Assignment - 7

1702001070 44

Q-1 Define big data and big data analytics. Explain key roles and their responsibilities for successful analytic project.

* Big Data :-

- Big data is a term used for any data that is large in quantity.

- It is used to refer to any kind of data that is difficult to be present using conventional methods like DBMS or MS Excel.

- Big data challanges include capturing data, data storage, data analysis, search, sharing, transfer, visualization of data, etc.

- Big data is an evolving term that describes any large amount of structured, semistructured and unstructured data that has the potential to be mined for information.

* Big Data Analytics :-

- Big Data Analysis is the often complex process of examining big data to uncover information.

- Such as hidden patterns, co-relations, market bonds and customer preferences.

- That can help organization make information business decisions.

- Big data analytics is a form of advanced analytics which involves complex applications with elements such as predictive models, statistical algorithm and what if analysis powered by analytics systems.

* Responsibilities of Big data analysis :-

- Big Data Analytics helps organizations harness their data and use it to identify new opportunities.

1) Cost Reduction :-
- Big Data technologies such as cloud based analytics bring significant cost advantages when it comes to storing large amount of data.

2) Faster and Better decision making :-
- With the speed of Hadoop on in-memory analytics combined with the ability to analyze new source of data.

3) New products and services :-
- With big data analytics, more companies are creating new products to meet customer's needs.

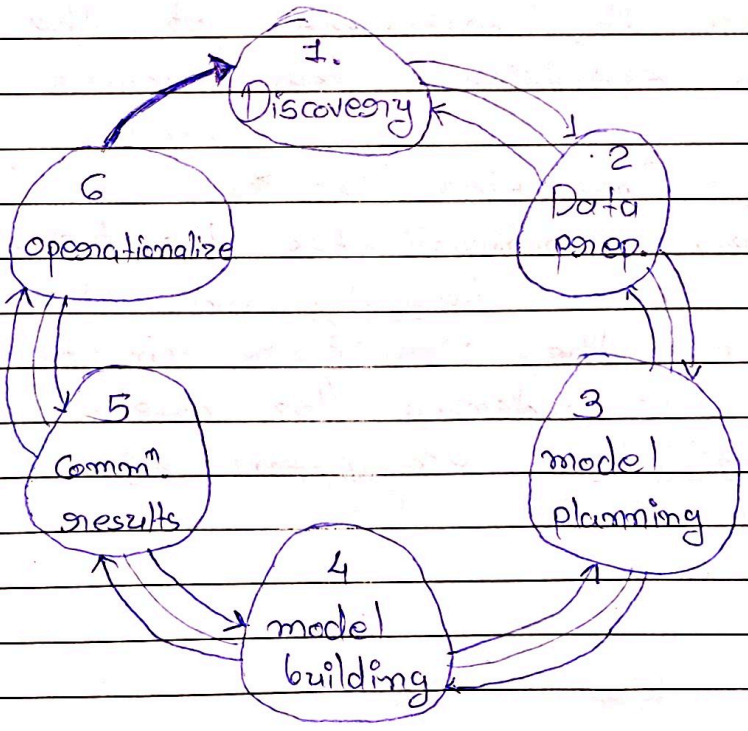
170200107044

Q-2 Briefly explain the life-cycle Data analytics and discuss the role of Data scientists.

* Data Analytics lifecycle :-
- To address the distinct requirements for performing analytics on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and representing data.

☆ Phase-1 : Discovery

- The data science team must learn and investigate the problem.
- Develop context and understanding.
- The team formulates initial hypotheses that can later be tested with data.



vision

★ Phase-2 : Data preparation
- It includes steps to explore, preprocess and condition data prior to modeling and analysis.

★ Phase-3 : Model planning
- In this phase, the team determines the methods, techniques and work flow it intends to follow for the subsequent model building phase.

★ Phase-4 : Model Building
- Data science team needs to develop data sets for training, testing and production purpose.
- These data sets enable the data scientist it, while holding aside some of the data for testing the model.

★ Phase-5 : Communicate Results
- After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure.

★ Phase-6 : Operationalize
- The team communicates the benefits of the project more broadly and sets up a pilot project to deploy the words in a controlled way before broadening the work to a full enterprise on ecosystem of users.

1702001070 44

* Data Scientist role :-
- A data scientist's main objective is to organise and analyze large amounts of data, often using specifically designed software.
- The final result of a data scientist's analysis needs to be easy enough for all stakholders to understand.
- Data scientists must have enough business domain expertise to translate company or databased deliverables such as prediction engines, pattern detection analysis, optimization algorithms and the like.

