Roll No												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

School Of Computer Science and Engineering & Information Science

Summer term End-Term Examinations, August 2024

Odd Semester: 2023 - 24

Course Code: CSE 3134

Course Name: Text Mining and Text Analytics

Department:

Date: 05-08-2024

Time: 1:00pm-4:00pm

Max Marks: 100

Weightage: 50%

Instructions:

(i) Read the all questions carefully and answer accordingly.

(ii) Do not write any matter on the question paper other than roll number.

<table>
<tr><th>Q.No</th><th>Questions</th><th>Marks</th><th>CO</th><th>RBT</th></tr>
<tr><td rowspan="3">1</td><td>a. Explain F-score evaluation</td><td>4</td><td>CO1</td><td>L1</td></tr>
<tr><td>b. <b>Textual data and analysis can derive new insights and bring valuable business insights. These insights can be further leveraged by making better future business decisions. Sources that are used for text analysis in financial industry vary from internal word documents, email to external sources like social media, websites or open data. The system described in this paper will utilize data from social media (Twitter) and tweets related to Italian banks, in Italian</b>".
Explain how topic extraction model will help to gather valuable information.</td><td>6</td><td>CO1</td><td>L2</td></tr>
<tr><td>c. A group of 50 college students are given a self-administered questionnaire and asked how often they have used recreational drugs in the past year: Often (more than 5 times), Seldom (1 to 4 times), and Never (0 times). On another occasion, the same group of students was asked the same question in an interview. The following table shows their responses. Determine how closely their answers agree.
<table>
<tr><th></th><th colspan="4">Questionnaire</th></tr>
<tr><th>Interview</th><th>Seldom</th><th>Often</th><th>Never</th><th>Total</th></tr>
<tr><td><b>Seldom</b></td><td>20</td><td>8</td><td>2</td><td></td></tr>
<tr><td><b>Often</b></td><td>12</td><td>32</td><td>4</td><td></td></tr>
<tr><td><b>Never</b></td><td>0</td><td>6</td><td>16</td><td></td></tr>
<tr><td><b>Total</b></td><td></td><td></td><td></td><td></td></tr>
</table>
Determine how closely their answers agree.</td><td>10</td><td>CO1</td><td>L3</td></tr>
</table>

OR

<table>
<tr><td rowspan="3">2</td><td>a. Explain Generative Probabilistic Models</td><td>4</td><td>CO1</td><td>L1</td></tr>
<tr><td>b. "<b>Textual data and analysis can derive new insights and bring valuable business insights. These insights can be further leveraged by making better future business decisions. Sources that are used for text analysis in financial industry vary from internal word documents, email to external sources like social media, websites or open data. The system described in this paper will utilize data from social media (Twitter) and tweets related to Italian banks, in Italian. This system is based on open source tools (R language) and topic extraction model was created to gather valuable information. This paper describes methods used for data ingestion, modelling, visualizations of results and insights.</b>"
Explain how text analytics is one of the most effective methods to analyse the accident reports in universities.</td><td></td><td>CO1</td><td>L2</td></tr>
<tr><td>c. Factory "ABC" produces very expensive and high quality chip rings that their qualities are measured in term of curvature and diameter. Resutl of quality control by experts is given in the table below:
<table>
<tr><th>Curative</th><th>Diameter</th><th></th></tr>
<tr><td>2.95</td><td>6.63</td><td>Passed</td></tr>
<tr><td>2.53</td><td>7.79</td><td>Passed</td></tr>
<tr><td>3.57</td><td>5.65</td><td>Passed</td></tr>
<tr><td>3.16</td><td>5.47</td><td>Passed</td></tr>
<tr><td>2.58</td><td>4.46</td><td>Not Passed</td></tr>
<tr><td>2.16</td><td>6.22</td><td>Not Passed</td></tr>
<tr><td>3.27</td><td>3.52</td><td>Not Passed</td></tr>
</table>
As a consultant to the factory, you get a task to set up the criteria for automatic quality control. Then, the manager of the factory also wants to test your criteria upon new type of chip rings that even the human experts are argued to each other. The new chip rings have curvature 2.81 and diameter 5.46.

