20-06-2023

**Chapter 2: Accessing Text Corpora and Lexical Resources**

1. **Gutenberg Corpus:**

Gutenberg Corpus NLTK includes a small selection of texts from the Project Gutenberg electronic text archive, which contains some 25,000 free electronic books, hosted at http://www.gu tenberg.org/.

* The raw () function gives us the contents of the file without any linguistic processing.
* Most NLTK corpus readers include a variety of access methods apart from words (), raw (), and sents (). Richer linguistic content is available from some corpora, such as part-of-speech tags, dialogue tags, syntactic trees, and so forth.

1. **Web and Chat Text:**

NLTK’s small collection of web text includes content from a Firefox discussion forum, conversations overheard in New York, the movie script of Pirates of the Caribbean, personal advertisements, and wine reviews.

There is also a corpus of instant messaging chat sessions, originally collected by the Naval Postgraduate School for research on automatic detection of Internet predators. The corpus contains over 10,000 posts, anonymized by replacing usernames with generic names of the form “UserNNN”, and manually edited to remove any other identifying information. The corpus is organized into 15 files, where each file contains several hundred posts collected on a given date, for an age-specific chatroom (teens, 20s, 30s, 40s, plus a generic adults’ chatroom). The filename contains the date, chatroom, and number of posts; e.g., 10-19-20s\_706posts.xml contains 706 posts gathered from the 20s chat room on 10/19/2006.

1. **Brown Corpus:**

The Brown Corpus was the first million-word electronic corpus of English, created in 1961 at Brown University. This corpus contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, and so on.

We can access the corpus as a list of words or a list of sentences (where each sentence is itself just a list of words). We can optionally specify particular categories or files to read.

1. **Reuters Corpus:**

The Reuters Corpus contains 10,788 news documents totalling 1.3 million words. The documents have been classified into 90 topics, and grouped into two sets, called “training” and “test”; thus, the text with fileid 'test/14826' is a document drawn from the test set. This split is for training and testing algorithms that automatically detect the topic of a document.

Unlike the Brown Corpus, categories in the Reuters Corpus overlap with each other, simply because a news story often covers multiple topics. We can ask for the topics 44 | Chapter 2: Accessing Text Corpora and Lexical Resources covered by one or more documents, or for the documents included in one or more categories. For convenience, the corpus methods accept a single fileid or a list of fileids.

1. **Inaugural Address Corpus:**

The corpus is actually a collection of 55 texts, one for each presidential address. An interesting property of this collection is its time dimension.

1. **Text Corpus Structure:**

The simplest kind lacks any structure: it is just a collection of texts. Often, texts are grouped into categories that might correspond to genre, source, author, language, etc. Sometimes these categories overlap, notably in the case of topical categories, as a text can be relevant to more than one topic. Occasionally, text collections have temporal structure, news collections being the most common example. NLTK’s corpus readers support efficient access to a variety of corpora, and can be used to work with new corpora.

Common structures for text corpora:

* The simplest kind of corpus is a collection of isolated texts with no particular organization.
* Some corpora are structured into categories, such as genre (Brown Corpus).
* Some categorizations overlap, such as topic categories (Reuters Corpus).
* Other corpora represent language use over time (Inaugural Address Corpus).

1. **Conditional Frequency Distributions:**

When the texts of a corpus are divided into several categories (by genre, topic, author, etc.), we can maintain separate frequency distributions for each category. This will allow us to study systematic differences between the categories. In the previous section, we achieved this using NLTK’s ConditionalFreqDist data type. A conditional frequency distribution is a collection of frequency distributions, each one for a different “condition.” The condition will often be the category of the text.

1. **Generating Random Texts with Bigrams:**

* We can use a conditional frequency distribution to create a table of bigrams.
* The bigrams () function takes a list of words and builds a list of consecutive word pairs.
* We treat each word as a condition, and for each one we effectively create a frequency distribution over the following words. The function generates\_model () contains a simple loop to generate text.
* When we call the function, we choose a word (such as 'living') as our initial context.
* Then, once inside the loop, we print the current value of the variable word, and reset word to be the most likely token in that context (using max ()); next time through the loop, we use that word as our new context.

1. **Lexical Resources Intro:**

A lexicon, or lexical resource, is a collection of words and/or phrases along with associated information, such as part-of-speech and sense definitions. Lexical resources are secondary to texts, and are usually created and enriched with the help of texts.

**A lexical entry** consists of a **headword** (also known as a **lemma**) along with additional information, such as the part-of-speech and the sense definition. Two distinct words having the same spelling are called **homonyms.**

The simplest kind of lexicon is nothing more than a sorted list of words. Sophisticated lexicons include complex structure within and across the individual entries.

1. **Wordlist Corpora:**

NLTK includes some corpora that are nothing more than wordlists. The Words Corpus is the /usr/dict/words file from Unix, used by some spellcheckers. We can use it to find unusual or misspelled words in a text corpus.

There is also a corpus of stop words, that is, high-frequency words such as the, to, and also that we sometimes want to filter out of a document before further processing. Stop words usually have little lexical content, and their presence in a text fails to distinguish it from other texts.

One more wordlist corpus is the Names Corpus, containing 8,000 first names categorized by gender. The male and female names are stored in separate files.

1. **Pronouncing Dictionary:**

A slightly richer kind of lexical resource is a table (or spreadsheet), containing a word plus some properties in each row. NLTK includes the CMU Pronouncing Dictionary for U.S. English, which was designed for use by speech synthesizers.

For each word, this lexicon provides a list of phonetic codes—distinct labels for each contrastive sound—known as phones. Observe that fire has two pronunciations (in U.S. English): the one-syllable F AY1 R, and the two-syllable F AY1 ER0. The symbols in the CMU Pronouncing Dictionary are from the Arpabet.

1. **Comparative Wordlists:**

Another example of a tabular lexicon is the comparative wordlist. NLTK includes so-called Swadesh wordlists, lists of about 200 common words in several languages. The languages are identified using an ISO 639 two-letter code.

We can access cognate words from multiple languages using the entries() method, specifying a list of languages.

1. **Wordnet:**

WordNet is a semantically oriented dictionary of English, similar to a traditional thesaurus but with a richer structure. NLTK includes the English WordNet, with 155,287 words and 117,659 synonym sets.

1. **Wordnet Hierarchy:**

WordNet synsets correspond to abstract concepts, and they don’t always have corresponding words in English. These concepts are linked together in a hierarchy.

* Some concepts are very general, such as **Entity, State, Event**; these are called unique beginners or root synsets.
* Others, such as **gas guzzler and hatchback**, are much more specific.

1. **More Lexical Relations:**

**Hypernyms and hyponyms** are called lexical relations because they relate one synset to another. These two relations navigate up and down the “is-a” hierarchy. Another important way to navigate the WordNet network is from items to their components (meronyms) or to the things they are contained in (holonyms).

1. **Semantic Similarity:**

Given a particular synset, we can traverse the WordNet network to find synsets with related meanings. Knowing which words are semantically related is useful for indexing a collection of texts, so that a search for a general term such as vehicle will match documents containing specific terms such as limousine.

Each synset has one or more hypernym paths that link it to a root hypernym.

Two synsets linked to the same root may have several hypernyms in common.

If two synsets share a very specific hypernym—one that is low down in the hypernym hierarchy—they must be closely related.