**Section1: Time allowed is 2 hours. Total points for this section is 100.**

1. (a) What is Big Data? Describe at least 4 major differences between Big Data and Traditional Data. 10pts  
     
   Big Data is huge, versatile, various set of data that can’t be manipulated/analyzed using traditional data tools.  
     
   Differences between Big Data and Traditional Data:

|  |  |  |
| --- | --- | --- |
|  | **Traditional Data** | **Big Data** |
| **Structure** | Usually structured | Might or might not be structured data |
| **Size** | Medium/Small GB/Terabytes | Huge - Petabytes |
| **Cost** | Low/Medium cost | High |
| **Database Architecture** | Relational RDBMS SQL | Non-relational NoSQL |
| **Query Language** | Usually SQL | Can be SQL, sql-like or other syntax |

(b) Create a personally meaningful SCI picture of Big Data. What SCI principles are represented in the picture. 10 pts (check the paper drawing please)

* 1. (a) What is Text Mining? What is Web Mining? Compare and contrast them, listing at least 2 major commonalities and 2 differences. (10 pts)  
       
     **Text Mining**: the art and science of extracting useful patterns from text/text documents.  
     **Web Mining**: the art and science of extracting patterns between world wide web pages and find measures like their relation to each other in order to improve the www or user experience or increase income of the businesses  
     Commonalities:
* Both are data mining techniques, both use big data tools
* Both will need to analyze text content and classify document/pages under specific topics.  
    
  Differences
* Text mining is about finding relationship between words on text documents; Web mining is about finding relationship between different web pages which might contain vast type of contents like (text, video, images, zip archives, programs) and/or studying/analyzing user engagement with those web pages.
* Text mining relies on TDM (terms document matrix) to find most common words in a set of document, Web Mining use different technique based on the goal of the mining, one technique is calculating PageRank based on number of outgoing links.

(b) What is Hadoop? What is Spark? Compare and contrast them, listing at least 2 major commonalities and 2 differences. (10 pts)  
- Hadoop is a set of tools for data mining and processing of big data in batches, the most common part of hadoop are: HDFS their distributed file system and MapReduce (their implementation of MapRedue algorithm)  
- Spark is a set of tools for data mining processing of big data in streaming or batches. It’s 10x~100x faster than hadoop due to usage of in-memory objects (RDDs), it’s commonly used to process streaming data.

Commonalities:

* Both are used for Big Data mining/processing
* Both are free and open source
* Both implements MapReduce algorithms

Differences:

* Hadoop does distributed file systems, Spark doesn’t implement that
* Hadoop is used for batch processing, doesn’t support streaming of data, Spark supports both and excels at streams.
* Spark has an implementation of machine learning algorithm, Hadoop doesn’t

1. A private university in the US wants to grow.They have accumulated a lot of data about students, faculty, departments, degrees, courses, grades, job placements, tuition, loans, alumni, marketing campaigns, revenue, costs, local competition, etc. for many decades. They want to use their data to generate new insights to grow and improve themselves.
2. Describe 3 important business objectives that they should set for data analytics. What data mining method(s) would be applicable for each of those objectives? (10 pts)
   1. **Find ways to retain existing students, use of big data regression and decision trees, to identify main reason for students to quit school and then prevent this.**
   2. **Get new student, use clustering technique and ANNs to see which category of students are interested more in studying majors that this university teaches, target those clusters with targeted ads.**
   3. **Qualify/Disqualify scholarship loans for students based on decision tree outcome.**
3. What is IBM Watson? How could it help this university? (10 pts).  
   watson is a machine learning/text mining/natural language processing tool by IBM which aims to study/analyze large number of documents and learns to make decisions based on that dataset.
4. (a) Write the R script for computing Pagerank for the network below. List nodes in order of their pageranks (10 pts).

