3. Processing Raw Text

3.1 Accessing Text from the Web and from Disk
1.1 Electronic Books
1.2 Dealing with HTML
1.3 Processing Search Engine Results
1.4 Processing RSS Feeds
1.5 Reading Local Files
1.6 PDF, MSWord
1.7 Capturing User Input
1.8 The NLP Pipeline

3.2 Strings: Text Processing at the Lowest Level
2.1 Basic Operations with Strings
2.2 Printing Strings
2.3 Accessing Individual Characters
2.4 Accessing Substrings
2.5 More operations on strings
2.6 The Difference between Lists and Strings

3.3 Text Processing with Unicode
3.1 What is Unicode?
3.2 Extracting encoded text from files
3.3 Using your local encoding in Python

3.4 Regular Expressions for Detecting Word Patterns
4.1 Using Basic Meta-Characters
4.2 Ranges and Closures

3.5 Useful Applications of Regular Expressions
5.1 Extracting Word Pieces
5.2 Doing More with Word Pieces
5.3 Finding Word Stems
5.4 Searching Tokenized Text

3.6 Normalizing Text
6.1 Stemmers
6.2 Lemmatization

3.7 Regular Expressions for Tokenizing Text
7.1 Simple Approaches to Tokenization
7.2 NLTK's Regular Expression Tokenizer
7.3 Further Issues with Tokenization

3.8 Segmentation
8.1 Sentence Segmentation
8.2 Word Segmentation

3.9 Formatting: From Lists to Strings
9.1 From Lists to Strings
9.2 Strings and Formats
9.3 Lining Things Up
9.4 Writing Results to a File
9.5 Text Wrapping