IRS - Unit - 2 Imp.

1. BSBI Algorithm:

With insufficient main memory, we need to use an external sorting algorithm, i.e., one that uses disk.
One solution is blocked-sort-based indexing algorithm or BSBI.

→ BSBI segments the collection into parts of equal size.
→ BSBI sorts term ID-doc ID pairs of each part in memory.
→ BSBI stores intermediate sorted results on disk and
→ BSBI merges all intermediate results into the final index.

Algorithm:

```
BSBINDEXCONSTRUCTION()
1. n ← 0
2. while (all documents have not been processed)
3. do n ← n+1
4.     block ← PARSENEXTBLOCK()
5.     BSBI-INVERT(block)
6.     WRITEBLOCKTODISK(block, fn)
7. MERGEBLOCKS(f1, ... fn; fmerged).
```

2. Spelling-Correction Algorithm:

Ans: Indexing for spelling correction is a bit different than for document retrieval.

Most common spelling errors are:

1. Insertion: adding an extra letter. Ex: truely.
2. Deletion: Missing a letter
3. Transposition: Interchanging of a letter's position
4. Substitution.

To overcome these errors & issues Spelling-Correction Algorithm is used.

Spelling Correction Algorithm:

1. Here, a Ternary Search tree stores dictionary.
2. For each stored word, we also keep frequency count, obtained from analysis of large corpora.
3. Queries entered in the search engine are parsed & individual terms are extracted with non word tokens ignored. Each word is then converted to lower case & checked to see if it is correctly spelled. These, correctly spelled words found in user queries are updated in the dictionary, by incrementing their frequency count.
4. Edit distance, k-gram indexes are the two types of basic Spelling correction.

Algorithm:

```
EDIT DISTANCE (S1, S2)
1.  int m[|S1|, |S2|] = 0
2.  for i ← 1 to |S1|
3.  do m[i,0] = i
4.  for j ← 1 to |S2|
5.  do m[0,j] = j
6.  for i ← 1 to |S1|
7.  do for j ← 1 to |S2|
8.     do m[i,j] = min{m[i-1,j-1] + if (S1[i] = S2[j]) then 0 else 1 fi, m[i-1,j]+1,
9.                     m[i,j-1]+1}.
10. return m[|S1|, |S2|].
```

Stemming Algorithm:

It is a process of linguistic normalisation, in which variant forms of a word are reduced to a common form. for example, connection connections connective → connect connected connecting.

→ It reduces a word to its base word using various approaches.

→ It is simple to develop.

→ There are two types of errors that occur. They are:

(i.) Over-Stemming:

It occurs when 2 words are stemmed from same root that are of different stems. It can also be regarded as false-positive.

(ii.) Under-Stemming:

It occurs when 2 words are stemmed from same root that are not of different stems. It can be interpreted as false-negatives.

→ There are various Stemming algorithms - Porter's Stemming Algorithm, Lovins Stemmer, Dawson Stemmer, Krovetz Stemmer etc.

APPLICATIONS:

Stemming is used in information retrieval systems like Search engines.

It is used to determine domain vocabularies in domain analysis.