Chapter No. 2 The term vocabulary and postings lists

<Food for Thoughts>

1. Explain what we mean by document processing pipeline for IR system.
2. What are the challenges in Document Processing? Explain each with an example.
3. Differentiate between Stemming and Lemmatization of a natural language token(word).
4. We have a two-word query. For one term the postings list consists of the following 16 entries:
   [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]
   and for the other it is the one entry postings list:
   [47]
   Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:
   a. Using standard postings lists
   b. Using postings lists stored with skip pointers, with a skip length of √P, as suggested in Section 2.3.
5. What do we mean by extended bi-words? How it is useful?
6. How positional indexing support phrase and proximity queries? Explain each with an example.
7. How a combination of bi-word indexing and positional indexing benefits an IR system?