DAV SAMPLE QUESTIONS

Module 1 – Data Analytics Lifecycle

2 Marks – Theory

1. What are the roles of key stakeholders of an analytics project?

Ans)Key roles for a successful analytics project

- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.
- **Project Sponsor:** Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.
- **Project Manager:** Ensures that key milestones and objectives are met on time and at the expected quality.
- **Business Intelligence Analyst :** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPis), key metrics, and business intelligence from a reporting perspective. Business Intelligence analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

- **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data Repositories.
- **Data Engineer:** Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox. The data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

- **Data Scientist:** Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.

2. What are the five main activities performed during *'identifying potential data sources sub-phase* under Discovery phase?

Ans) The team should perform five main activities during this step of the discovery phase:

o **Identify data sources:** Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase. Make an inventory of the datasets currently available and those

that can be purchased or otherwise acquired for the tests the team wants to perform.

o **Capture aggregate data sources:** This is for previewing the data and providing high-level understanding. It enables the team to gain a quick overview of the data and perform further exploration on specific areas. It also points the team to possible areas of interest within the data.

o **Review the raw data:** Obtain preliminary data from initial data feeds. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.

o **Evaluate the data structures and tools needed:** The data type and structure dictate which tools the team can use to analyze the data. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.

• **Scope the sort of data infrastructure needed for this type of problem:** In addition to the tools needed, the data influences the kind of infrastructure that's required, such as disk storage and network capacity.

3. What is Data Conditioning?

Ans)

Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data. A critical step within the Data Analytics Lifecycle, data conditioning can involve many complex steps to join or merge data sets or otherwise get datasets into a state that enables analysis in further phases. It is viewed as a preprocessing step. It involves many operations on the dataset before developing models to process or analyze the data. The data-conditioning step is performed only by IT, the data owners, a DBA, or a data engineer. It is also important to involve the data scientist in this step because many decisions are made in the data conditioning phase that affects the subsequent analysis.

4. In which phase would the team expect to invest most of the project time? Why?

Ans) Understanding the data in detail is critical to the success of the project. The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis. The team may perform data visualizations to help team members understand the data, including its trends, outliers,etc. Each of these are steps of the data preparation phase. Data preparation tends to be the most labor-intensive step in the analytics lifecycle. So, it is common for teams to spend at least 50% of a data science project's time in this critical phase.

5. What are the common questions that are helpful to ask during the Discovery phase when interviewing the project sponsor?

Ans) Following is a brief list of common questions that are helpful to ask during the discovery phase when interviewing the project sponsor. The responses will begin to shape the scope of the project and give the team an idea of the goals and objectives of the project.

• What business problem is the team trying to solve?
• What is the desired outcome of the project?
• What data sources are available?
• What industry issues may impact the analysis?

• What timelines need to be considered?
• Who could provide insight into the project?
• Who has final decision-making authority on the project?
• How will the focus and scope of the problem change if the following dimensions change:
• Time: Analyzing 1 year or 10 years' worth of data?
• People: Assess impact of changes in resources on project timeline.
• Risk: Conservative to aggressive
• Resources: None to unlimited (tools, technology, systems)
• Size and attributes of data: Including internal and external data sources

6. Mention the list of activities in Phase 1.

Ans)

- Learning the Business Domain
- Resources
- Framing the Problem
- Identifying key Stakeholders
- Interviewing the Analytics Sponsor
- Developing Initial Hypothesis
- Identifying Potential Data Sources

7. How to prepare the Analytic Sandbox?

Ans)

The first subphase of data preparation requires the team to obtain an analytic sandbox (also commonly referred to as a workspace), in which the team can explore the data without interfering with live production databases. Consider an **example** in which the team needs to work with a company's financial data. The team should access a copy of the financial data from the analytic sandbox rather than interacting with the production version of the organization's main database, because that will be tightly controlled and needed for financial reporting.

When developing the analytic sandbox, it is a best practice to collect all kinds of data there, as team members need access to high volumes and varieties of data for a Big Data analytics project. This can include everything from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs, depending on the kind of analysis the team plans to undertake. This expansive approach for attracting data of all kind differs considerably from the approach advocated by many information technology (IT) organizations. Because of these differing views on data access and use, it is critical for the data science team to collaborate with IT, make clear what it is trying to accomplish, and align goals. The analytic sandbox enables organizations to undertake more ambitious data science projects and move beyond doing traditional data analysis and Business Intelligence to perform more robust and advanced predictive analytics.

