Question-1

- **Assumptions:**
  - Entered query will be a Boolean formula.
  - No spelling errors in the query.
  - No punctuation marks will be present in the query.
  - No stop words present in the query.
  - No single characters present in the query.
- **Preprocessing of Data:**
  - Text of the file converted to lower case.
  - Metadata removal.
  - Removal of punctuation marks.
  - Removal of stop words.
  - Removal of single characters.
  - Tokenization of text stream.
  - Lemmatization of tokens.
- **Preprocessing of Query:**
  - Text converted to lower case.
  - Tokenization of query text.
  - Lemmatization of query text.
- **Methodology:**
  - After the preprocessing of data, inverted index was created using the standard steps.
    - It has been stored as a dictionary.
  - For retrieving the documents, the query was first split on the basis of the “or” operator.
  - The individual parts of the split query which will contain only “and” and “not” operator were handled first using the merge algorithm.
  - These intermediate results were stored and then the “or” operator was handled.
  - Number of comparisons were reduced using the frequency count of words in the index. The two smallest postings were selected at every intermediate step for retrieval of required documents.
  - Comparisons recorded were the ones performed in the merge operation.