Explain employing the Discriminant Analysis to solve the problem.</td><td>10</td><td>CO1</td><td>L3</td></tr>
</table>

<table>
<tr><td rowspan="3">3</td><td>a. Explain the probability of observing a word with an example</td><td>4</td><td>CO2</td><td>L1</td></tr>
<tr><td>b. "<b>Patent application is one of the important ways to protect innovation achievements that have great commercial value for enterprises; it is the initial step for enterprises to set the business development track, as well as a powerful means to protect their core competitiveness. The emergence of a large amount of patent data makes the effective detection of patent data difficult, and patent infringement cases occur frequently. Manual measurement in patent detection is slow, costly, and subjective, and can only play an auxiliary role in measuring the validity of patents. Protecting the inventive achievements of patent holders and realizing more accurate and effective patent detection were the issues explored by academics.</b>"
Explain a method to fuse the similarity of patent text and image.</td><td>6</td><td>CO2</td><td>L2</td></tr>
<tr><td>c. Let C1 and C2 be two coins.
Ø1 be the probability of getting head with C1.
Ø2 be the probability of getting head with C2.</td><td>10</td><td>CO2</td><td>L3</td></tr>
</table>

Chosing any of the coin randomly, toss for 5 times.
Each selected coin has to toss for 10 mins.

C2	H	T	T	T	H	H	T	H	T	H
C1	H	H	H	H	T	H	H	H	H	H
C1	H	T	H	H	H	H	H	T	H	H
C2	H	T	H	T	T	T	H	H	T	T
C1	T	H	H	H	T	H	H	H	T	H

Find value of Ø1 and Ø2 by tossing C1 and C2 for 10 times by assuming the probabilities Ø1 = 0.6 and Ø2 = 0.5

OR

4

a. Explain the idea of Mixture Model. 4 CO2 L1

b. **The Volume of text resources have been increasing in digital libraries and internet. Organizing these text documents has become a practical need. For organizing great number of objects into small or minimum number of coherent groups automatically, Clustering technique is used. These documents are widely used for information retrieval and Natural Language processing tasks. Different Clustering algorithms require a metric for quantifying how dissimilar two given documents are. This difference is often measured by similarity measure such as Euclidean distance, Cosine similarity etc.**"
Explain the similarity measure process in text mining can be used to identify the suitable clustering algorithm for a specific problem. 6 CO2 L2

c. Explain any 5 Generative Probabilistic Models 10 CO2 L3

5

a. Explain Syntagmatic Relation with an example. 4 CO3 L1

b. "**The Probabilistic Latent Semantic Analysis has been related with the Singular Value Decomposition. Several problems occur when this comparative is done. Data class restrictions and the existence of several local optima mask the relation, being a formal analogy without any real significance. Moreover, the computational difficulty in terms of time and memory limits the technique applicability. The Nonnegative Matrix Factorization with the Kullback–Leibler divergence to prove, when the number of model components is enough and a limit condition is reached, that the Singular Value Decomposition and the Probabilistic Latent Semantic Analysis empirical distributions are arbitrary close. Under such conditions, the Nonnegative Matrix Factorization and the Probabilistic Latent Semantic Analysis equality is obtained.**"
Explain how the Singular Value Decomposition of every nonnegative entries matrix converges to the general case Probabilistic Latent Semantic Analysis results and constitutes the unique probabilistic image. 6 CO3 L2

c. **Research Lab "NanoTech" produces high-precision nanomaterials, with their qualities measured in terms of length and density. The results of quality control by experts are given in the table below:** 10 CO3 L3

Length	Density	Quality Control Result
0.56	2.13	Passed
0.42	2.45	Passed

0.63	1.98	Passed
0.59	1.91	Passed
0.35	1.76	Not Passed
0.48	1.89	Not Passed
0.57	1.67	Not Passed

As a consultant to the research lab, you are tasked with setting up the criteria for automatic quality control. Then, the head of the lab wants to test your criteria on a new type of nanomaterial that experts are undecided about. The new nanomaterial has a length of 0.49 and a density of 2.01.

Explain how you would employ Discriminant Analysis to solve the problem.

