

# Applied Text Mining in Python

*Identifying Features from Text*

# 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 (I)

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

# Types of Textual Features (2)

- **Characteristics of words : Capitalization**
- **Parts of speech of words in a sentence**
- **Grammatical structure, sentence parsing**
- **Grouping words of similar meaning, semantics**
  - **{buy, purchase}**
  - **{Mr., Ms., Dr., Prof.}; Numbers / Digits; Dates**

# 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”, ...

# How would you do it?

- Recall lectures from previous week