Lecture 2: What will we discuss today?

Lecture 2: Overview

Today, we'll look at **words**:
— How do we identify words in text?
— Word frequencies and Zipf's Law
— What is a word, really?
— What is the structure of words?
— How can we identify the structure of words?

To do this, we'll need a bit of linguistics,
some data wrangling, and a bit of automata theory.

Later in the semester we'll ask more questions about words:
How can we identify different word classes (parts of speech)?
What is the meaning of words? How can we represent that?

Lecture 2: Reading

Most of the material is taken from Chapter 2 (3rd Edition)

I won’t cover regular expressions (2.1.1) or edit distance (2.5), because I assume you have all seen this material before.
I you aren’t familiar with regular expressions, read this section because it’s very useful when dealing with text files!

The material on finite-state automata, finite-state transducers and morphology is from the 2nd Edition of this textbook, but everything you need should be explained in these slides.

Lecture 2: Key Concepts

You should understand the distinctions between
— Word forms vs. lemmas
— Word tokens vs. word types
— Finite-state automata vs. finite-state transducers
— Inflectional vs. derivational morphology

And you should know the implications of Zipf's Law for NLP (coverage!)