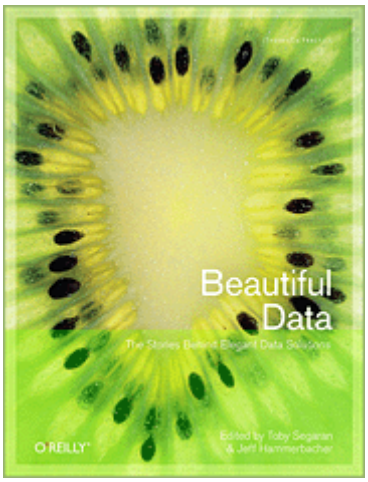
Natural Language Corpus Data: Beautiful Data



O'Reilly / Amazon / Google Books

This directory contains code and data to accompany the chapter *Natural Language Corpus Data* from the book *Beautiful Data* (Segaran and Hammerbacher, 2009). If you like this you may also like: *How to Write a Spelling Corrector*.

Data files are derived from the *Google Web Trillion Word Corpus*, as described by Thorsten Brants and Alex Franz, and distributed by the Linguistic Data Consortium.

To run this code, download either the zip file (and unzip it) or all the files listed below. Then from a shell execute `python -i ngrams.py` (or start a Python IDE and `import ngrams`), and if you want to test if everything works, call `test()`. Note that the `hillclimbing` function has a random component, so if you have bad luck it is possible that some of the tests will fail, even if everything is correctly installed. (It is unlikely that they will fail twice in a row.)

Files for Download

.

6.6MB	ngrams.zip	A zip file of all the files below. Get this *or* the files below.
0.7MB	ch14.pdf	The chapter from the book.
0.0 MB	ngrams.py	The Python code for everything in the chapter.
0.0 MB	ngrams-test.txt	Unit tests; run by the Python function `test()`.
4.9 MB	count_1w.txt	The 1/3 million most frequent words, all lowercase, with counts. (Called `vocab_common` in the chapter, but I changed file names here.)
5.6 MB	count_2w.txt	The 1/4 million most frequent two-word (lowercase) bigrams, with counts.
0.0 MB	count_2l.txt	Counts for all 2-letter (lowercase) bigrams.
0.2 MB	count_3l.txt	Counts for all 3-letter (lowercase) trigrams.
0.0 MB	count_1edit.txt	Counts for all single-edit spelling correction edits, from the file `spell-errors.txt`.
0.5 MB	spell-errors.txt	A collection of "right: wrong1, wrong2" spelling mistakes, collected from Wikipedia and Roger Mitton.
		The following files are not referenced in the chapter, but may be useful to you.
6.5	big.txt	File of running text used in my spell correction article.

MB		
1.0 MB	smaller.txt	Excerpt of file of running text from my spell correction article. Smaller; faster to download.
0.3 MB	count_big.txt	A word count file (29,136 words) for big.txt.
1.5 MB	count_1w100k.txt	A word count file with 100,000 most popular words, all uppercase.
.02 MB	words4.txt	4360 words of length 4 (for word games)
.04 MB	sgb-words.txt	5757 words of length 5 (for word games) from Knuth's Stanford GraphBase
.03 MB	words.js	1000 most common words of English from xkcd Simple Writer (more than 1,000 words because plurals are included)
4.3 MB	shakespeare.txt	The complete works of Shakespeare, tokenized so that there is a space between words and punctuation. From John DeNero.
3.0 MB	sowpods.txt	The SOWPODS word list (267,750 words) -- used by Scrabble players (except in North America) and in other word games.
1.9 MB	TWL06.txt	The Tournament Word List (178,690 words) -- used by North American Scrabble players.
1.9 MB	enable1.txt	The ENABLE word list (172,819 words) -- also used by word game players. Words with Friends uses a variant of this.
2.7 MB	word.list	The YAWL (Yet Another Word List) word list (263,533 words) -- formed by combining the above.
		(See Internet Scrabble Club for more lists.)

Peter Norvig, 8 July 2008; updated 22 Nov 2011