8. List of questions to consider after phase 4.

Ans) Creating robust models that are suitable to a specific situation requires thoughtful consideration to

ensure the models being developed ultimately meet the objectives outlined in Phase 1
Questions to consider include these:
• Does the model appear valid and accurate on the test data?
• Does the model output/behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?
• Do the parameter values of the fitted model make sense in the context of the domain?
• Is the model sufficiently accurate to meet the goal?
• Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance.
• Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?
• Will the kind of model chosen support the runtime requirements?
• Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

5 Marks – Theory

1. Give a brief overview of the main phases of the Data Analytics Lifecycle (with diagram)

Ans)

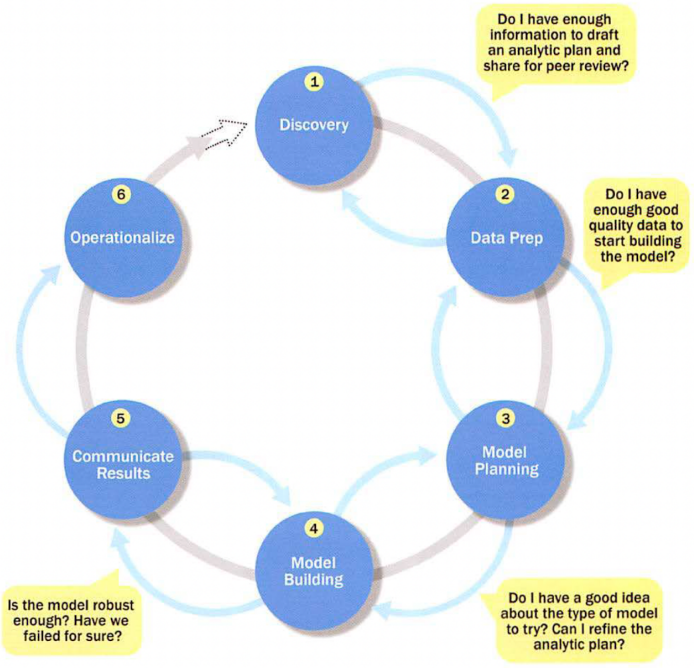


FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Here is a brief overview of the main phases of the Data Analytics Lifecycle:
• **Phase 1- Discovery:** In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business

problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

• **Phase 2- Data preparation:** Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

• **Phase 3-Model planning:** Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

• **Phase 4-Model building:** In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

• **Phase 5-Communicate results:** In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

• **Phase 6-Operationalize:** In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production Environment.

Once team members have run models and produced findings, it is critical to frame these results in a way that is tailored to the audience that engaged the team. Moreover, it is critical to frame the results of the work in a manner that demonstrates clear value. If the team performs a technically accurate analysis but fails to translate the results into a language that resonates with the audience, people will not see the value, and much of the time and effort on the project will have been wasted.

2. What are the common tools used in Phase 2, 3 and 4?

Ans)

Common Tools for the Data Preparation Phase (Phase 2)

Several tools are commonly used for this phase:

• **Hadoop:** can perform massively parallel ingest and custom analysis for web traffic parsing, GPS location analytics, genomic analysis, and combining of massive unstructured data feeds from multiple sources.

• **Alpine Miner :** provides a graphical user interface (GUI) for creating analytic workflows, including data manipulations and a series of analytic events such as staged data-mining techniques on Postgresql and other Big Data sources.

• **Open Refine (formerly called Google Refine):** It is "a free, open source, powerful tool for working with messy data." It is a popular GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.

• **Data Wrangler:** It is an interactive tool for data cleaning and transformation. It can be used to perform many transformations on a given dataset.

Common Tools for the Model Planning Phase (Phase 3)

Many tools are available to assist in this phase. Here are several of the more common ones:

o **R:** has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code.

o **SQL Analysis services:** can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.

• **SAS/ACCESS :** provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, etc.

Common Tools for the Model Building Phase (Phase 4)

There are many tools available to assist in this phase, focused primarily on statistical analysis or data mining
software. Common tools in this space include, but are not limited to, the following:

• **Commercial Tools:**

• **SAS Enterprise Miner**: allows users to run predictive and descriptive models based on large volumes of data from across the enterprise. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.

