https://www.analyticsvidhya.com/blog/2015/04/comprehensive-guide-data-exploration-r/

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Dataset Understanding				
Technical Knowledge				
Signature & Date				

Practical No. 6	**Enrollment No:**
Practical Problem	(A) Use the same dataset given in Practical 5 to perform the following tasks. 1. Calculate Central Tendency Measurements on the attribute Sales and Price and write interpretation about it. 2. Calculate Standard Deviation, Variation and Range and write interpretation about it. 3. What is the impact of one or few values different from the others on the variance? Justify for both the attribute Sales and Price. 4. What will be the impact on variance if all price value multiply by 2? Justify the answer. (B) Find the mean, median and mode for the following collection of responses to the question: "How many online lectures have you attended this semester (260 days)?" 1, 1, 0,1, 2, 2, 0, 0, 0, 3, 3,0, 3, 3, 0,2, 2, 2, 1, 1,4, 1, 1,0,3, 0, 0, 0, 1, 1, 2, 2, 2, 2,1, 1, 1, 1, 4, 4, 4,1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,1, 1, 1, 1, 1, 3,3,0, 3, 3, 1, 1, 1, 1,0, 0, 1, 1, 1, 1, 3, 3, 3, 2, 3, 3, 1, 1, 1,2, 2, 2,4, 5, 5, 4, 4, 1, 1, 1, 4,1, 1, 1,3, 3, 5,3, 3, 3, 2,3, 3, 0, 0, 0, 0, 3, 3, 3, 3, 3, 3, 0, 2, 2, 2, 2, 1, 1, 1,3, 1, 0, 0, 0,1, 1, 3,1, 1, 1, 2, 2, 2, 4, 2, 2, 2, 1, 1, 1, 1,0, 0, 2, 2, 3, 3,2, 2, 3,2, 0, 0, 1, 1,3, 3, 3, 1, 1, 1, 1, 1,2, 2, 2, 2, 1, 1, 1, 1, 0,1, 1, 1, 3,1, 1, 1, 2, 2, 2, 1, 1, 1,2, 1, 1, 1,3, 3,5, 3, 3, 1, 1, 1, 3, 3, 3, 3, 1, 1, 1,4, 1, 1, 4, 4, 4, 4, 4, 4,1, 1, 1,2, 2,5, 5, 2, 3, 3, 4, 4,3,2, 2, 2, 1,5, 1,2, 2, 1, 1, 1, 2, 2, 2, 2, 2,1, 1, 0,1, 1, 1,3, 3, 3, 3, 3
Objective(s)	To understand Central Tendency and Scale of Variation in the data
Pre-requisite	Central Tendency, Frequency Distribution
Duration for completion	4 hours
PO(s) to be achieved	PO1, PO2

CO(s) to be achieved	CO3
Units mapped	3
Skill mapped	Statistical Skill
Nature of workbook submission	Practical Solution with Output
References for solving the problem	https://www.statisticshowto.com/probability-and-statistics/variance/

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Statistical Skill				
Result Interpretation Skill				
Technical Knowledge				
Signature & Date				

Practical No. 7	**Enrollment No:**
Practical Problem	Analyse the following data are the number of pages in 40 books on a shelf to solve the below given tasks. 136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512 1. Show the five number summary based on the given data and write the interpretation for each number. 2. Generate the IQR and construct the Box-plot graph. 3. Identify the represent the outlier on a Box-plot graph and discuss its impact on the given data. 4. Which other techniques are used rather than Box-plot to identify the outliers? 5. If we present Price and Rating data given in Practical 4 using Scatter Plot, what analysis will be observed? How it will be useful to find out the correlation between this two attributes? 6. Write the trend that you have observed in solution of Question 5.
Objective(s)	To understand Outliers identification and representation
Pre-requisite	Outliers characteristics
Duration for completion	3 hours
PO(s) to be achieved	PO1, PO2
CO(s) to be achieved	CO5
Units mapped	3
Skill mapped	Statistical Skill and Analysis Skill

