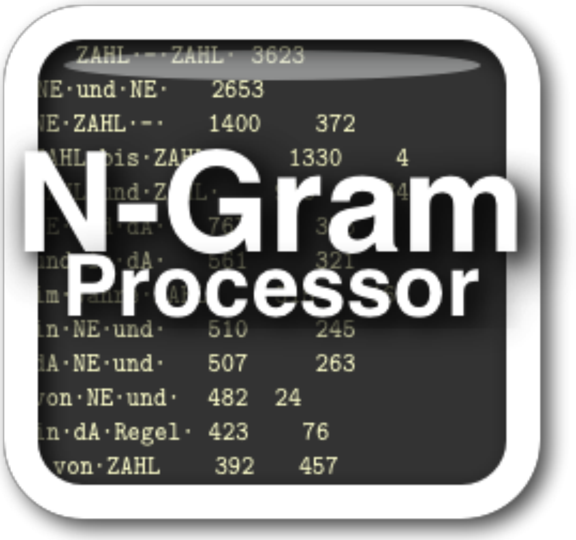

N-Gram Processor (NGP)

see the included manual_0.6.pdf for detailed information, including a tutorial

DESCRIPTION

The N-Gram Processor is a set of scripts and a Perl module allowing the creation and processing of n-gram lists out of text files. It features the following:

- creation of word n-gram lists out of input text, incl. frequencies
- listing of document counts (in how many documents an n-gram occurs)
- unicode support
- support for processing of reasonably large (10 million words or more) corpora, given appropriately powerful hardware
- support for processing of annotated corpora

Please refer to the PDF-manual for a more detailed description.

The codebase of the NGP is partly a branch of two versions of the Ngram Statistics Package (NSP) by Ted Pedersen and collaborators (also known as Text::NSP). The N-Gram Processor can be used for broadly the same purposes as the NSP – differences lie in the following areas:

- support for unicode-encoded in- and output and multi-language awareness
- generating document counts for n-grams
- modifications to allow the processing of larger amounts of data
- no statistics module included (but compatible with the NSP's statistics module under certain conditions, see manual)

The two software packages are otherwise completely separate, so both can be installed on the same machine without causing conflicts.

N-Gram Processor was tested under MacOS X, Xubuntu Linux and the Cygwin environment under Windows (cf. http://cygwin.org). It should also work well on any other platform that can run Perl code and bash shell code.

INSTALLATION

Using the supplied installers (recommended):

Double-clickable installers are provided for OS X, Xubuntu-Linux and the Cygwin environment under Windows. For other environments, please follow the manual installation instructions further down.

OS X / Xubuntu

Inside the `ngramprocessor_0.0` directory, double-click on `Xubuntu_installer` (for Xubuntu), `OSX_installer` (OS X). Follow the instructions of the installer. OS X might prompt users to install the command line tools – this is a free download from Apple and is needed to install the NGP. A video walk through the installation process for MacOS 10.15 is found here

Windows

The Cygwin environment needs to be installed first. During the installation procedure for Cygwin, the following optional packages need to the installed:

- 'bc' from the 'maths' category
- 'make' from the 'devel' category
- 'makemaker' from the 'perl' category
- 'diffutils', 'ncurses' and 'cygutils-extra' from the 'utils' category

A guide on how to install Cygwin is found here. After Cygwin has been installed, double click on the `Cygwin_installer` or `Cygwin64_installer` (try both if one does not work) to start the installation process.

Manual installation / other flavours of Linux

1. open a Terminal window

   OS X: in Applications/Utilities

   Xbuntu Linux: via menu Applications>Accessories>Terminal

   Cygwin: via the link on the Windows desktop to Cygwin Terminal

2. drop the `install.sh` script (located inside the `ngramprocessor_0.0 directory`) onto the terminal window and press ENTER. This should start the installation process.

If an entirely manual installation is necessary, type the following commands into a terminal window while in the `ngramprocessor_0.0` directory:

```
perl Makefile.PL
make
make test
make install
```

The last command requires administrative privileges, so it might need to be run as `sudo make install`, for example on OS X. This installs the files in the standard locations. The N-Gram Processor can then be launched from a terminal window by typing `NGP.sh` or by double-clicking on its icon (if the installers were used for installation). For more details, see the PDF-manual.

AUTHOR

Andreas Buerki, buerkiA@cardiff.ac.uk
Authors of the N-gram Statistics Package (NSP), of which N-Gram Processor is a branch: (v1.09) Ted Pedersen, Satanjeev Banerjee, Amruta Purandare, Bridget Thomson-McInnes, Saiyam Kohli; (the v1.10 re-write) Bjoern Wilmsmann.

SEE ALSO

http://buerki.github.io/ngramprocessor/

http://buerki.github.io/SubString/

http://www.d.umn.edu/~tpederse/nsp.html

https://github.com/BjoernKW/Publications/blob/master/Re-write*of*Text-NSP.pdf

COPYRIGHT