• **SPSS Modeler:** offers methods to explore and analyze data through a GUI.

• **Matlab:** provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

• **Free or Open Source tools:**

• **Rand PL/R [14)** R was described earlier in the model planning phase, and PL!R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database. This technique
provides higher performance and is more scalable than running R in memory.

• **Octave [22),** a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities

when teaching machine learning.
• **WEKA [23)** is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
• **Python** is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib, seaborn,etc.
• **SQL** in-database implementations, such as MADlib . provide an alterative to in-memory desktop analytical tools.
MADlib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

3. Explain: Why Phase 5 – Communicate Results is critical.
Ans)
Phase 5, "Communicate Results", is critical in the data analytics lifecycle because it involves presenting the analysis results to stakeholders in a way that is easy to understand and actionable. Effective communication of the analysis results is essential for making informed decisions based on the insights gained from the data.

Here are some reasons why Phase 5 is critical:
Making decisions: The primary objective of data analytics is to help stakeholders make informed decisions. Without effective communication of the analysis results, stakeholders may not be able to understand the insights gained from the data or how they can be used to inform decisions.

Trust and credibility: Effective communication of the analysis results builds trust and credibility with stakeholders. If the analysis results are not communicated effectively, stakeholders may question the validity of the analysis or the expertise of the analytics team.

Actionable insights: Effective communication of the analysis results ensures that stakeholders can understand the insights gained from the data and how they can be used to inform decisions. If the analysis results are not communicated effectively, stakeholders may not be able to take action based on the insights gained from the data.

Continued engagement: Effective communication of the analysis results keeps stakeholders engaged in the data analytics process. By presenting the results in an understandable and actionable way, stakeholders are more likely to continue to support the data analytics efforts and invest in future projects.

Eg: If accurate insights are communicated → Stakeholders will work on their limitations and will gain profits.
If accurate insights are not communicated → Stakeholders will work on other things and will not gain profits.

4. What are the benefits of doing a pilot program before a full-scale rollout of a new analytical methodology?
Ans)There are several benefits to doing a pilot program before a full-scale rollout of a new analytical methodology. Here are some of the most important ones:

Identify and resolve issues: A pilot program provides an opportunity to identify and resolve issues before the methodology is implemented on a larger scale. This can help to prevent potential problems that could negatively impact the effectiveness of the methodology and the accuracy of its results.

Test the methodology in a controlled environment: A pilot program allows you to test the methodology in a controlled environment, which can help to identify areas where improvements can be made. This can help to ensure that the methodology is effective and reliable when implemented on a larger scale.

Evaluate the feasibility of the methodology: A pilot program provides an opportunity to evaluate the feasibility of the methodology in terms of resources, costs, and time. This can help to identify any limitations or constraints that may need to be addressed before the methodology is implemented on a larger scale.

Gather feedback from stakeholders: A pilot program provides an opportunity to gather feedback from stakeholders, including users, management, and other stakeholders. This feedback can be used to improve the methodology and ensure that it meets the needs of all stakeholders.

Build support and buy-in: A successful pilot program can build support and buy-in for the methodology, making it easier to implement on a larger scale. This can help to ensure that the methodology is widely adopted and effectively used throughout the organization.

Overall, a pilot program can help to ensure that analytical methodology is effective, reliable, and feasible before it is implemented on a larger scale. This can help to reduce risks and increase the chances of success when rolling out the methodology more broadly.

5. Explain the sub-phase – Performing ETLT under Data Preparation phase.

Ans)

- In ETL, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore.
- However, the analytic sandbox approach differs slightly; it advocates extract, load, and then transform. In this case, the data is extracted in its raw form and loaded into the datastore, where analysts can choose to transform the data into a new state or leave it in its original, raw condition.
- The reason for this approach is that there is significant value in preserving the raw data and including it in the sandbox before any transformations take place.
- Following the ELT approach gives the team access to clean data to analyze after the data has been loaded into the database and gives access to the data in its original form for finding hidden nuances in the data.
- This approach is part of the reason that the analytic sandbox can quickly grow large. The team may want clean data and aggregated data and may need to keep a copy of the original data to compare against or look for hidden patterns that may have existed in the data before the cleaning stage.

- This process can be summarized as ETLT to reflect the fact that a team may choose to perform ETL in one case and ELT in another.