OR

6

a. Define word prediction with an example. 4 CO3 L1

b. **"In recent years, it has become common to analyze purchase history data and take advantage of the effect on business policies. An XYZ company is in the stage of introducing a membership system. Probabilistic Latent Semantic Analysis(PLSA) is well-known as an analytical model for analysis of co-occurenece of variables in data. However, the relationship between customers and items for the customer purchase behavior of each stage based on PLSA has not shown good performance. The purchase behaviors may be slightly different between customer stages, and accordingly, the purchase behavior of customers in different stages should be represented by a similar but different model. In addition, the higher the membership stage, the fewer customers there are; therefore, it becomes difficult to accurately understand the features of the customers's purchase behavior within high membership stages."**
Explain how PLSA model at a lower-stage, to estimate the parameters of models at a higher-stage to which few people belong. 6 CO3 L2

c. **Let D1 and D2 be two dice. Ø1 be the probability of rolling a 6 with D1. Ø2 be the probability of rolling a 6 with D2. Choosing any of the dice randomly, roll for 5 times. Each selected die has to roll for 10 mins.** 10 CO3 L3

Rol	**D 1**	**D 2**	**D 1**	**D 1**	**D 2**	**D 2**	**D 2**	**D 1**	**D 1**	**D 2**
1	6	2	6	4	5	3	6	1	6	3
2	3	6	2	6	1	4	2	3	4	6
3	6	3	6	5	6	6	4	6	2	2
4	4	5	5	6	2	5	3	2	3	4
5	2	6	6	3	6	1	5	6	6	5

Find the value of Ø1 and Ø2 by rolling D1 and D2 for 10 times, assuming the probabilities Ø1 = 0.2 and Ø2 = 0.1.

7

a. Describe the general problem of Text Mining 4 CO4 L1

b. "**Coincidental correctness occurs when the program happens to produce the correct output for some input even though it has executed a fault; the program is coincidentally correct rather than correct. One of the causes of coincidental correctness is known as Failed Error Propagation (FEP). FEP is known to hamper software testing, yet it remains poorly understood. FEP can occur for several reasons. For example, it might be that the faulty state is simply never inspected by the test oracle. In this case, the failure to propagate the error is caused by an inadequate oracle rather than by any inherent property of the program under test. Such failures of error propagation could be addressed by oracle improvement**".
Explain how conditional Entropy opens up the possibility in the longer term of devising inexpensive information theory-based metrics that allow us to minimize the effect of FEP. — 6 — CO4 — L2

c. **Medical Device Company "MediPrecision" manufactures high-accuracy medical sensors, with their qualities measured in terms of sensitivity and response time. The results of quality control by experts are given in the table below:**

Sensitivity	Response Time	Quality Control Result
0.85	1.30	Passed
0.78	1.45	Passed
0.92	1.10	Passed
0.87	1.20	Passed
0.70ß	1.60	Not Passed
0.75	1.55	Not Passed
0.82	1.80	Not Passed

As a consultant to the company, you are tasked with setting up the criteria for automatic quality control. Then, the manager of the company wants to test your criteria on a new type of medical sensor that experts are undecided about. The new sensor has a sensitivity of 0.80 and a response time of 1.35.

Explain how you would employ Discriminant Analysis to solve the problem. — 10 — CO4 — L3

OR

8

a. Describe the Landscape of Text Mining and Analytics — 4 — CO4 — L1

b. "**Data-driven soft sensors have been extensively studied in the process industry to provide an accurate online estimation of quality-related variables with easy-to-measure variables. For chemical processes with massive process variables, the performance of soft sensor models could be significantly improved by variable selection because part of these measurements is redundant or independent of quality-related variables. Generally, the variable selection is achieved by ranking process variables in order of their importance to the quality-related variables by correlation analysis. However, considering that correlation analysis methods are relative measures of variable dependence, the determination of the final variable set is quite subjective because there are several user-defined parameters.** " Explain how a conditional entropy-based feature selection method can help in overcome the limitation. — 6 — CO4 — L2

c. **Let B1 and B2 be two biased dice. Ø1 be the probability of rolling a 6 with B1. Ø2 be the probability of rolling a 6 with B2. Choosing any of the dice randomly, roll for 5 times. Each selected die has to roll for 10 rounds.** — 10 — CO4 — L3

Roll	B1	B2	B1	B1	B2	B2	B2	B1	B1	B2
1	6	1	4	6	5	3	6	6	2	4
2	3	6	5	2	2	4	3	3	4	6
3	6	2	6	6	6	6	2	5	3	5
4	1	3	3	4	1	2	5	6	6	6
5	5	6	6	5	3	5	4	1	6	2

Find the value of Ø1 and Ø2 by rolling B1 and B2 for 10 times, assuming the probabilities Ø1 = 0.25 and Ø2 = 0.15.

