|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  Roll No |  |  |  |  |  |  |  |  |  |  |  |

PRESIDENCY UNIVERSITY BENGALURU

 SCHOOL OF ENGINEERING

 MAKE-UP EXAMINATION JULY 2024

|  |  |
| --- | --- |
| **Semester : VII** | **Date : 01 JULY 2024** |
| **Course Code : CSE3134** | **Time : 09.30am to 12.30pm** |
| **Course Name : Text Mining and Text Analytics** | **Max Marks : 100** |
| **Program : BTech(CSE)** | **Weightage : 50%** |

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Question paper consists of 3 parts.*
3. *Scientific and non-programmable calculator are permitted.*
4. *Do not write any information on the question paper other than Roll Number.*

|  |
| --- |
| **PART A** |
|  **ANSWER ANY 4 QUESTIONS 4Q X 5M=20M** |
| 1 | Explain F-score evaluation. | (CO 4) | [Knowledge] |
|  |
| 2 | Explain the probability of observing a word with an example | (CO 3) | [Knowledge] |
|  |
| 3 | Explain Syntagmatic Relation with an example. | (CO 2) | [Knowledge] |
|  |
| 4 | Describe the general problem of Text Mining | (CO 1) | [Knowledge] |
|  |
| 5 | Describe the Landscape of Text Mining and Analytics | (CO 1) | [Knowledge] |
|  |
|  |

|  |
| --- |
| **PART B** |
|  **ANSWER ANY 5 QUESTIONS 5Q X 10M=50M** |
| 6 | **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**". Explain how topic extraction model will help to gather valuable information.  | (CO3) | [Comprehension] |
|  |
| 7 | "**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.**." Explain a method to fuse the similarity of patent text and image.  | (CO4) | [Comprehension] |
|  |
| 8 | **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.  | (CO4) | [Comprehension] |
|  |
| 9 | "**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 | (CO3) | [Comprehension] |
|  |
| 10 | "**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. | (CO2) | [Comprehension] |
|  |
| 11 | "**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**." Explain how the clustering task as a means of evaluating the ability of text representations to produce meaningful groups. | (CO1) | [Comprehension] |
|  |  |  |  |
|  |

|  |
| --- |
| **PART C** |
|  **ANSWER ANY 2 QUESTIONS 2Q X 15M=30M** |
| 12 | 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.

|  |  |
| --- | --- |
|   | **Questionnaire** |
| **Interview** | **Seldom** | **Often** | **Never** | **Total** |
| **Seldom** | 20 | 8 | 2 |   |
| **Often** | 12 | 32 | 4 |   |
| **Never** | 0 | 6 | 16 |   |
| **Total** |   |   |   |   |

Determine how closely their answers agree. | (CO2) | [Application] |
|  |
| 13 | Explain any 5 Generative Probailistic Models. | (CO3 & CO4) | [Application] |
|  |
| 14 | 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:

|  |  |  |
| --- | --- | --- |
| Curative | Diameter | Quality Control Result |
| 2.95 | 6.63 | Passed |
| 2.53 | 7.79 | Passed |
| 3.57 | 5.65 | Passed |
| 3.16 | 5.47 | Passed |
| 2.58 | 4.46 | Not Passed |
| 2.16 | 6.22 | Not Passed |
| 3.27 | 3.52 | Not Passed |

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. | (CO2) | [Application] |
|  |
|  |