Module 2 – Regression Analysis

2 Marks – Theory

9. Differentiate Correlation and Regression analysis

Ans)

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Eg: Regression :- If student studies more no of hours → Scores good marks (i.e marks dependent on no of hours)

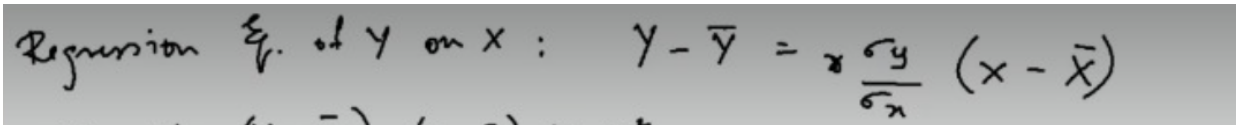
Correlation:- There exists a positive relationship between the sales of ice cream and the climate temperature. This implies that the sales of ice cream are higher in hotter weather conditions. Obviously one tends to crave the ice cream in summers more than in winters.

Extra points:

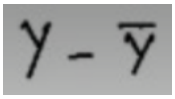
- Regression establishes how *x* causes *y* to change, and the results will change if *x* and *y* are swapped. With correlation, *x* and *y* are variables that can be interchanged and get the same result.
- Correlation is a single statistic, or data point, whereas regression is the entire equation with all of the data points that are represented with a line.
- Correlation shows the relationship between the two variables, while regression allows us to see how one affects the other.
- The data shown with regression establishes a cause and effect, when one changes, so does the other, and not always in the same direction. With correlation, the variables move together.

10. What is the regression equation of Y on X

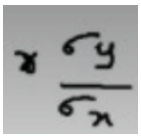
Ans)



Where,



= standard deviation from mean of y



= regression coefficient of y on x

R = pearson's coefficient

Sigma x = SD of x

11. Explain the concepts: Fitted Values and Residuals.

Ans)

A fitted value is a statistical model's prediction of the mean response value when you input the values of the predictors, factor levels, or components into the model. Suppose you have the following regression equation: y = 3X + 5. If you enter a value of 5 for the predictor, the fitted value is 20. Fitted values are also called predicted values.

The *fitted* values, also referred to as the *predicted* values, are typically denoted by $\hat{Y}_i$ (Y-hat). These are given by:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

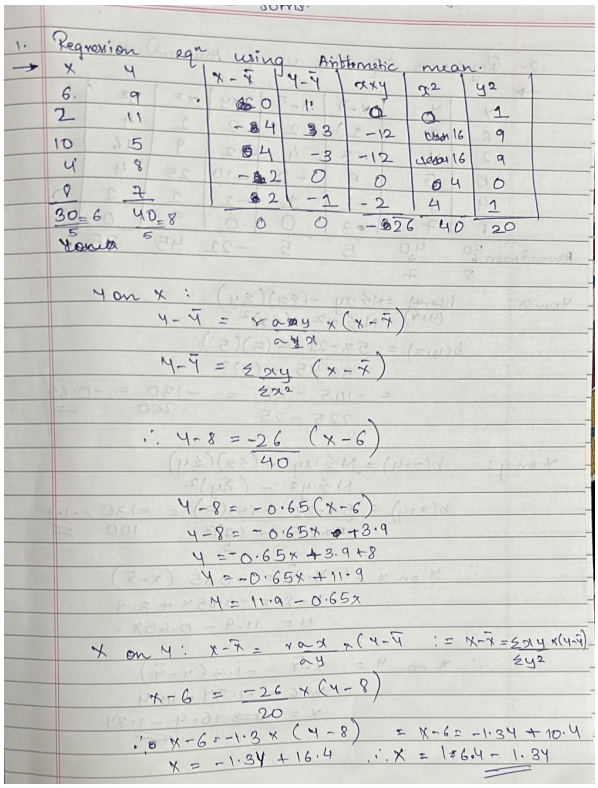
Residuals in a statistical or machine learning model are the differences between observed and predicted values of data. They are a diagnostic measure used when assessing the quality of a model. They are also known as errors.

