Reg. No.										

MANIPAL ACADEMY OF HIGHER EDUCATION

THIRD TERM POST GRADUATE DIPLOMA IN DATA SCIENCE (FULL TIME) DEGREE EXAMINATION – SEPTEMBER 2018

SUBJECT: DSC 417.1 – UNSTRUCTURED DATA ANALYSIS

Friday, September 28, 2018

Time: 09:30 – 12:30 Hrs. Max. Marks: 100

✍ **Answer ALL the questions:**

1A. While building corpus how to convert all the words to lowercase

i) docs = tm_map(docs, tolower)

ii) docs = tm_map(docs, content_transformer(tolower))

iii) docs = tm_map(docs, content_transformer(tolower()))

iv) docs = tm_map(docs, content_transformer(toLower))

1B. Which of the following is used to calculate inverse document frequency of a term?

i) log(No. of documents in which the term is appearing/No. of documents in the corpus)

ii) No. of documents in the corpus/No. of documents in which the term is appearing

iii) log(No. of documents in the corpus/No. of documents in which the term is appearing)

iv) log(No. of documents in the corups/Frequency of the term in all the documents)

1C. Assume the following data. What is the probability P(game | Sports)

Text	**Category**
A great game	Sports
The election was over	Not sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not sports

i) 2/14 ii) 4/14 iii) 2/11 iv) 2/5

1D. Which of the following is true about sparsity?

i) Percentage of zeros against total number of values in Document Term Matrix

ii) Percentage of non-zeros in Document Term Matrix

iii) Percentage of zero with non-zero values in Document Term Matrix

iv) None of the above

1E. Extracting word cloud in text mining is an example of

i) Unsupervised Learning

ii) Network Analysis

iii) Classification

iv) Supervised Learning

1F. Which of the following is used to get the terms total frequency in a term document matrix

i) Row sums

ii) Column sums

iii) Column sums divided by total number of documents

iv) Row sums divided by total number of documents

1G. What is the command to restrict the words in a word cloud based on their count of occurrences?

i) wordcloud(words_counts$words, words_counts$Freq, .freq = 50)

ii) wordcloud(words_counts$words, words_counts$Freq, min.freq = 50)

iii) wordcloud(words_counts$words, words_counts$Freq, count = 50)

iv) wordcloud(words_counts$words, words_counts$Freq, min.count = 50)

1H. Which of the following is used for document clustering

i) Term Document Matrix

ii) Document Term Matrix

iii) Bag of words

iv) Inverse document frequency

1I. In two documents, the word "the" occurs ten times together across both documents. The word "an" occurs in one document but occurs ten times.

Which of the following is well suited for this scenario.

i) Both have same TFIDF

ii) TFIDF of the is higher

iii) TFIDF of an is higher

iv) Cannot compute TFIDF

1J. In LexRank algorithm, the edges between the sentences are based on

i) cosine similarity

ii) Correlation of words

iii) Frequency of words

iv) None of the above

(2 marks × 10 = 20 marks)

2A. Analyze the challenges in unstructured data analysis?

2B. In the following table, the text column contains the tweets from a hashtag #iphone. How does the Term Document Matrix and Document Term Matrix looks for this?

Screenname	**Text**	**Created_at**
Sam	Waiting for the new #iphone release	23/06/2017
Madhan	New #iphone technical specifications	22/06/2017
Veera	#iphone 6 prices slashed. Grab yours soon	21/06/2017
Sundeep	Any idea about new #iphone launch date in India? I heard it is on 22/07/2017. Not sure though	22/06/2017

2C. Write an R code to compute top 10 most frequent words from a corpus?

2D. Explain the need for sentiment analysis? Quote at least two practical applications of it in mobile manufacturing companies?

2E. Write MongoDB shell commands for the following

i) Switch to a new database “manipal”

ii) Show to all collections in database “manipal”

iii) Count all the documents in the collection “tweets”

iv) Assume each document in “tweets” collection has a field called words. The value of the field “words” is array of words in that particular tweet. Write a query to filter those documents which has the hashtag “#datascience”

2F. How to remove custom and common stop words from corpus. Explain with one example.

2G. Explain Vector Representations of Words with an example.

2H. Define polarity in sentiment analysis? Explain with one example

2I. How can we classify data assets in organizations? Provide at least two example for each type of asset

2J. Explain the importance of TF-IDF transformation with an Example.

(4 marks × 10 = 40 marks)

3A. The following table contains the subject of different emails. The category columns denotes whether it is a sports category or not. Using naiye baye’s algorithm compute the necessary probabilities to classify the new sentence “A very close game”, whether it belong to category sports or not? Explain the process and calculations involved in detail.

Text	**Category**
A great game	Sports
The election was over	Not sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not sports

3B. A leading multiplex wanted to use social media as reference to decide whether to screen a particular movie or not based on people reaction to the movie’s song release, promo, teaser etc.

i) Explain the steps involved in scrapping tweets from Twitter API using R for the hashtag “#pari” and storing the same in MongoDB using R. Provide sample R code

ii) Explain the steps involved in building sentiment analysis based on polarities along with R code

3C. Explain the steps involved in text cleaning along with R – Code. Also explain atleast 3 challenges involved in text cleaning.

3D. Explain the following terms

i) Stemming
ii) TF-IDF
iii) Sentiment polarity
iv) Collections in MongodB

(10 marks ×4 = 40 marks)