Page No.

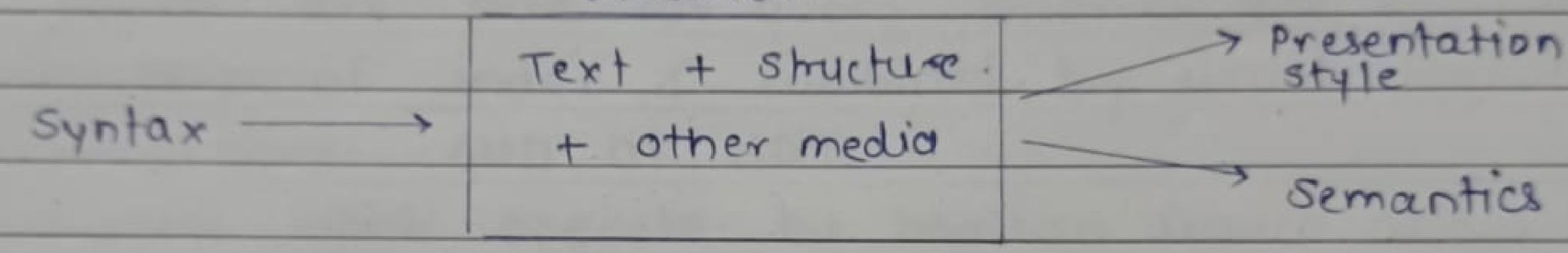
Date : / /

Name : Dhiraj Bodake

Roll No : 18141216

# TUTORIAL NO : 05

Q.1 Explain the characteristics of documents with the help of suitable diagram.




Syntax of document can express structure, presentation style, semantics are even external actions. Syntax of a doc can be implicit in its context or expressed in a simple declarative language or even in programming language.

- Text can be written in natural language. However at present semantics of natural language is still not easy for computer to understand.
- Current trend is to use languages which provide information or document structure. Format & semantic while being readable by human as well as computers.
- The style of doc defines how the doc is visualised in computer window or printed page

media such as audio or video.

Q.2) Explain metadata with respect to information retrieval.

→ Metadata is information on the organization of data, the various data domains & the relationship between them. Metadata is 'data about the data'.

Common forms of metadata associated with text include author, date of publication the source of publi$^{n}$, doc. length & doc. genre Descriptive metadata is the metadata that is external to meaning of doc., & pertains more to how it was created.

- Another type of metadata characterizes the subject matter that can be found within docs. content. This is semantic metadata.
- An important metadata format is the Machine Readable cataloging Record (MRC) which is the most used format for library records. MARC has several fields for different attributes of bibilographic entry.
- In the web, metadata can be used for many purposes. New standard for web metadata is Resource Description framework (RDF) which provides interoperability bet$^{n}$ appl$^{n}$
- This framework allows description of web resources to facilitate automated processing of info.

Page No.:
Date:

Q.3. Explain information theory in details with help of mathematical expressions.

→ Written text has certain semantics & is a way to communicate info. Although it is difficult to formally capture how much info. is there in given text, the distribution of symbols is related to it.

e.g. a text where one symbol appears almost all time does not convey much informat$^{n}$.

Information theory defines a special concept entropy to capture information content.

Entropy of text is defined as

$$E = \sum_{i=1}^{\sigma} P_i \log_2 P_i$$

In this formula, $\sigma$ symbols of alphabet are coded in binary, so entropy is measured in bits.

The definition of entropy depends on the probabilities we need text model. So we can say that amount of info. in a text is measured with regard to text model.