We compute the residuals $\hat{e}_i$ by subtracting the *predicted* values from the original data:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

5 Marks – Sums

6. Compute the regression equations for the given data using the Arithmetic mean.



1. Regression eqn using Arithmetic mean.

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	x^2	y^2
6	9	0	1	0	0	1
2	11	-4	3	-12	16	9
10	5	4	-3	-12	16	9
4	8	-2	0	0	4	0
8	7	2	-1	-2	4	1
$\frac{30}{5}=6$	$\frac{40}{5}=8$	0	0	-26	40	20

y on x :

$$y-\bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x-\bar{x})$$

$$y-\bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x-\bar{x})$$

$$\therefore y-8 = \frac{-26}{40}(x-6)$$

$$y-8 = -0.65(x-6)$$

$$y-8 = -0.65x+3.9$$

$$y = -0.65x+3.9+8$$

$$y = -0.65x+11.9$$

$$y = 11.9-0.65x$$

x on y : $x-\bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y-\bar{y})$ $\therefore x-\bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y-\bar{y})$

$$x-6 = \frac{-26}{20}(y-8)$$

$\therefore x-6 = -1.3(y-8)$ $\therefore x-6 = -1.3y+10.4$

$x = -1.3y+16.4$ $\therefore x = 16.4-1.3y$

7. Compute the regression equations for the given data using the Assumed mean.

2. Regression equation using Assumed mean.

→

x	y	$x-\bar{x}$	$y-\bar{y}$	xy	x^2	y^2
6	9	1	2	2	1	4
2	11	-3	4	-12	9	16
10	5	5	-2	-10	25	4
4	8	-1	1	-1	1	1
P	7	3	0	0	9	0
Assumed mean 30 / 5	40 / 7	5	5	-21	45	25

y on x:

$$b_{(y,x)} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N\Sigma x^2 - (\Sigma x)^2}$$

$$b_{(y,x)} = \frac{5 \times -21 - (5)(5)}{5 \times 45 - (5)^2}$$

$$= \frac{-105 - 25}{225 - 25} = \frac{-130}{200} = -0.65$$

x on y:

$$b_{(x,y)} = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{N\Sigma y^2 - (\Sigma y)^2}$$

$$b_{(x,y)} = \frac{5 \times -21 - (5)(5)}{5 \times 25 - (7)^2} = \frac{-130}{100} = -1.3$$

$\therefore$ y on x: $y - \bar{y} = -0.65\,(x - \bar{x})$

$y - 8 = -0.65x + 3.9$

$y = 11.9 - 0.65x.$

$\therefore$ x on y: $x - \bar{x} = -1.3\,(y - \bar{y})$

$x - 6 = -1.3y + 10.4$

$x = 16.4 - 1.3y$

Module 4 – Text Analytics

2 Marks – Theory

12. Explain how can you Summarize Text.

Ans)

One of the use case of text mining is the extraction of meaning when the goal is to quickly summarize one or a few very large documents.

There are two types of text summarizations.

- One type summarizes themes across the chapters or paragraphs of the text, in which case the individual paragraphs or chapters can be considered different documents of a larger corpus (the entire text). The goal of this type is to identify the different themes across the various documents (e.g., as just described) or to identify common dimensions or relationships among individuals, events, and so on.
- The second type summarizes the contents of a large text document into a meaningful narrative which cannot be accomplished effectively (yet) using automatic text mining methods and algorithms. In short, it is not realistic to expect that present computer algorithms are capable of summarizing the "essence" of a very large book into a single paragraph. At present, this can be done only in other highly subjective ways.

13. What do you mean by topic in Documents by Topics? How is it useful?

Ans) A topic consists of a cluster of words that frequently occur together and share the same theme. The topics of a document are not as straightforward as they might initially appear.

For example,

ACME wants to categorize the reviews by topics.

We Consider the following review:

1. While I love ACME's bPhone series, I've been quite disappointed by the bEbook. The text is illegible, and it makes even my old NBook look blazingly fast.

For machines, it is difficult to answer whether the review is about bPhone series, bEbook or NBook.

Since a document typically consists of multiple themes running through the text in different proportions, document grouping is required. It can be achieved with various clustering or classification methods. However, a more feasible and prevalent approach is to use topic modeling. Topic models are statistical models that examine words from a set of documents, determine the themes over the text, and discover how the themes are associated or change over time. Thus, it is very useful as it provides tools to automatically organize, search, understand, and summarize from vast amounts of information.

14. What is a caveat of IDF? How does TFIDF address the problem?

Ans)

The inverse document frequency (IDF) is a widely used metric in natural language processing and information retrieval for measuring the importance of a word in a corpus of documents. However, there are some caveats to be aware of when using IDF.

