2017 봄학기

Natural Language Processing with Python

Chapter 2. Accessing Text Corpora and Lexical Resources

융합과학기술대학원

인간중심컴퓨팅연구실

H U M A N
C E N T E R E D
C O M P U T I N G
L A B O R A T O R Y

목차

HUMAN
CENTERED
COMPUTING
LABORATORY

Corpora

- Large bodies of linguistic data

1. What are some useful text corpora and lexical resources, and how can we access them with Python?
2. Which Python constructs are most helpful for this work?
3. How do we avoid repeating ourselves when writing Python code?

Corpora in NLTK

- Gutenberg Corpus
- Web and Chat Text
- Brown Corpus
- Reuters Corpus
- Inaugural Address Corpus
- Annotated Text Corpora
- …

Conditional Frequency Distributions

- A collection of frequency distributions, each for a specific “condition”
  - ConditionalFreqDist()

Example	Description
`cfdist = ConditionalFreqDist(pairs)`	Create a conditional frequency distribution from a list of pairs
`cfdist.conditions()`	Alphabetically sorted list of conditions
`cfdist[condition]`	The frequency distribution for this condition
`cfdist[condition][sample]`	Frequency for the given sample for this condition
`cfdist.tabulate()`	Tabulate the conditional frequency distribution
`cfdist.tabulate(samples, conditions)`	Tabulation limited to the specified samples and conditions
`cfdist.plot()`	Graphical plot of the conditional frequency distribution
`cfdist.plot(samples, conditions)`	Graphical plot limited to the specified samples and conditions
`cfdist1 < cfdist2`	Test if samples in `cfdist1` occur less frequently than in `cfdist2`

Let’s Not Repeat Code

- Text Editors
- Libraries - Packages - Modules - Functions

Some Important Vocabularies

- Lexical Entry: Headword (Lemma) + Additional Information

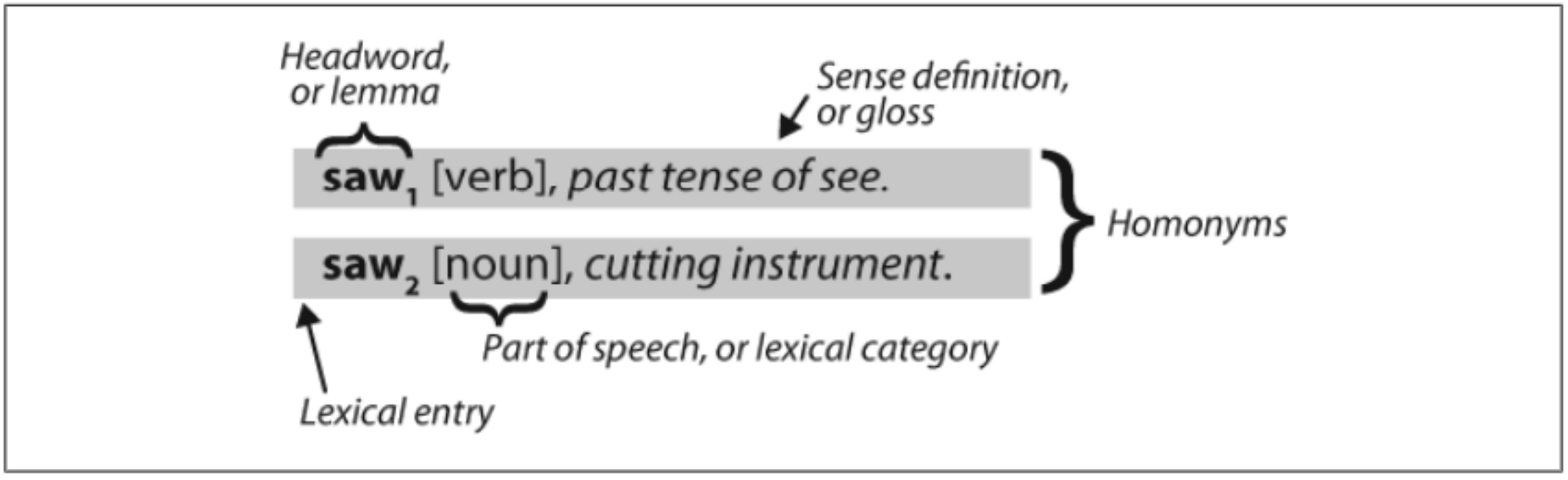
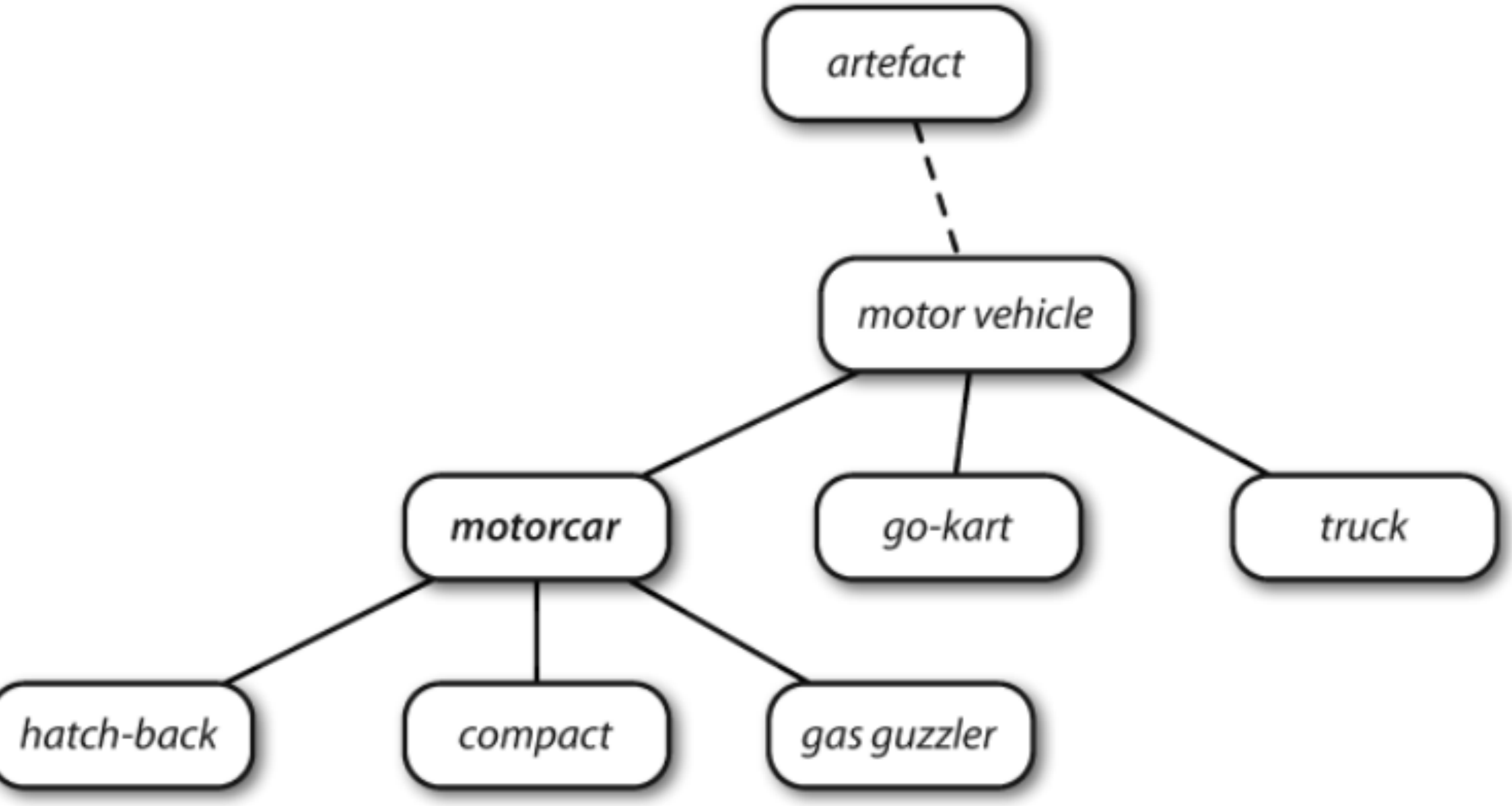


