**REVA Academy for Corporate Excellence**

PGDM/MBA in Business Analytics 06 – Batch 2019-20

**Trimester 02 Examination**

|  |  |  |  |
| --- | --- | --- | --- |
| **Module Name** | **Text Analytics | Module 4** | | |
| **Module Code** | PGDBA19P210/MBA19P210 | | |
| **Date** | 25th April, 2020 | **Time** | 180 Min |
| **Exam Type** | Open Book | **Marks** | 100 |
| **Instructions** | * All questions are compulsory. * Submit the final answer files in appropriate format in a zipped folder. * You may also submit the hand written document scanned. * Data sets are attached for Question 2 and Question 3 | | |

**Section 1**

1. Explain your approach to create an analytics solution for understanding the Twitter users with a good data pipeline. (Use the Tweets.csv). Assume the necessary details and state the assumptions. (This is a descriptive question which can be answered in a .doc format or hand written). **(1x30 = 30 Marks)**

.

1. Prepare the given Tweet data set (Tweets.csv). The notebook must be well commented. **(1x30 = 30Marks)**
   1. Remove stop words
   2. Remove words less than 3 characters
   3. Remove hyperlinks
   4. Explore the given data set for insights through
      1. Word frequency
      2. Word cloud

****

1. Write a Python/R program to compare the lexicon-based and naive Bayesian approach to the given data set (AFINN.csv) for text classification. Explain the outcomes of the results with good inference(s).  (**1x40 = 40Marks)**



\*\*\*\*\*\*\*End\*\*\*\*\*\*\*