- Bias towards rare words.
- Inability to capture word context.
- Sensitivity to document size.
- Difficulty in comparing IDF values across different corpora.
- Vulnerability to noise

In tf-idf, the frequency of each word in a document is weighted by its inverse document frequency (IDF) score, which gives more weight to words that are rare in the corpus and less weight to words that are common. This helps to address the bias towards rare words that can occur with IDF alone.

In addition, tf-idf also takes into account the term frequency (TF) of each word in a document. This helps to address the limitation of IDF in capturing word context. By considering both the TF and IDF of each word, tf-idf can give higher weight to words that are both rare and important in a specific document.

15. Explain IDF and TFIDF. How does TFIDF overcomes issue of using IDF?

Ans)

IDF measures the rarity of a word in a corpus of documents. The intuition behind IDF is that words that appear in many documents are less informative than words that appear in few documents.

The TFIDF (or TF-IDF) is a measure that considers both the prevalence(context) of a term within a document (TF) and the scarcity(rarity) of the term over the entire corpus (IDF).

Tf-idf overcomes some of the limitations of using IDF alone. Here are some of the ways in which tf-idf addresses the issues with using IDF:

- Addresses the bias towards rare words.
- Captures word context.
- Less sensitive to document size.
- Can be used to compare scores across different corpora.
- Less vulnerable to noise.

16. Explain Precision and Recall. Calculate precision and recall from above given confusion matrix.

Ans)

Precision and Recall

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

- **Recall:** completeness – what % of positive tuples did the classifier label as positive?

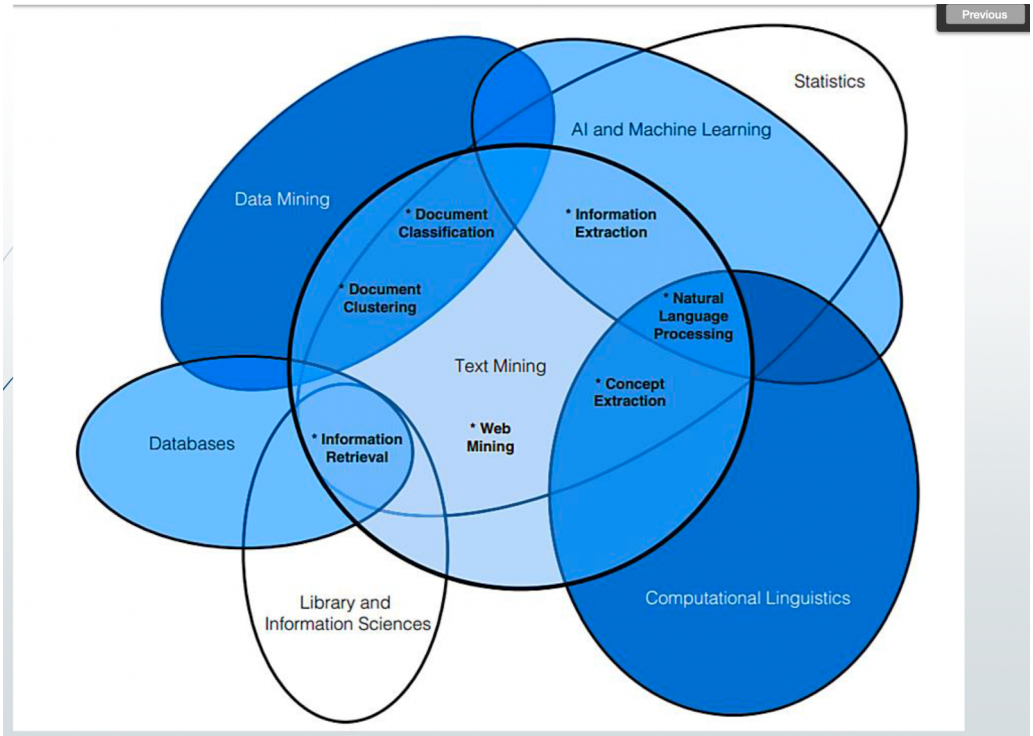
$$recall = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall

