Linguistics 696f
Hammond
Spring '03

Toolkits

A. Overview

(1) a. Basic unix commands
b. Obtaining texts
c. Massaging
d. Using the toolkit

(2) There are two free statistical NLP packages that have been installed on the u-cluster: CMU-Cambridge and SRILM. This handout treats the latter.

B. Basic unix stuff

(3) Logging in: use the `ssh` program to connect to `shell.u.arizona.edu`.

(4) Logging out: enter the command `logout` at the prompt.

(5) Basic navigation commands:
a. `ls`: list files in the current directory.
b. `cd`: change directories.
c. `pwd`: what is the current directory?

(6) File manipulation commands:
a. `mv`: rename or move a file.
b. `cp`: copy a file.

(7) Information about files:
a. `more`: scroll the file screen by screen.
b. `cat`: print the whole file to the screen.
c. `head`: print out the first few lines of a file.
d. `tail`: print out the last few lines of a file.
e. `wc`: count the lines, words, characters in a file.

(8) Editing a file:
a. `pico`: the editor used in `pine`.
b. `emacs`: very powerful arcane editor.
c. `vi`: more powerful more arcane editor.

(9) Other important commands:

a. `man`: additional help on any command.

b. `quota -v`: how much free space do you have?

b. `xdisk`: temporarily augment your disk quota if you have big files to work with.

(10) Piping/chaining:

a. `command`$_1$ | `command`$_2$: takes the output of one command and sends it on to another.

b. `command` > `filename`: takes the output of a command and puts it into a new file.

c. `command` >> `filename`: takes the output of one command and appends it to an existing file.

C. Obtaining texts

(11) Tons of corpora are available for free over the web. (See the links on the course website: `http://linguistics.arizona.edu/~hammond/ling696f-sp03/`.) For this demonstration, we use literary texts available from Project Gutenberg.

(12) Assume you are working on a computer with a fast direct web connection, e.g. at school:

a. Use Netscape/IE to connect to `http://promo.net/pg/`.

b. Navigate to a text of interest, either by author or title.

c. Download relevant files to your desktop computer

d. Use `ssh` to upload those files to the u-cluster.

e. Use the `unzip` command to uncompress files with a `.zip` extension.

(13) Assume you are working on a computer with a slow indirect web connection, e.g. at home:

a. Use `lynx` to connect to `http://promo.net/pg/`.

b. Navigate to a text of interest, either by author or title.

c. Download relevant files directly to the u-cluster.

d. Use the `unzip` command to uncompress files with a `.zip` extension.

D. Massaging

(14) Text files from Project Gutenberg must be "massaged" into a form suitable for analysis with the SRILM toolkit:

a. Use one of the text editors to remove the legal stuff at the beginning of the file.
b. Download the `forngram.pl` program from the course website to your u-cluster account using either of the methods above.
c. The `forngram.pl` file will download as `forngram.txt`. Rename it to `forngram.pl` with `mv`.
d. Use the `forngram.pl` program from the website to split the file into sentences, e.g. `perl forngram.pl textfile > sentencefile`.

(15) Do this with *two* texts of interest.

E. Using the SRILM toolkit

(16) There are two steps in the analyses we will perform. You must first create a language model using some (training) text. You can then use that model with a new (test) text.

(17) Create three n-gram models for one text: trigram, bigram, and unigram models:

a. `ngram-count -text sentencefile1 -lm m3.lm`
b. `ngram-count -text sentencefile1 -order 2 -lm m2.lm`
c. `ngram-count -text sentencefile1 -order 1 -lm m1.lm`

(18) Calculate perplexity for those models with respect to another text:

a. `ngram -lm m3.lm -ppl sentencefile2`
b. `ngram -lm m2.lm -ppl sentencefile2`
c. `ngram -lm m1.lm -ppl sentencefile2`

(19) Use the `man` command with `ngram-count` and `ngram` to find out other options, e.g. different smoothing choices, interpolation, etc.

References

Clarkson, P.R., & R. Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech*.

Stolcke, Andreas. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Denver.