Q-3 How Data mining is useful for Business intelligence application -
→ Balanced scorecard, Fraud Detection, Clickstream Mining, Market Segmentation, Retail Industry, Telecommunications Industry, Banking, Finance and CRM.

* Balanced Scorecard :-
- It is a framework for managing business performance.
- BSC provides executives and managers with a method for reporting and analyzing key performance.
- It is a management technique for structuring their scorecards and displays financial, internal process, customer and learning & growth data.

vision

170200107044

* Fraud Detection :-
- Fraud detection for telecommunication industry.
- With the increasing number of mobile phone access, global mobile phone, fraud is also set to rise.
- At a low level, simple rule-based detection systems are ruled such as the apparent use of the same phone in two very distant geographical locations in quick succession.
- At a higher level, statistical summarise of call distributions are compared with thresholds determined either by experts or by application of supervised learning methods to known fraud cases.

* Clickstream Mining :-
- The approach which is used by most of the people for surfing information on websites is difficult to analyze and understand.
- Quantitative data can lack information about what a users actually intends to get, while quantitative data tends to be localized and is impractical to gather for large samples.

* Retail Industry :-
- It is a major application area for data mining since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption and service.
- The quantity of data collected continue to expand rapidly.

vision

170700107044

* Telecommunications Industry :-
- The telecommunication market is rapidly expanding and highly competitive.
- This creats a great demand for data mining in order to help understand the business involved. Identify telecommunication patterns, catch fraudulent activities, make better use of resources and improve the quality of services.

* Banking and Finance :-
- Most banks and financial institutions offer a wide variety of banking services, credit and investment services.
- Data warehouse needs to be constructed.
- Multidimensional data analysis methods should be used.
- Attribute selection and attribute relevance ranking used.
- Critical factor analysis used for loans.

* CRM :-
- Customer Relationship Management (CRM) emerged in the last decade to reflect the central role of the customer for the strategic positioning of company.
- Data mining helps marketing professionals improve their understanding and customer behaviour.
- In turn, this better understanding allows them to target marketing campaigns more accurately and to align campaigns more closely with needs, wants of customers and prospects.

vision

Assignment - 8

170200107044

Q-1 Discuss text mining, web mining and spatial mining with example.

* Text Mining :-

- Text mining, also referred as text data mining, roughly equivalent to text analytics is the process of deriving high-quality information from text.
- Among the data, most of the data is in unstructured textual form.
- In modern culture, text is the most common way for the formal exchange of information.
- It has became an essential part to develop better techniques and algorithms to extract useful and interesting information from large textual data.

  L> Areas of text mining :-
  => information retrival
  => Natural Language Processing (NLP)

* Web Mining :-

- It is the use of data mining techniques to automatically discover and extract info. from web docs & services.
- There are general classes of information that can be discovered in web mining : web activity, from server logs and web browser history tracking.
- Web mining can be broadly divided into :
  L> Web context mining
  L> Web structure mining
  L> Web usage mining

VISION

1702001070444

* Spatial Mining :-

- It is the application of data mining to spatial models.
- It is based on geographical analysis.
- In spatial data mining, analysis use geographical or spatial information to produce business intelligence or other results.
- It requires specific techniques and measures to get the geographical data into relevant and useful formats.
- The task is to search for spatial patterns.

Q-2 What is Web log? Explain web structure mining and web usage mining in detail.

* Web log :
- A web log is sometimes written as weblog, it is a website that consists of a series of entries arranged in recursive chronological order.

* Web structure Mining :
- It is the application of discovery structure info. from the web.
- The structure of the web graph consists of web pages as nodes and hyperlinks as edges connecting related pages.
- Structure mining basically shows the structural summary of a particular website.
- To determine the connection between two commercial websites, web structure mining can be very useful.

VISION

1702001070044

* Web usage mining :-
- It is the application of identifying and discovering interesting usage patterns from large data sets.
- And these patterns enable you to understand the user behaviour or something like that.
- In web usage mining, user access data on the web and collect data in form of logs.
- So, web usage mining is also called log mining.

Q-3 Both k-means and k-medoids algorithm can perform effective clustering.
- Illustrate the strength and weakness of both.

* Advantages of k-means :-
- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the position of centroids.
- Easily adopts to new examples.

* Disadvantages of k-means :-
- Choosing k manually - "loss v/s clusters"
- Being dependent on initial values.
- Clustering data of varying sizes and density clustering outlines.
- Scaling with number of dimensions.

VISION

170200107044

* Comparision :-

k-means	k-medoids
- Complexity is O(ikm)	- Complexity is $O(ik(n-k)^2)$
- More efficient.	- Compesratively less efficient.
- Sensitive to outlines.	- Not sensitive to outlins.
- Convex shape is required	- Convex shape is not must.