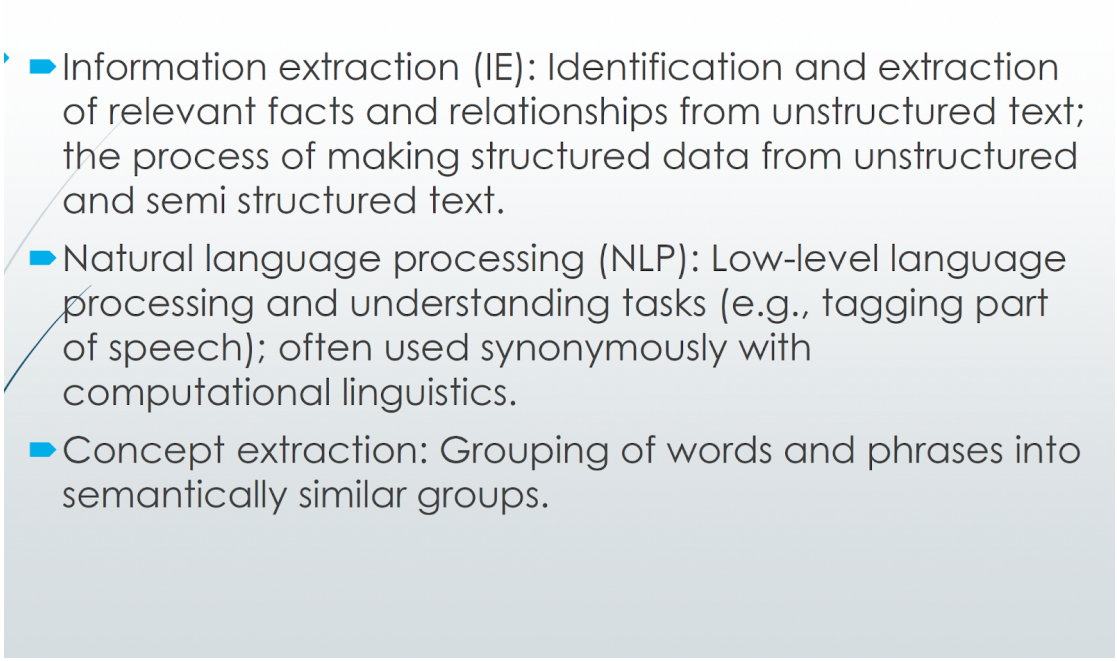
5 Marks – Theory

8. Explain seven practices of text analytics. Give example application in that area.

Ans)



- Search and information retrieval (IR): Storage and retrieval of text documents, including search engines and keyword search.
- Document clustering: Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
- Document classification: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
- Web mining: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.



Examples:
IR: one of the real time applications of IR is that it is used in search engines like google,bing,etc
Document Clustering: A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.
Document Classification: It can be used in Gmail as a spam classifier making it easy for users and keeping their inbox clear.
Web Mining: It can be used to collect data from the web i.e retrieve relevant information using mining.
IE: One of the example of IE is as follows-
“Shanaya lives in Mumbai city.”
Through information extraction, the following basic facts can be pulled out of the free-flowing text and organized in a structured, machine-readable form:

Person: Shanaya

Location: Mumbai

NLP: For example, it is used in Chatbot like ChatGPT.

Concept Extraction: For example, when searching Google images for “fluffy cat”, I have defined a set of features *F={fluffy, cat}*, and the response Google gives—a collection of fluffy cat images—is the concept *Y* extracted for the feature set *F*.

9. Explain three important steps for text analysis problem in detail.

Ans)