|  |  |
| --- | --- |
| Page | Pagerank |
| 3 | 0.2902176 |
| 2 | 0.2429125 |
| 4 | 0.2139663 |
| 1 | 0.1264518 |
| 5 | 0.1264518 |

code:

library(igraph)

myGraph <- graph(

c(

"1","1", "1","2", "1","5",

"2","3",

"3","2", "3","4",

"4","4", "4","5", "4", "1",

"5","2", "5","3"

), directed = TRUE)

plot(myGraph)

sort(page.rank(myGraph)$vector, decreasing = T)

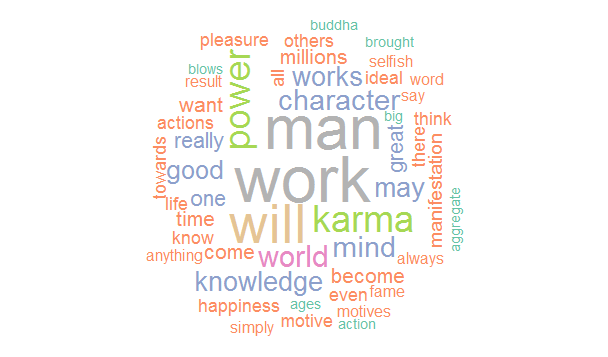
(b) For the network below, compute the PageRanks using Excel. List all nodes and their Pageranks. (copy the matrix and 5 iterations from excel into the word document) (10 pts)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **R1** | **R2** | **R3** | **R4** | **R5** | **init val** | **iteration 1** | **iteration 2** | **iteration 3** | **iteration 4** | **iteration 5** |
| **R1** | 0.33 | 0.00 | 0.50 | 0.33 | 0.00 | 0.20 | 0.25 | 0.32 | 0.27 | 0.25 | 0.27 |
| **R2** | 0.33 | 0.00 | 0.00 | 0.00 | 0.50 | 0.30 | 0.12 | 0.13 | 0.18 | 0.18 | 0.15 |
| **R3** | 0.00 | 1.00 | 0.00 | 0.00 | 0.50 | 0.30 | 0.35 | 0.17 | 0.21 | 0.27 | 0.25 |
| **R4** | 0.00 | 0.00 | 0.50 | 0.33 | 0.00 | 0.10 | 0.18 | 0.24 | 0.16 | 0.16 | 0.19 |
| **R5** | 0.33 | 0.00 | 0.00 | 0.33 | 0.00 | 0.10 | 0.10 | 0.14 | 0.19 | 0.14 | 0.13 |
|  | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** |

.

5)  
 (a) Create using R a wordcloud for the attached document. Clean up the data, removing all filler words. (10 pts)

code:  
library(tm)



library(wordcloud)

library(RColorBrewer)

articleCorpus <- readLines("karma.txt")

mycorpus = Corpus(VectorSource(articleCorpus))

# Transformation

mycorpus1 = tm\_map(mycorpus, removePunctuation)

mycorpus2 = tm\_map(mycorpus1, removeWords, stopwords("english"))

mycorpus3 = tm\_map(mycorpus2, tolower)

mycorpus4 = tm\_map(mycorpus3, stripWhitespace)

mycorpus5 = tm\_map(mycorpus4, PlainTextDocument)

mycorpus5 = tm\_map(mycorpus5, removeWords, c("the","of","which","but","without","every","like","among","let","can","called","just"))

tdm <- TermDocumentMatrix(mycorpus5)

m1 <- as.matrix(tdm)

v1<- sort(rowSums(m1), decreasing=TRUE)

d1 <- data.frame(word=names(v1), freq=v1)

d1 <- head(d1,50)

wordcloud(d1$word,d1$freq,col=brewer.pal(8,"Set2"), min.freq="5",random.order=FALSE)

(b) Show the ten most frequent words in the word list. Show the wordcloud of top 50 words. (10 pts).   
  
**10 most frequent words:**

d1.word d1.freq

1 work 35

2 man 33

3 will 29

4 karma 18

5 power 18

6 world 14

7 character 13

8 mind 13

9 knowledge 12

10 may 12  
  
**code**: same as above with this modification:

d1 <- head(d1,10)  
 data.frame(d1$word, d1$freq)