Coding it in java

1. Using Stanford’s Core NLP library
2. We can create pipelines in that

Approach

1. Linguistic Preprocessing
   1. Tokenization
   2. Stemming/Lemmatization
   3. Stop word/punctuation removal
2. Using the tokens from above to create Index
3. Classes
   1. Preprocessing
      1. Attributes: annotator, ..
      2. Operations: all preprocessing
   2. Term
      1. Attributes: Term id, term, posting lists, size of posting list
      2. Operations:
   3. Index
      1. Attributes:
      2. Operations: merge, intersect, phrase query, query, ...