Figure 2-5. Lexicon terminology: Lexical entries for two lemmas having the same spelling (homonyms), providing part-of-speech and gloss information.

Some Corpora

- Wordlist Corpora
  - Stopwords: High-frequency words that we want to filter out before processing
- Pronouncing Dictionary
- Comparative Wordlists
- Shoebox and Toolbox Lexicons
  - Toolbox: A collection of entries, of which entries consist of one or more fields
    - e.g. P-O-S, gloss information, other languages, etc.



2-8. Fragment of WordNet concept hierarchy: Nodes correspond to synsets; edges indicate
ym/hyponym relation, i.e., the relation between superordinate and subordinate concepts.

More WordNet

- Meronyms: Items and their components
- Holonyms: Items and the things they are contained in
- Antonyms: Items that other items that are semantically opposite
- Semantic Similarity:
  - path_similarity()

Chapter 2 Summary

- A text corpus is a large, structured collection of texts. NLTK comes with many corpora, e.g., the Brown Corpus, `nltk.corpus.brown`.
- Some text corpora are categorized, e.g., by genre or topic; sometimes the categories of a corpus overlap each other.
- A conditional frequency distribution is a collection of frequency distributions, each one for a different condition. They can be used for counting word frequencies, given a context or a genre.
- Python programs more than a few lines long should be entered using a text editor, saved to a file with a *.py* extension, and accessed using an `import` statement.
- Python functions permit you to associate a name with a particular block of code, and reuse that code as often as necessary.
- Some functions, known as "methods," are associated with an object, and we give the object name followed by a period followed by the method name, like this: `x.funct(y)`, e.g., `word.isalpha()`.
- To find out about some variable `v`, type `help(v)` in the Python interactive interpreter to read the help entry for this kind of object.
- WordNet is a semantically oriented dictionary of English, consisting of synonym sets—or synsets—and organized into a network.
- Some functions are not available by default, but must be accessed using Python's `import` statement.

감사합니다

박소현

4/13/2017

HUMAN
CENTERED
COMPUTING
LABORATORY