Identifying Features from Text saved

Why is textual data unique?

- Textual data presents a unique set of challenges
- All the information you need is in the text
- But features can be pulled out from text at different granularities!

Types of textual features (1)

- Words
  - By far the most common class of features
  - Handling commonly-occurring words: Stop words
  - Normalization: Make lower case vs. leave as-is

Types of textual features (2)

- Characteristics of words : Capitalization
- Parts of speech of words in a sentence

Types of Textual features (3)

- Depending on classification tasks, features may come from inside words and word sequences
  - bigrams, trigrams, n-grams: “White House”
  - character sub-sequences in words: “ing”, “ion”, …