Module-4:

Q7. a). Explain the semantically guided model for effective text mining.

→• We developed a semantically guided model for evolutionary Text Mining which is domain-independent but genre-based.

- In order to deal with issues regarding representation and new genetic operations so to produce an effective KDT process, our working model has been divided into two phases.

- The first phase is the preprocessing step aimed to produce both training information for further evaluation and the initial population of the GA (Genetic Algorithms).
- The second phase constitutes the knowledge discovery itself, in particular this aims at producing and evaluating explanatory unseen hypotheses.

- The whole processing starts by performing the IE task which applies extraction patterns and then generates a rule-like representation for each document of the specific domain corpus. After processing a set of n documents, the extraction stage will produce n rules, each one representing the document's content in terms of its conditions & conclusions. Once generated, these rules, along with other training data, become the "model" which will guide the GA-based discovery.

1. Text Preprocessing and Training: The preprocessing phase has two main goals: to extract important information from the texts and to use that information to generate both training data & the initial population for the GA.

2. Knowledge Discovery and Automatic Evaluation of Patterns: (Semantic and Rhetorical information)

The GA will start from a initial population, which in this case, is a set of semi random hypotheses built up from the preprocessing phase. Next, constrained GA operations are applied and the hypotheses are evaluated.

- Hypothesis Discovery: Using the semantic measure above and additional constraints, we propose new operations to allow guided discovery such that unrelated new knowledge is avoided:
  - Selection
  - Crossover
  - Mutation (small random changes).

- Evaluation: In order to establish evaluation criteria, we have taken into account different issues concerning plausibility and quality itself.

  Accordingly we have defined eight evaluation criteria to assess the hypotheses given by: relevance, structure, cohesion, interestingness, coherence, coverage, simplicity, plausibility of origin.

b). Define the following with an example for each.

- Cohesion
- Co-Matrix
- LSA.

→ • Cohesion:

Cohesion is the degree to which ideas in the text are explicitly related to each other and facilitate a unified situation model for the reader.

Eg: Recent research in text processing has emphasized the importance of the cohesion of a text in comprehension

- "The sun was shining brightly in the sky."
- "People were enjoying the warm weather."

- The word "weather" in sentence 2 is related To the word "sun" in sentence 1 creating cohesion b/w 2 sentences.

• Coh-Metrix:

Coh-Metrix assesses characteristics of texts by using parts of speech classifiers and latent semantic analysis.

The indices generated from Coh-Metrix offer an assessment of the cohesiveness and readability of any given text.

Eg: Coh-Metrix is used for word similarity, word clustering and semantic analysis for text identification.

• **LSA:**

Latent Semantic Analysis is used to assess the inter-relatedness of text sections.

LSA is a technique that uses a large corpus of texts together with singular value decomposition to derive a representation of world knowledge.

Eg: chair|chairs or run|ran are managed to rate the relative semantic similarity between terms such as chair|table, table|wood, wood|forest.

Q8. a). Describe Text Coherence. Discuss the significance of Text Coherence in Discourse Segmentation.

→. Text coherence refers to the logical and meaningful connection b/w sentences and paragraphs in a text.

- The significance of Text Coherence in Discourse Segmentation:

1. Readibility → Coherent Texts are easier to read and understand, as they present information in a logical manner.
2. Sense-making → Coherent texts helps readers make sense of the information represented.
3. Identifying boundaries → Text coherence assists in determining where one segment or paragraph ends and another begins.
4. Facilitating analysis → For NLP tasks, such as sentiment analysis or others, coherent segments are vital.

