Natural Language Processing

Natural Language Processing

- Install tools

```
$ conda install nltk
```

Language

A language is a set of strings.

- Symbols
  - Lexicon (alphabet)
- Tokens
  - Words: strings of symbols
  - Specified with regular grammar (regular expressions)
- Sentences
  - Strings of tokens
  - Specified with context-free grammar

Natural and Artificial Languages

Artificial language

- Prescribed by regular and context free grammars.

Natural language

- Described by regular and context free grammars.

Analyzing Natural Languages

Symbolic NLP

- Parsing

Statistical NLP

- N-grams
- HMMs
- Vector space model

Bag of Words

- Documents as vectors of word counts
- Document categorization
- Information retrieval