Week 2

- Representing words and meanings
- Language modeling

Agenda

-
-
-

Tokenization

Tokenization lies at the hearth of any natural language processing analysis. Hence, it's.

Naive tokenizer

NLTK tokenizers

N-grams & tokenizers

Modeling natural language: Main challenges

Processing raw text intelligently is difficult:

- it's common for words that look completely different to mean almost the same thing
- the same words in a different order can mean something completely different
- most words are rare
- even splitting text into useful word-like units can be difficult in many languages (see Japanese)

Source is spaCy website

Pre-DL models of the language

Post-DL models of the language

The architecture of Natural Language Models