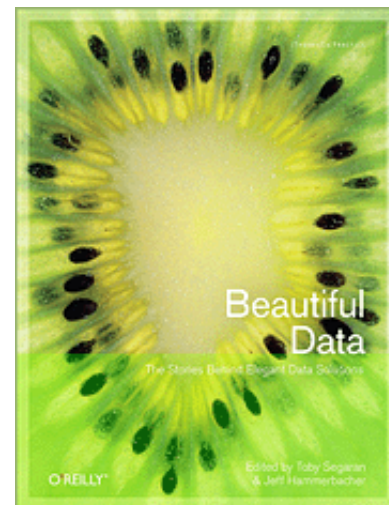


# Natural Language Corpus Data: Beautiful Data

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.



[O'Reilly](#) / [Amazon](#) / [Google Books](#)

Code copyright (c) 2008-2009 by Peter Norvig. You are free to use this code under the [MIT license](#).

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	<a href="#">ngrams.zip</a>	A zip file of all the files below. Get this <i>or</i> the files below.
0.7MB	<a href="#">ch14.pdf</a>	The chapter from <a href="#">the book</a> .
0.0 MB	<a href="#">ngrams.py</a>	The Python code for everything in the chapter.
0.0 MB	<a href="#">ngrams-test.txt</a>	Unit tests; run by the Python function <code>test()</code> .
4.9 MB	<a href="#">count_1w.txt</a>	The 1/3 million most frequent words, all lowercase, with counts. (Called <code>vocab_common</code> in the chapter, but I changed file names here.)
5.6 MB	<a href="#">count_2w.txt</a>	The 1/4 million most frequent two-word (lowercase) bigrams, with counts.
0.0 MB	<a href="#">count_2l.txt</a>	Counts for all 2-letter (lowercase) bigrams.
0.2 MB	<a href="#">count_3l.txt</a>	Counts for all 3-letter (lowercase) trigrams.
0.0 MB	<a href="#">count_1edit.txt</a>	Counts for all single-edit spelling correction edits, from the file <code>spell-errors.txt</code> .
0.5 MB	<a href="#">spell-errors.txt</a>	A collection of "right: wrong1, wrong2" spelling mistakes, collected from <a href="#">Wikipedia</a> and <a href="#">Roger Mitton</a> .

*The following files are not referenced in the chapter, but may be useful to you.*

6.5 MB	<a href="#">big.txt</a>	File of running text from my <a href="#">spell correction</a> article.
1.0 MB	<a href="#">smaller.txt</a>	Excerpt of file of running text from my <a href="#">spell correction</a> article. Smaller; faster to download.
0.3 MB	<a href="#">count_big.txt</a>	A word count file (29,136 words) for big.txt.
1.5 MB	<a href="#">count_1w100k.txt</a>	A word count file with 100,000 most popular words, all uppercase.
4.3 MB	<a href="#">shakespeare.txt</a>	The complete works of Shakespeare, tokenized so that there is a space between words and punctuation. From John DeNero.
3.0 MB	<a href="#">sowpods.txt</a>	The <a href="#">SOWPODS</a> word list (267,750 words) -- used by Scrabble players (except in North America) and in other word games.
1.9 MB	<a href="#">TWL06.txt</a>	The <a href="#">Tournament Word List</a> (178,690 words) -- used by North American Scrabble players.
1.9 MB	<a href="#">enable1.txt</a>	The <a href="#">ENABLE</a> word list (172,819 words) -- also used by word game players. <a href="#">Words with Friends</a> uses a variant of this.
2.7 MB	<a href="#">word.list</a>	The YAWL (Yet Another Word List) word list (263,533 words) -- formed by combining the above.  (See <a href="#">Internet Scrabble Club</a> for more lists.)

---

[Peter Norvig](#), 8 July 2008; updated 22 Nov 2011