Text Analysis Steps

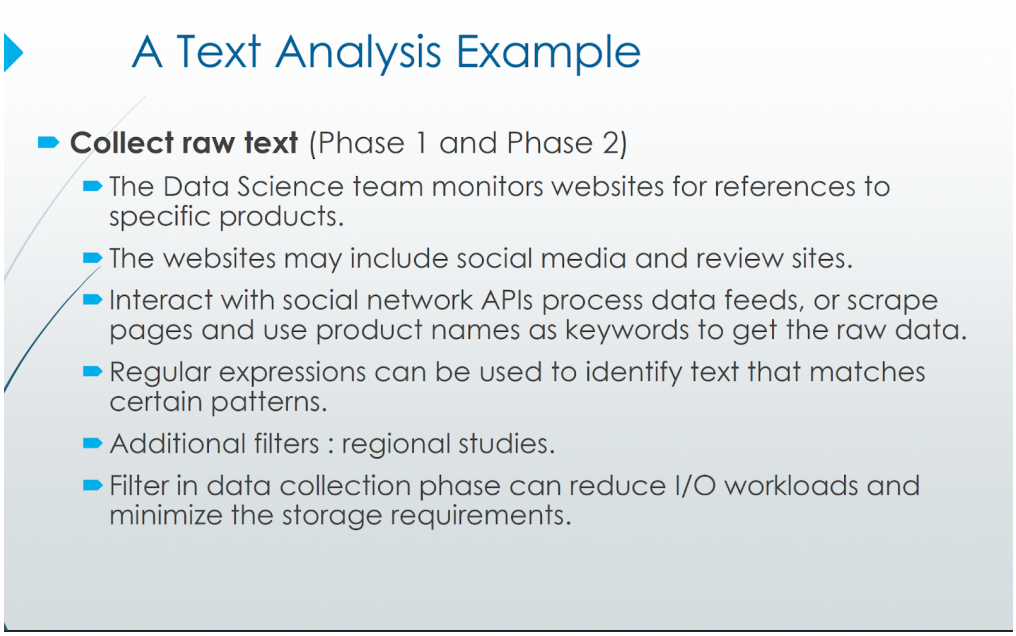
A text analysis problem usually consists of three important steps: parsing, search and retrieval, and text mining.

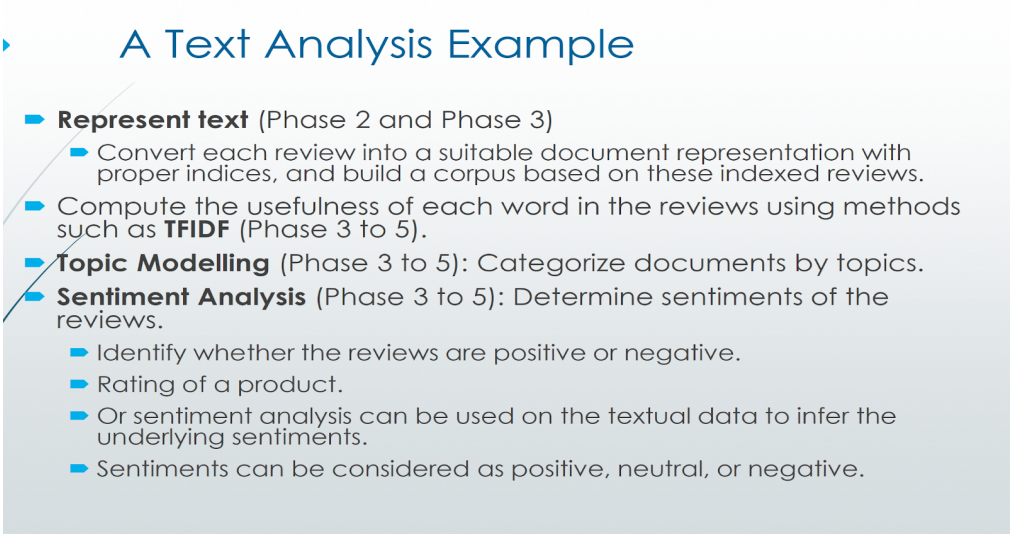
1. **PARSING** is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Markup Language (XML) file, a HyperText Markup Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.
2. **SEARCH AND RETRIEVAL** is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines.
3. **TEXT MINING** uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques, such as clustering and classification, can be adapted to text mining. K-means can be modified to cluster text documents into groups, where each group represents a collection of documents with a similar topic. The distance of a document to a centroid represents how closely the document talks about that topic. Classification tasks such as sentiment analysis and spam filtering are prominent use cases for the naïve Bayes classifier.

All three steps may not be included in all projects. Also they may have any sequence.

10. Explain how can you perform Collection Raw Text and Representing that Text using any example application.

Ans)





Let's take the example of an application that collects and represents news articles. This application could collect news articles from various sources using web scraping tools or APIs and store them in a database. Once the news articles are collected, the application could represent them using a bag-of-words approach. The bag-of-words approach involves creating a matrix of word frequencies in the news articles. For example, the application could use the bag-of-words matrix to perform sentiment analysis on the news articles to determine the overall sentiment of each article. The application could also use the matrix to perform topic modeling to identify the main topics discussed in the news articles. The application could then use this information to provide users with personalized news recommendations based on their interests.