NLP: Building (small) corpora, or "Where to get lots of not-too-specialized English-language text files?"

Asked 16 years, 2 months ago Modified 10 years, 10 months ago

Viewed 2k times Part of NLP Collective

5

Does anyone have a suggestion for where to find archives or collections of everyday English text for use in a small corpus? I have been using Gutenberg Project books for a working prototype, and would like to incorporate more contemporary language. A recent answer here pointed indirectly to a great archive of usenet movie reviews, which hadn't occurred to me, and is very good. For this particular program technical usenet archives or programming mailing lists would tilt the results and be hard to analyze, but any kind of general blog text, or chat transcripts, or anything that may have been useful to others, would be very helpful. Also, a partial or downloadable research corpus that isn't too marked-up, or some heuristic for finding an appropriate subset of wikipedia articles, or any other idea, is very appreciated.

(BTW, I am being a good citizen w/r/t downloading, using a deliberately slow script that is not demanding on servers hosting such material, in case you perceive a moral hazard in pointing me to something enormous.)

UPDATE: User S0rin points out that wikipedia requests no crawling and provides this export tool instead. Project Gutenberg has a policy specified here, bottom line, try not to crawl, but if you need to: "Configure your robot to wait at least 2 seconds between requests."

UPDATE 2 The wikpedia dumps are the way to go, thanks to the answerers who pointed them out. I ended up using the English version from here: http://download.wikimedia.org/enwiki/20090306/ , and a Spanish dump about half the size. They are some work to clean up, but well worth it, and they contain a lot of useful data in the links.

NLP nlp linguistics corpus

Share

Improve this question

Follow

edited May 23, 2017 at 12:01

Community Bot

1 ● 1

asked Sep 26, 2008 at 2:15

unmounted

34.3k ● 16 ● 63 ● 61

7 Answers

Sorted by: Highest score (default)

- Use the Wikipedia dumps

8

- needs lots of cleanup

- See if anything in nltk-data helps you
  - the corpora are usually quite small
- the Wacky people have some free corpora
  - tagged
  - you can spider your own corpus using their toolkit
- Europarl is free and the basis of pretty much every academic MT system
  - spoken language, translated
- The Reuters Corpora are free of charge, but only available on CD

You can always get your own, but be warned: HTML pages often need heavy cleanup, so restrict yourself to RSS feeds.

If you do this commercially, the LDC might be a viable alternative.

Share Improve this answer

Follow

answered Sep 26, 2008 at 8:32

Torsten Marek

86.3k • 21 • 93 • 98

Also, 40 million words is an ideal size, and what I have been using until now, the Europarl data is perfect. – unmounted Sep 26, 2008 at 9:05

RSS feeds might not be so clean either if they use CDATA, which allows HTML to be embedded into the feed.

– VirtuosiMedia Mar 4, 2009 at 1:10

4

Wikipedia sounds like the way to go. There is an experimental Wikipedia API that might be of use, but I have no clue how it works. So far I've only scraped Wikipedia with custom spiders or even `wget`.

Then you could search for pages that offer their full article text in RSS feeds. RSS, because no HTML tags get in your way.

Scraping mailing lists and/or the Usenet has several disatvantages: you'll be getting AOLbonics and Techspeak, and that will tilt your corpus badly.

The classical corpora are the Penn Treebank and the British National Corpus, but they are paid for. You can read the Corpora list archives, or even ask them about it. Perhaps you will find useful data using the Web as Corpus tools.

I actually have a small project in construction, that allows linguistic processing on arbitrary web pages. It should be ready for use within the next few weeks, but it's so far not really meant to be a scraper. But I could write a module for it, I guess, the functionality is already there.

Share Improve this answer

Follow

answered Sep 26, 2008 at 8:07

Aleksandar Dimitrov

9,457 ● 3 ● 43 ● 49

Just a note, Wikipedia ask that people do not use web crawlers to view pages see: en.wikipedia.org/wiki/… They provide runtime access through 'export' pages, see: en.wikipedia.org/wiki/Special:Export – user63899 Mar 31, 2009 at 19:06

1

If you're willing to pay money, you should check out the data available at the Linguistic Data Consortium, such as the Penn Treebank.

Share Improve this answer

Follow

answered Oct 10, 2008 at 22:32

Robert Elwell

6,668 ● 1 ● 30 ● 32

1

Wikipedia seems to be the best way. Yes you'd have to parse the output. But thanks to wikipedia's categories you could easily get different types of articles and words. e.g. by parsing all the science categories you could get lots of science words. Details about places would be skewed towards geographic names, etc.

Share Improve this answer

Follow

answered May 16, 2009 at 18:26

Amandasaurus

60.5k ● 74 ● 193 ● 255

I did find a great dataset from wikipedia a couple of weeks ago, I will update the answer. Wikimedia foundation makes

some incredible resources available. – unmounted May 17, 2009 at 1:46

0

You've covered the obvious ones. The only other areas that I can think of too supplement:

1) News articles / blogs.

2) Magazines are posting a lot of free material online, and you can get a good cross section of topics.

Share Improve this answer

Follow

answered Sep 26, 2008 at 4:24

torial

13.1k ● 9 ● 65 ● 89

0

Looking into the wikipedia data I noticed that they had done some analysis on bodies of tv and movie scripts. I thought that might interesting text but not readily accessible -- it turns out it is everywhere, and it is structured and predictable enough that it should be possible clean it up. This site, helpfully titled "A bunch of movie scripts and screenplays in one location on the 'net", would probably be useful to anyone who stumbles on this thread with a similar question.

Share Improve this answer

Follow

answered Sep 27, 2008 at 0:37

unmounted

34.3k ● 16 ● 63 ● 61

One problem with TV and movie scripts is that they would be copyrighted. So you'd have to be careful about copyright.
– Amandasaurus May 17, 2009 at 11:14

0

You can get quotations content (in limited form) here: http://quotationsbook.com/services/

This content also happens to be on Freebase.

Share Improve this answer

Follow

answered Jan 30, 2014 at 12:29

Amit Kothari
530 ● 8 ● 27