9

a. Describe the landscape of Text Mining and Analytics.. 4 CO5 L1

b. "**With an overwhelming amount of textual information in biology and biomedicine, the needs for automatically extracting information, eg, protein protein interaction, from biomedical documents increase.**" With the development of computational linguistic tools, large quantities of text can be analyzed efficiently concerning their syntactic structure. Some researchers have used phrases and multiple words rather than individual words as indexing terms. Moreover, to eliminate the synonym problem in natural language, researchers also tried to use word meanings or term clustering to represent text. Furthermore, to capture the semantic relationships between words ignored by using the bag-of-words representation, the researchers also included a hypernym-based representation under WordNet. Explain the different ways to represent the bio-medical text. 6 CO5 L2

c. **Medical Device Company "HeartGuard" manufactures high-accuracy heart rate monitors, with their qualities measured in terms of accuracy and response time. The results of quality control by experts are given in the table below:** 10 CO5 L3

Accuracy	Response Time	Quality Control Result
95%	1.2s	Passed
92%	1.5s	Passed
98%	1.0s	Passed
97%	1.1s	Passed
85%	1.8s	Not Passed
88%	1.6s	Not Passed
90%	1.7s	Not Passed

As a consultant to the company, you are tasked with setting up the criteria for automatic quality control. Then, the manager of the company wants to test your criteria on a new type of heart rate monitor that experts are undecided about. The new monitor has an accuracy of 91% and a response time of 1.4s.

Explain how you would employ Discriminant Analysis to solve the problem.

OR

<table>
<tr><td rowspan="3">10</td><td>a. List the features of Deep NLP</td><td>4</td><td>CO5</td><td>L1</td></tr>
<tr><td>b. "<b>In the last few years, several studies have been devoted to dissecting dense text representations to understand their effectiveness and further improve their quality. Particularly, the anisotropy of such representations has been observed, which means that the directions of the word vectors are not evenly distributed across the space but rather concentrated in a narrow cone. This has led to several attempts to counteract this phenomenon both on static and contextualized text representations. However, despite this effort, there is no established relationship between anisotropy and performance</b>."
Explain how the clustering task as a means of evaluating the ability of text representations to produce meaningful groups.</td><td>6</td><td>CO5</td><td>L2</td></tr>
<tr><td>c. <b>Let M1 and M2 be two machines producing bolts. Ø1 be the probability of producing a defective bolt with M1. Ø2 be the probability of producing a defective bolt with M2. Choosing any of the machines randomly, produce 5 bolts. Each selected machine has to produce bolts for 10 rounds.</b>
<table>
<tr><th>Round</th><th>M1</th><th>M2</th><th>M1</th><th>M1</th><th>M2</th><th>M2</th><th>M2</th><th>M1</th><th>M1</th><th>M2</th></tr>
<tr><td>1</td><td>N</td><td>D</td><td>N</td><td>N</td><td>N</td><td>D</td><td>N</td><td>N</td><td>D</td><td>D</td></tr>
<tr><td>2</td><td>N</td><td>N</td><td>N</td><td>D</td><td>N</td><td>N</td><td>D</td><td>D</td><td>N</td><td>N</td></tr>
<tr><td>3</td><td>D</td><td>N</td><td>N</td><td>N</td><td>D</td><td>D</td><td>N</td><td>N</td><td>N</td><td>N</td></tr>
<tr><td>4</td><td>N</td><td>D</td><td>N</td><td>D</td><td>N</td><td>N</td><td>D</td><td>N</td><td>N</td><td>D</td></tr>
<tr><td>5</td><td>N</td><td>N</td><td>N</td><td>N</td><td>N</td><td>D</td><td>N</td><td>N</td><td>N</td><td>N</td></tr>
</table>
<b>(N: Non-defective, D: Defective)</b>
<b>Find the value of Ø1 and Ø2 by observing the results of M1 and M2 for 10 rounds, assuming the probabilities Ø1 = 0.1 and Ø2 = 0.2.</b></td><td>10</td><td>CO5</td><td>L3</td></tr>
</table>