Feature Construction for Text Categorization

- Feature design affects categorization accuracy significantly
- A combination of machine learning, error analysis, and domain knowledge is most effective
  - Domain knowledge → seed features, feature space
  - Machine learning → feature selection, feature learning
  - Error analysis → feature validation
- NLP enriches text representation ➔ enriches feature space (more likely overfitting!)
- Optimizing the tradeoff between **exhaustivity** and **specificity** is a major goal

high coverage (frequent) **discriminative (infrequent)**