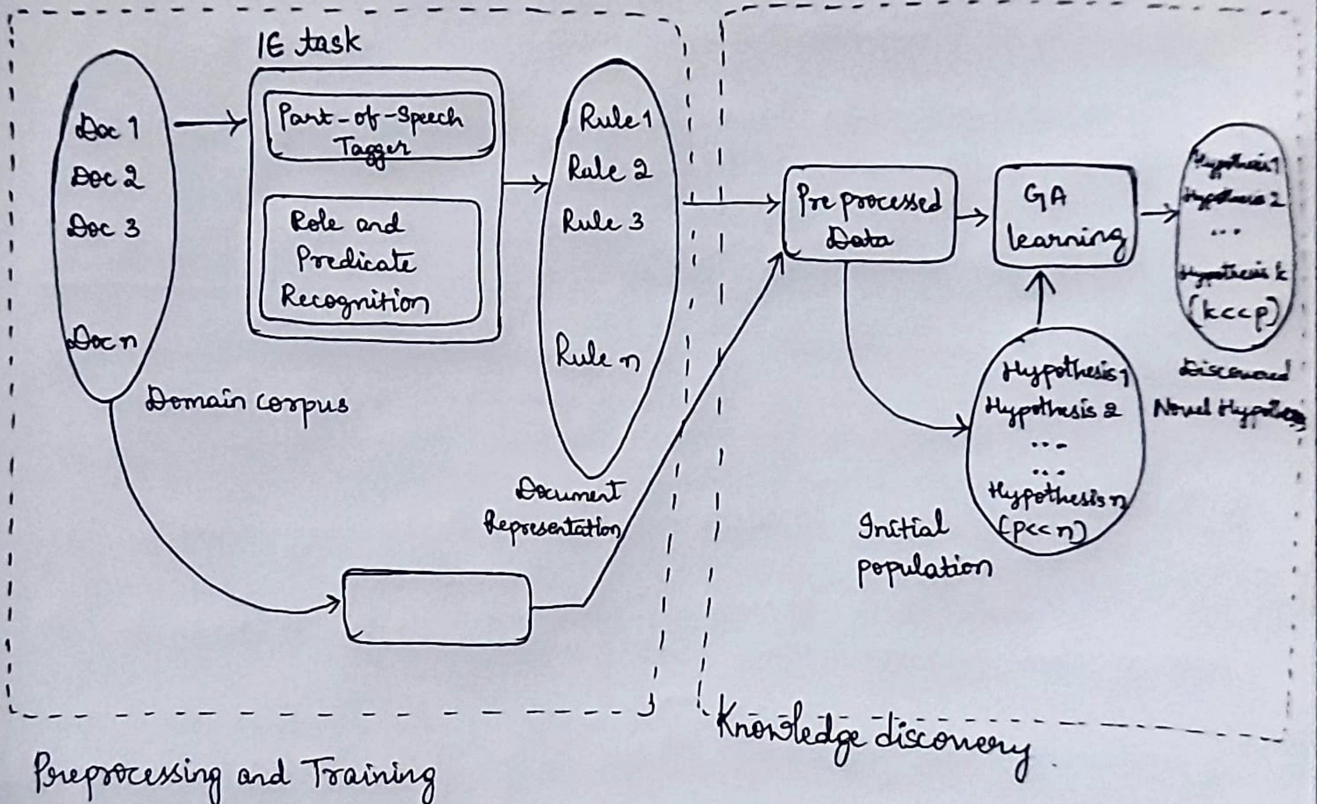
5 Language understanding → Text coherence is a fundamental aspect of language comprehension. Discourse segmentation helps in understanding the language more precisely.

6. Natural flow → Coherent discourse creates a natural flow of information, making it more effective To read.

- In discourse segmentation, Text coherence serves as a guiding principle for dividing larger texts into smaller, manageable, enhancing overall readability and understanding the content, meaningful units.

b). With the neat diagram explain the evolutionary model for KDT (Knowledge Discovery from Text)(Ans- 7a.)

→.



(Before steps add this):

In order to generate an initial set of hypotheses, an initial population is created by building random hypotheses from the initial rules, i.e., hypotheses contain predicate and rhetorical information from the rules are constructed.

The GA then runs for a number of generations until a fixed number of generations is achieved. At the end, a small set of the best hypotheses are obtained.