2.9 N-Gram Data Structures

- N-Grams can be viewed as a special technique for conflation (stemming).
- It is an unique data structure in information systems that ignores words and treats the input as a continuous data, optionally limiting its processing by interword symbols.
- N-Grams are a fixed length consecutive series of "n" characters or fixed length overlapping symbol segments that define the searchable processing tokens.
- These tokens have logical linkages to all the items in which the tokens are found.
- Inversion lists, document vectors and other proprietary data structures are used to store the linkage data structure and are used in the search process.
- Unlike stemming that generally tries to determine the stem of a word that represents the semantic meaning of the word, n-grams do not care about semantics.

- Examples of bigrams, trigrams and pentagrams are given in Figure 4.7 for the word phrase "sea colony".
- For n-grams, with n greater than two, some systems allow interword symbols to be part of the n-gram set usually excluding the single character with interword symbol option.
- The symbol # is used to represent the interword symbol which is anyone of a set of symbols (e.g., blank, period, semicolon, colon, etc.).

se ea co ol lo on ny	Bigrams (no interword symbols)
sea col olo lon ony	Trigrams (no interword symbols)
#se sea ea# #co col olo lon ony ny#	Trigrams (with interword symbol #)
#sea# #colo colon olony lony#	Pentagrams (with interword symbol #)

Figure 4.7 Bigrams, Trigrams and Pentagrams for "sea colony"

- Each of the n-grams created becomes a separate processing token and are searchable.
- It is possible that the same n-gram can be created multiple times from a single word.

- The advantage of n-grams
  - they place a finite limit on the number of searchable tokens.
  - maximum number of unique n-grams can be generated.
  - implementation techniques allow for fast processing on minimally sized machines
- disadvantage
  - false hits can occur under some architectures.
  - increased size of inversion lists (or other data structures) that store the linkage data structure.
  - In effect, use of n-grams expands the number of processing tokens by a significant factor.
  - There is no semantic meaning in a particular n-gram since it is a fragment of processing token and may not represent a concept.
  - Thus n-grams are a poor representation of concepts and their relationships.