Nature of workbook submission	Practical Solution with Output

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Statistical Skill				
Result Interpretation Skill				
Technical Knowledge				
Signature & Date				

Practical No. 8	**Enrollment No:**
Practical Problem	Refer the case study report to solve the following tasks. 1. Apply Spelling Error Correction Technique on any two words from your introduction part of the case study report. Write all the steps and identify the minimum step to correct the word. 2. Apply Tokenization on any two statements from the conclusion part of the case study report. 3. Discuss any two applications where Tokenization is highly used for text analysis. 4. Identify the stop words from the sentences and remove it from the sentence that you have used to solve the Q. 2 and write any two reasons for removing the stop words in Text Analysis process. 5. Write the Python code to solve the Q. 2 and Q. 4.
Objective(s)	To understand the basic operation for text processing
Pre-requisite	-
Duration for completion	4 hours
PO(s) to be achieved	PO1, PO2
CO(s) to be achieved	CO6
Units mapped	5
Skill mapped	Technical Writing Skill, Analysis Skill
Nature of workbook submission	Practical Solution with Output

Assessment				
Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Technical Writing Skill				
Technical Knowledge				
Signature & Date				

Practical No. 9	Enrollment No:
Practical Problem	Refer below given text to solve the following tasks. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. If confronted with the token saw, stemming might return just s, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun. The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma. Linguistic processing for stemming or lemmatization is often done by an additional plug-in component to the indexing process, and a number of such components exist, both commercial and open-source. The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is Porter's algorithm (Porter, 1980). The entire algorithm is too long and intricate to present here, but we will indicate its general nature. Porter's algorithm consists of 5 phases of word reductions applied sequentially. Within each phase, there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix. In the first phase, 1. Analyze the given text and write at least two reasons for processing this text. 2. Perform the sentence tokenization and write the Python code for the same. 3. Perform word tokenization and write the Python code for the same. Why Word Tokenization overweighs the usage of Sentence Tokenization in Text Analysis, justify the answer. 4. Identify the default stopwords in the English Language corpus and print it.

	5. Add "English", "Algorithm" and "crude" words to the default stopwords list. 6. Remove all the stopwords from the given text and print the all important tokens from the given text. 7. Identify all the libraries that used to perform the task Q. 2 to Q. 6 in Python with its purpose.
Objective(s)	To understand the basic operation for text processing
Pre-requisite	-
Duration for completion	4 hours
PO(s) to be achieved	PO1, PO2
CO(s) to be achieved	CO6
Units mapped	6
Skill mapped	Technical Writing Skill, Analysis Skill
Nature of workbook submission	Practical Solution with Output

Assessment

Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Technical Writing Skill				
Technical Knowledge				
Signature & Date				

Practical No. 10	Enrollment No:
Practical Problem	Refer below given document's text to solve the following tasks. •D1: the best Italian restaurant enjoy the best pasta •D2: american restaurant enjoy the best hamburger 1. Prepare the Term Document Matrix table for both documents. 2. Prepare a Vector Space Model and represent the Term Document Matrix table as Document as Vector and Words as Vector. 3. Prepare an Inverted Index for the tokens from documents. 4. Find Term Frequency and Inverse Document Frequency and Calculate Term Weights and interpret the results. 5. Solve Que. 4 in using Python code. 6. Create Bag of Words to find out the similarity between both documents using Cosine similarity.
Objective(s)	To understand Term Frequency, Term Weights and Document Similarity in NLP
Pre-requisite	-
Duration for completion	4 hours
PO(s) to be achieved	PO1, PO2
CO(s) to be achieved	CO6
Units mapped	6
Skill mapped	Technical Writing Skill, Analysis Skill
Nature of workbook submission	Practical Solution with Output

Assessment				
Parameter	**4- Mastery**	**3- Apprentice**	**2- Developing**	**1- Beginning**
Technical Writing Skill				
Technical Knowledge				
Signature & Date				