Commonly Used Text Features (cont.)

- Word classes
  - Syntactic (= POS tags)
  - Semantic Concept: e.g., thesaurus/ontology, recognized entities
  - Empirical word clusters (e.g., cluster of paradigmatically or syntagmatically related words)
- Frequent patterns in text (e.g., frequent word set; collocations)
  - More specific/discriminative than words
  - May generalize better than pure n-grams
- Parse tree-based (e.g., frequent subtrees, paths)
  - Even more discriminative, but need to avoid overfitting
- Pattern discovery algorithms are very useful for feature construction